DISCUSSION PAPER / 2020.03
ISSN 2294-8651

Big Data for poverty measurement
insights from a scoping review

Michaëla **Stubbers**
Nathalie **Holvoet**

IOB
Institute of Development Policy
University of Antwerp

Comments on this Discussion Paper are invited.

Instituut voor Ontwikkelingsbeleid en -Beheer
Institute of Development Policy

Postal address:
Prinsstraat 13
B-2000 Antwerpen
Belgium

Visiting address:
Lange Sint-Annastraat 7
B-2000 Antwerpen
Belgium

Tel: +32 (0)3 265 57 70
Fax: +32 (0)3 265 57 71
e-mail: iob@uantwerp.be
http://www.uantwerp.be/iob

DISCUSSION PAPER / 2020.03

Big Data for poverty measurement

insights from a scoping review

Michaëla **Stubbers***

Nathalie **Holvoet**

September 2020

* former IOB Master student

** Professor at the Institute of Development Policy (IOB), University of Antwerp.

IOB
Institute of Development Policy
University of Antwerp

Table of Contents

Abstract

This research presents a scoping review of 53 systematically selected studies that employ Big Data to measure and monitor poverty concepts. The primary aim of the review is to explore if and how Big Data can be used as a replacement or complementary to national and international statistics to identify, measure, and monitor poverty, economic development and inequality on a macro level. The analysis reveals that (1) the relevance of the field so far is driven by data availability, (2) researchers from different fields are involved as data types and analytics employed stem from various research domains, however, researchers from the global south are underrepresented, (3) the main data types used are Call Detail Records (CDR) and satellite image data while night-light is frequently associated with economic development, (4) the choice for certain data types is based on the hypothesis that the manifestations of poverty and development leave traces that are captured by Big Data sources (5) Big Data techniques are so far mainly applied for feature extraction while classical statistical techniques are preferred for analysis.

With this in mind, the review highlights challenges and opportunities of using Big Data for development statistics and briefly discusses the implications for monitoring and evaluation showing that it is highly unlikely that Big Data statistics will replace traditionally generated development data any time soon. Many barriers need to be overcome, including some technical challenges, stability and sustainability issues as well as institutional and legal aspects. In the meantime, Big Data offers undoubtfully a major opportunity to play a role to improve accuracy, timeliness and relevance of socio-economic indicators especially where no data is available, or where quality is highly disputable.

Keywords

Big data, poverty measurement, scoping review, monitoring, evaluation

1. Introduction

The Sustainable Development Goals (SDG) (United Nations, 2015) confirm that eradicating poverty in all its dimensions remains the number one priority. Whereas the consensus on poverty alleviation is clear, little agreement exists on how to define poverty. Poverty definitions and their measurement approaches are central to many theoretical discussions. These definitions matter a great deal as different approaches lead to different policy and targeting implications (Ruggeri Laderchi, Saith, & Stewart, 2003). That said, to fight poverty it is crucial to collect information on where people in poverty live. The geographic distribution of poverty helps to make decisions about targeting resources for poverty alleviation programmes and provides a foundation for studying the determinants of poverty and economic development. Moreover, these data are critical for monitoring poverty rates over time which is in turn pivotal to reach SDG1 (Steele et al., 2017).

Traditional data sources for measuring and monitoring poverty rely on censuses, sample surveys and administrative data (Hackl, 2016; Marchetti et al., 2015). Census data are exhaustive but come with enormous financial costs (Hackl, 2016) which consequently often leads to untimely data collection (Blumenstock, Cadamuro, & On, 2015). This limits their influence on policy making and deters monitoring of short-term effects. Sample surveys can complement census data to overcome major time lags but also involve high costs, especially if local level estimates are needed which ask for oversampling (Marchetti et al., 2015). Furthermore, resistance against the response burden is growing (Hackl, 2016) while obsolete census data also bare the risk of limited representativeness of sampling frames. Administrative data on the other hand are mostly owned by public authorities and generally contain information on the full population of well-defined units. However, users of administrative data need to be aware of the specific quality issues for each data source including the issue of representativeness (Hackl, 2016). In the developing world low capacity of ministries and sometimes conflicting interests can pose additional challenges to quality (Jerven, 2013; Elvidge et al., 2009).

Collecting data to produce high-resolution estimates of poverty, economic development and inequality indicators thus involves a lot of theoretical and practical decisions. In low-income countries the data gathering is furthermore hampered by limited resources and capacities which may result in uncertain information (Steele et al., 2017). As discussed in Jerven (2013), a great deal of what we know about a substantial part of the developing world is likely to be shaped by poor numbers. However, in the production process statistics and indicators achieve an air of certainty and objectivity which is typically associated with knowledge generation and as such feeds into decision making. An indicator may dictate which policies to implement, continue or discontinue, and point out different resource allocation scenarios within the country or, at a global level (Jerven, 2013). Moreover, indicators are often used for transparency and accountability. Given their widespread usage and consequences sound indicator selection and data collection is consistently high on the research and policy agenda. Recently, with the digitalisation of our society and the exponential growth of data globally, there are promising signs that novel Big Data sources can help providing cost- and time efficient, accurate and up-to-date indicators (Steele et al., 2017).

The popularisation of the term Big Data over the past decade led the notion Big Data to be used for several and inconsistent meanings though. As discussed in Hackl (2016) and De Mauro, Greco and Grimaldi (2016), the notion lacks a formal definition. The latter propose to define Big Data as: " The Information asset characterised by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value" (De Mauro, Greco, & Grimaldi, 2016, p. 131). This definition includes the first 3 V's frequently used to describe Big Data and adds the importance to create value by using specific technologies and analytics. Another definition, geared towards the use of Big Data in development, focusses on how the quantitative growth of digital capacity evoked five new qualitative features for treating this kind of data: (1) Big Data are a by-product of other services and produced anyway, (2) Big Data do not use random sampling but uses mostly all there is, (3) Big Data are often available real-time, (4) Big Data often merge different sources, and (5) the full name of Big Data is Big Analytics (Hilbert, 2015, p.137). This definition deemphasizes the volume aspect arguing that the kind of data and how the data is produced are more relevant. Altogether, Big Data are less a matter of volume than the capability to gather information and create knowledge where previously it was economically or technically not feasible to do so (Sonka, 2016).

Big Data are currently gathering momentum as an opportunity for monitoring and evaluating progress towards the SDGs, in particular SDG1, where it is considered a useful alternative or supplement to official statistics. Yet, so far, Big Data analysis for monitoring the SDGs, and more broadly, development outcomes, is still in its infancy. Against this background, our article presents the findings of a scoping review taking stock of case studies which employed Big Data to measure SDG1 related concepts . These concepts are linked to national and international statistics which are used for monitoring and evaluation on a macro level and for policy making. Whereas other studies focus on the use of particular Big Data sources such as DMSP/

OLS Nighttime Light Images (e.g. Huang, Yang, Gao, Yang, & Zhao, 2014), our review aims to cover all studies that try to quantify concepts of poverty, development or inequality through the use of Big Data. The guiding question is "Which Big Data types and methodologies can be used as a replacement or complementary to existing national and international statistics to identify, measure, and monitor poverty, economic development and inequality and what are the most important challenges?". In the next section, we provide a detailed account of the scoping review methodology.

2. Methodology

To map out different Big Data types and methodologies that have so far been used for poverty monitoring, a scoping review was performed along the lines suggested by Pham et al. (2014) and Levac, Colquhoun, & O'Brien (2010). To identify relevant cases, three multidisciplinary electronic databases were searched: EBSCOhost, Web of Science and Scopus using search terms related to poverty, development, inequality and Big Data. All searches included journal articles and conference proceedings published (in English) between January 2000 and March 2018. Table 1 presents further details about the search strategy and process.

After deduplication, 451 references remained for subsequent title, abstract and free full text availability[1] screening. The latter led to a further reduction to 49 references which were then subject to a full text screening. To be included in the final list for more in-depth review and analysis the cases needed to use Big Data to measure, monitor or map poverty, inequality or development with the aim to inform policy making. To that end, research using remote sensing specifically aimed at measuring or mapping urban poverty or inequality were retained while case studies related to e.g. the mapping of slums but focusing on application areas such as land management, urban planning, crime studies or environmental hazards were excluded. As a result of this step in the selection process, the pool of publications was further reduced to 28.

In a second phase, a backward snowball technique was adopted in which the reference lists of all 28 retained articles were manually searched to identify other articles meeting the inclusion criteria. The resulting 74 additional unique references were screened in line with the criteria used in phase one. Finally, 53 studies were selected to be included in the review (see Appendix A for the flow diagram depicting detailed information about the different phases of the selection process).

Whereas qualifying the use of Big Data was straightforward and based on its definition, judging the alignment with SDG1 proved to be more challenging given the range of existing poverty measurements, the philosophical debates around its definition and the many facets of well- and ill-being directly or indirectly related to poverty and socio-economic development.

Table 1: Search strategy

Search Strategy – 1		
Search term title		
Problem	"Poverty" OR "SDG" OR "socio-economic"	
Interest	"Big Data" or "mobile" or "sensing" or "satellite" or "night-time light" or "search" or "crowdsourcing" or "social media" or "scraping" or "internet"	
Databases	Date	Number of retained references
EBSCOhost	11/07/2018	104

[1] Free full text availability with institutional account.

Web of Science	10/07/2018	107
Scopus	10/07/2018	131

Search Strategy – 2		
Search term title		
Problem	"Inequality"	
Interest	“Big Data” or “mobile” or “sensing” or “satellite” or “night-time light” or “search” or “crowdsourcing” or “social media” or “scraping” or “internet”	
Databases	Date	Number of retained references
EBSCOhost	11/07/2018	69
Web of Science	11/07/2018	164
Scopus	11/07/2018	155

Search Strategy – 3		
Search term title		
Problem	(“Vulnerable” or “vulnerability”) and (“Population” or “people” or “groups” or “victims”)	
Interest	“Big Data” or “mobile” or “sensing” or “satellite” or “night-time light” or “search” or “crowdsourcing” or “social media” or “scraping” or “internet”	
Databases	Date	Number of retained references
EBSCOhost	12/07/2016	19
Web of Science	12/07/2016	16
Scopus	12/07/2016	19

Search Strategy – 4		
Search term title		
Interest	“social mobility” or “access to services” or “access to basic services” or “access to public services” or “social exclusion”	
Context	“Big Data” or “mobile” or “sensing” or “satellite” or “night-time light” or “search” or “crowdsourcing” or “social media” or “scraping” or “internet”	
Databases	Date	Number of retained references
EBSCOhost	12/07/2016	9
Web of Science	12/07/2016	2
Scopus	12/07/2016	0

3. Findings and Analysis

3.1. Overall research trends

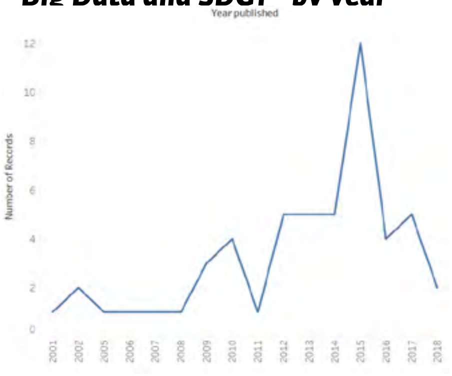
Figure 1 highlights the distribution of publications over time. An upward trend is noticed between 2009 and 2012, with a clear upsurge in 2015. Although based on a rather small number of publications, the trend noticed is not in line with the steady yearly increases one would expect and which was e.g. described by Huang et al. (2014) and Kuffer, Pfeffer, & Sliuzas (2016) (Kuffer et al., 2016) who both covered satellite image data only. The upsurge in 2015 is especially driven by an emergence of cases based on mobile phone data which is related to the ‘Data for Development’ (D4D) Challenges[2] launched by Orange in 2012 and 2014. In general, mobile phone operators and corporate institutions overall, are rather reluctant to share data related to mobile phone and other technology usage out of privacy and competitive concerns (Njuguna & Mcsharry, 2017). The D4D Challenges made anonymous data, extracted from the mobile networks in Ivory Coast and Senegal, open to the public with the objective to contribute

[2] ‘Data for Development’ was a global innovation challenge organised by Sonatel and the Oranje Group that allowed researchers to use anonymized, aggregated Call Detail Record (CDR) data to help solve various development problems.

to the development and welfare of populations ("data for development," n.d.).

Figure 1: Number of publications related to "Big Data and SDG1" by year

Number of publications related to "Big Data and SDG1" by year



Source: author's own compilation

In total 173 authors contributed to the 53 publications selected, with a majority of 87% contributing only to one publication. The articles are published in 32 different journals coming from a wide variety of scientific backgrounds amongst others economics, geography, ecology, space research and development research. 84% of these journals published only one of the selected studies. Focusing on country of authorship's affiliation, the USA and the UK are most prominently represented with 43% and 23% of publications respectively, closely followed by China responsible for 17% of the papers.

Inspired by the influential article by C. D. Elvidge et al. published in 1997 which investigated for the first time the relationship between night-time light and the Gross Domestic Product (GDP), 25% of the selected papers focus on 'global' development. About 49% relate to countries in the Global South[3], 17% focus on China while only 5 publications (9%) analyse poverty or inequality in Western countries. Countries and/or regions studied are listed in detail in Table Annex B.

In sum, Big Data analytics for development statistics engages researchers of various backgrounds which does not come as a surprise as the types of Big Data and methodologies to analyse various data types originate from different scientific domains. Added to this, poverty can be approached from multiple angles and as poverty or development is linked to many areas of life, measuring poverty using novel technologies is influenced by different disciplines which underscores the importance of a holistic approach (Rindfuss & Stern, as cited in Taubenböck et al., 2009, p.1). That said, the discrepancy between the involvement of authors of the Global South, combined with the nature of Big Data that generally does not require any personal involvement, and the number of publications affecting the Global South illustrates an increasing distance between researchers and populations researched (Taylor & Broeders, 2015).

3.2. Big Data sources used

As regards the different sources of Big Data used by the papers under study, we notice the effect of ICTs' development which has generated a stream of fresh and digitized data related to human and non-human activity. Turning all these data into knowledge asks research-

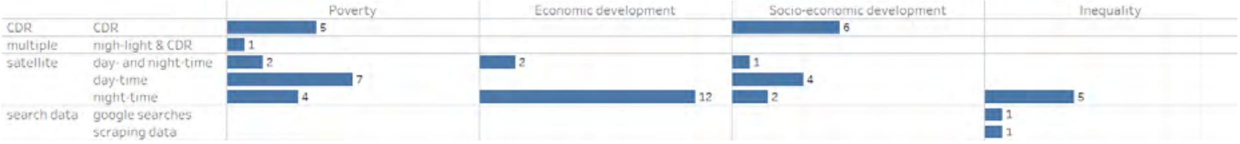
[3] The term global South is used to refer to countries classified by the World Bank as low or middle income and that are located in Africa, Asia, Oceania, Latin America and the Caribbean.

ers to deal with a huge amount of unstructured and heterogeneous data. This requires a careful planning and organisation of the data analysis process, including an assessment of the wide variety of heterogeneous data sources available (Blazquez & Domenech, 2018). To this end, manifold general and domain specific architectures and frameworks have been developed. This review applies the 2013 UNECE framework developed by the United Nations Global Working Group on Big Data for Official Statistics Task Team on Cross-Cutting Issues (TTCC), which was still under progress at the time of our scoping review. The UNECE framework was developed with a possible use of Big Data for compiling SDG indicators in mind. The exercise was more challenging than expected and revealed the complex and fast evolving nature of Big Data as the UNECE taxonomy fails to cover all Big Data types used in our case studies. The review identifies for example CDRs, i.e. the traces generated by mobile phone usage, as one of the major data sources used for poverty mapping. CDRs are not included in the UNECE framework. Other data employed in the case studies and missing in the framework are data generated by extracting data from websites ("web scraping") and data referring to readily available data sources based on at least one Big Data source to compile databases for the use by business or communities.

Our classification exercise highlights that the cases only employed a relatively limited number of data types: 74% are based on satellite image data, another 22% on Call Detail Records (CDRs) and 4% on other data types. The search terminology also included search terms such as "search", "social media" or "internet" and did not focus on numerical data only. In spite of this, not one publication was identified utilizing "qualitative" Big Data such as social media data while only 3 studies included other than CDR or satellite image Big Data types; Walasek & Brown (2016) used search data, while Kholodilin & Siliverstovs (2012) and Sultan et al. (2015) added data gathered by web scraping in their analysis.

Figure 2 illustrates the use of different data sources classified by poverty concept. While economic development is mainly captured by night-light data, CDRs are more frequently used in relation to overall socio-economic development and poverty indicators are correlated to both CDR and sensing data.

Figure 2: Main Big Data sources used by poverty concept – Number of studies



Source: author's own compilation

3.3. Big Data and knowledge generation

Big Data analytics describe techniques for analysing data that are so large and complex that they "require advanced storage, management, analysis, and visualization technologies" (H. Chen & Storey, 2012, p1166). These techniques have evolved considerably over the past decades and are still unfolding due to the ongoing digitalisation of our societies and the scaling-up of earth observation data. In what follows, we provide a stepwise description of how knowledge is generated from Big Data in the reviewed publications.

3.3.1. Choosing the data source

In his 2008 provocative piece "The End of Theory", Anderson refers to the historical quote of George Box, "all models are wrong, but some are useful" when introducing the new era that comes with increased computer power and Big Data. The combination of both offers novel, never seen methods for knowledge generation while the need to speculate and hypothesise is becoming obsolete. Anderson highlights that computers will show us patterns, trends, and relationships also within the field of social, economic and political sciences. In his opinion, correlation will supersede causation and science will be able to advance without coherent models or unifying theories (Anderson, 2008). Reversely, opponents such as Smith, Mashhadi, & Capra (2013), argue that this kind of 'black-box' knowledge generation is of limited value for policy making as the lack of causal insights does not allow to provide guidance on what can be done or what should be avoided. Algorithms based on such a theory-free research paradigm can at its best lead to amusing results but at its worst initiate disastrous decisions, which is also related to the concept of 'past data' (see Hilbert, 2015). More specifically, predictions based on data generated in the past work fine when the future resembles the past. In case of significant changes to the modus operandi of the system, the future will not reflect any longer the past and the logic does not work. As discussed in Hilbert (2015), to predict a never been future or to simulate what-if scenario's, theory-driven modelling still is a necessity. Especially in a developmental context it is important to avoid adverse consequences and even status-quo solutions and increase our understanding of what drives and hinders development to contribute to sound policy making.

Amongst the reviewed papers, even the most 'black-box' like studies, did not use just any kind of data available. The choice for a particular Big Data type is in all cases driven by a hypothesis that relates the data to the concept of interest. The data captures traces of human behaviour or activity that are assumed to correlate with poverty or development and as symptoms of poverty can vary by context, these traces may vary as well. To be relevant for policymaking it is crucial to know which behaviours or activities the data are assumed to represent. Furthermore, analysing the underlying theories can reveal potential research gaps both for (1) the manifestations of poverty and/or development as for (2) the traces these manifestations leave behind in the digital world.

Table 2 and Table 3 give a summative account of the concepts captured in our case studies and the theories linking these concepts to night and day-time satellite image data. The relationship between night-light and economic development is heavily disputed as this relationship is highly complex. Night-light reflects outdoor and some indoor use of light and as nearly all human consumption and investment activities require light when it is dark, it is expected that the usage of light increases with economic development. However, as night-light does not capture activities of economic sectors for which no activity takes place in the dark (Ghosh, Powell, Elvidge, et al., 2010), satellites cannot detect light used inside cavernous production plants while they can also not distinguish between plants that are running full or mostly empty capacity nor between high valued and low valued economic activity (Mellander, Lobo, Stolarick, & Matheson, 2015).

Table 2: Theories underpinning the use of satellite night-time imagery

Concept	Theory	References
Economic development (GDP)	• Night-light intensity reflects human economic and consumption activity.	(X. Chen & Nordhaus, 2011; Dai, Hu, & Zhao, 2017; Doll, Muller, & Morley, 2006; Ebener, Murray, Tandon, & Elvidge, 2005; Li, Xu, Chen, & Li, 2013; Mellander et al., 2015; Shi et al., 2014; Sutton, 2007; Zhao et al., 2017)

Informal economic activity	• Night-light better reflects total economic activity than official statistics as it also captures informal economy.	(Ghosh, Anderson, Powell, Sutton, & Elvidge, 2009; Ghosh, Powell, Anderson, Sutton, & Elvidge, 2010; Ghosh, Powell, Elvidge, et al., 2010; Henderson, Storeygard, & Weil, 2012)
Poverty	• Night-light is a proxy estimate for wealth.	(Elvidge et al., 2009; Kimijiama & Nagai, 2014; Noor et al., 2008; Wang et al., 2012)
	• Poverty is connected to access to basic services such as electricity which in turn is captured in night-light.	(Doll & Pachauri, 2010; Xu et al., 2015).
Socio-economic development	• Night-light reflects social-economic development levels.	(Levin & Duke, 2012; Nischal, Radhakrishnan, Mehta, & Chandani, 2015; Proville, Zavala-Araiza, & Wagner, 2017)
Inequality	• Spatial inequalities of development are reflected in the spatial disparity of night-light.	(Elvidge et al., 2012; Lessmann & Seidel, 2017; Wu et al., 2018; Zhou et al., 2015)

Table 3: Theories underpinning the use of satellite day-time imagery

Concept	Theory	References
Economic development (GDP)	• The environment supports human life and production. High-resolution day-light satellite imagery contains an abundance of information about environmental morphology that potentially correlate directly, or indirectly through night-light, to economic activity.	(Jean et al., 2016; Sutton & Costanza, 2002; Zhao et al., 2017)
Poverty	• Environmental morphology may indicate the presence of people living in poverty but is context specific and may differ in urban as opposed to rural settings.	(Hall, Malcolm, & Piwowar, 2001; Zhao et al., 2017)
Socio-economic development	• Environmental morphology is related to socio-economic status in urban settings.	(Biggs, Anderson, & Pombo, 2015; Taubenböck et al., 2009)

Mobile phone usage reveals patterns linked to poverty such as ‘call activity’ and socio-economic opportunities such as ‘introversion’ or ‘network advantage’. These are captured in CDR data that contain a huge and varied amount of rich information. Data modelling converts this information in a myriad of metrics. Overall the researchers exploring mobile phone meta data to estimate poverty or socio-economic development analyse particular combinations of these CDR features (Table 4).

Table 4: CDR features used to estimate poverty or socio-economic development

Concept	Feature	Theory
Poverty	- Call activity or consumption	- Sometimes also referred to as consumption is related to call volume. This is likely to reflect wealth.
	- Introversion	- Measures the tendency of the population within a region to communicate within instead of outside the region which is connected to access to resources. More introverted areas have fewer access to resources and thus less economic opportunities leading to higher poverty levels.
	- Network advantage	- Captures the opportunity for economic development afforded by an advantageous position in the network.
	- Gravity residuals	- Compute the difference between observed and expected communication flows between locations. This difference reflects the level of wealth in an area.
	- Social network variables	- These include the diversity of contacts. Poverty or wealth of a community can be affected by the structure of the social relations between individuals.
	- Phone ownership	- Calculates an estimate for the fraction of the population owning a mobile phone which is connected to wealth.
Socio-economic development	- Call activity or consumption	- Sometimes also referred to as consumption is related to call volume. Mobile phone usage reveals patterns linked to socio-economic development.
	- Social network variables, including diversity	- Socio-economic development of a community can be affected by the structure of the social relations between individuals.
	- Mobility patterns	- Reveal information about the whereabouts of individuals which are strongly connected to economic opportunity and access to basic services which in turn have an impact on socio-economic development.
	- Phone model	- The phone model is linked to its price. Average prices payed for consumer goods within regions are an indication of spatial socio-economic status.

3.3.2. Data structuring and feature extraction

Big Data come in all shapes and sizes. Data transformation is required to wrangle that variety of data into structured formats and features for analysis. Night-light data is available in pre-processed publicly available datasets from which the features of interest can be extracted. Usually the log of the night-light metrics is applied which can be interpreted as night-light elasticity (e.g. Lessmann & Seidel, 2017). Table 5 lists all night-light features used by the studies reviewed.

Day-light satellite images consist of unstructured data and need image processing. Machine learning may identify day-light image features such as land coverage including vegetation commonly used in connection with environmental indicators (Imran, Stein, & Zurita-Milla, 2014; Levin & Duke, 2012; Morikawa, 2014; Sedda et al., 2015; Watmough, Atkinson, Saikia, & Hutton, 2016; Biggs et al., 2015; Imran et al., 2014), roof tops (Sutton & Costanza, 2002) or structural and textural features (Duque, Patino, Ruiz, & Pardo-Pascual, 2015; Biggs et al., 2015). Jean et al. (2016) train the model to extract features of day-light imagery that predict night-light. Other metrics derived from day-light images include distance measures (Morikawa, 2014; Watmough, Atkinson, & Hutton, 2013; Watmough et al., 2016). Boyd et al. (2018) employed expert visual interpretation in combination with crowd sourcing to code areas of interest and crowd sourcing proved to be a very efficient and accurate way to interpret the images.

CDRs are delivered as structured data logs that require data modelling to retrieve, compute and/or model meaningful metrics. The majority of cases uses a limited set of a-priori motivated metrics which are listed in table 6. Two studies did not define the predictors upfront but use automated methods to select the potential and final covariates (Steele et al., 2017; Blumenstock et al., 2015). Modelling mobile phone data mostly employs combinations of CDR features.

Table 5: Features extracted of night-light images

Feature	N	Reference
▪ Area (un)lit (%)	9	(Doll et al., 2006; Elvidge et al., 2012; Ghosh et al., 2009; Ghosh, Powell, Anderson, et al., 2010; Henderson et al., 2012; Lessmann & Seidel, 2017; Noor et al., 2008; Proville et al., 2017; Sutton, 2007)
▪ Sum of lights	9	(X. Chen & Nordhaus, 2011; Dai et al., 2017; Ghosh, Powell, Elvidge, et al., 2010; Li et al., 2013; Shi et al., 2014; Sutton, 2007; Sutton & Costanza, 2002; Xu et al., 2015; Zhao et al., 2017)
▪ Light intensity	8	(Henderson et al., 2012; Lessmann & Seidel, 2017; Mellander et al., 2015; Nischal et al., 2015; Noor et al., 2008; Wang et al., 2012; Zhao et al., 2017; Zhou et al., 2015)
▪ Frequency saturated pixels	2	(Henderson et al., 2012; Lessmann & Seidel, 2017)
▪ Population in lit area	2	(Ghosh et al., 2009; Ghosh, Powell, Anderson, et al., 2010)
▪ Light present or not	1	(Doll & Pachauri, 2010)
▪ Night time light index	1	(Zhao et al., 2017)
▪ Distance to nearest night-time light	1	(Noor et al., 2008)
▪ Night time light index	1	(Zhao et al., 2017)
▪ Unspecified or unclear	4	(Ebener et al., 2005; Elvidge et al., 2009; Levin & Duke, 2012; Wu et al., 2018)

Notes: the night-time light index is a compound measure of various of the aforementioned night-light indexes.

Table 6: Features extracted of CDR data

Feature	N	Reference
Call activity or consumption	9	(Eagle, Macy, & Claxton, 2010; Frias-Martinez & Virseda, 2012; Gutierrez, Krings, & Blondel, 2013; Mao, Shuai, Ahn, & Bollen, 2015; Njuguna & McSharry, 2017; Pokhriyal, Dong, & Govindaraju, 2015; Chris Smith-Clarke & Capra, 2016; Christopher Smith-Clarke, Mashhadi, & Capra, 2014; Smith et al., 2013)
Introversion	5	(Mao et al., 2015; Pokhriyal et al., 2015; Chris Smith-Clarke & Capra, 2016; Christopher Smith-Clarke et al., 2014; Smith et al., 2013)
Network advantage	4	(Mao et al., 2015; Pokhriyal et al., 2015; Chris Smith-Clarke & Capra, 2016; Christopher Smith-Clarke et al., 2014)
Gravity residuals	4	(Pokhriyal et al., 2015; Chris Smith-Clarke & Capra, 2016; Christopher Smith-Clarke et al., 2014; Smith et al., 2013)
Social network variables, including diversity	3	(Eagle et al., 2010; Frias-Martinez & Virseda, 2012; Smith et al., 2013)
Mobility patterns	2	(Frias-Martinez & Virseda, 2012; Šćepanović, Mishkovski, Hui, Nurminen, & Ylä-Jääski, 2015)
Ownership	1	(Njuguna & McSharry, 2017)
Phone model	1	(Sultan et al., 2015)

Finally, for data matching, GIS techniques are frequently used to match satellite data imagery with ground truth data or additional data sources such as population data. As CDRs are mostly provided anonymously (for privacy concerns), they can only be analysed at aggregate levels and matching usually employs a spatial key defined by the location of the cell tower (e.g. Njuguna & McSharry, 2017). Only two studies link CDRs to consumer data directly at the individual level. This was only possible because the researchers added an independent sample survey to their data collection which allowed to obtain respondents' informed consent (Steele et al., 2017; Blumenstock et al., 2015).

3.3.3. Data analysis

Whereas the data source has a major impact on the technique used for data extraction and feature engineering, the impact is considerably less for choosing an analytical model. Most researchers choose for classical statistical techniques such as correlation and regression analysis: 66% (N=35) use the generalized linear model for both analysing relations and predictive purposes, 13% (N=7) only rely on correlation analysis, while 15% (N=8) use correlations to inform the regression analysis.

The regression models range from very simple (e.g. Noor et al., 2008) to complex multivariate models (e.g. Ebener et al., 2005). Fixed effects models (Lessmann & Seidel, 2017) and distinct sets of models (Doll et al., 2006; Ghosh, Powell, Elvidge, et al., 2010) are presented to account for context effects. Other authors explicitly consider the spatial properties through the use of spatial techniques (Biggs et al., 2015; Imran et al., 2014; Mellander et al., 2015; Steele et al., 2017; Sedda et al., 2015). Only a handful of studies present alternative methods specifically developed for Big Data; geostatistical modelling (Watmough et al., 2016), index calculation based on supervised classification (Hall et al. 2001), and machine learning(Blumenstock et al. 2015). Independent of the complexity of the model, most researchers train the models based on official development statistics.

Most models are fitted or trained on classical indicators such as poverty indexes and GDP. The results of the analyses are highly disparate and difficult to compare. Some case studies are rather exploratory in nature whereas others focus on fine-tuning the predictive power. Predictive power overall increases with the quality of data and the number of features which in turn increases the risk for overfitting and limits its value for policy making. Details of the analyses, including quality of the models, are summarised in Table Annex B.

4. Discussion

Big Data are clearly promising, yet with potential benefits and applications come challenges and dilemmas. This section first discusses challenges and opportunities related to the use of Big Data for measuring SDG1. Table 7 provides a summative account drawing upon the five qualitative features of Big Data as defined by Hilbert (2015). Secondly, the implications for monitoring and evaluating socio-economic development are discussed.

4.1. Challenges and opportunities

4.1.1. Big Data are a by-product of other services and produced anyway

All Big Data sources employed in the retained studies originate as a by-product of a commercial or a public service. The origin of the data leads to several challenges when used for research and monitoring and evaluation.

Firstly, verification and reliability are critical for rigorous research and monitoring and evaluation, and even more so when informing policies in low income countries since the stakes are high. Amongst the reviewed studies all but one (Walasek & Brown, 2016) employed data sources of unvolunteered nature which hampers verification. Further, the data is delivered by a single corporate or public authority. CDR data is released under restricted conditions (Taylor & Schroeder, 2015) while satellite imagery data are pre-processed before being made publicly available and the resulting datasets may differ which can confuse researchers and M&E experts (Huang et al., 2014). Many authors consider the data quality, reliability and consistency over time of night-light imagery as one of its major challenges (Duque et al., 2015; Huang et al., 2014; Jean et al., 2016; Li et al., 2013; Shi et al., 2014; Noor et al., 2008; Sutton, 2007). Likewise, as to date publicly available satellite imagery are too coarse to provide good local estimates (Duque et al., 2015; Noor et al., 2008; Sutton, 2007), some authors additionally process the data (e.g. Jean et al., 2016; Li et al., 2013). However, so far, no standardised methodologies or best practices exist which renders verification and replication very difficult if not impossible.

Table 7: Challenges and opportunities for the use of Big Data to measure socio-economic development

Characteristics	Challenges	Opportunities
(1) Big Data are a by-product of other services and produced anyway	Verification issues. Risk to alienate the researcher from the researched. No support of a sound theory of change. Debatable what the data represents. Data quality sometimes questionable.	Limited financial and labour costs. Appearance of consistency, independence and objectivity. Data can be collected for specific purposes. Big and Small Data can be combined.
(2) Big Data do not use random sampling but uses mostly all there is	Risk to misrepresent the population of interest.	The population of interest can be covered with maximum accuracy. Opens possibilities for new levels of analysis including global, small area and pixel level. Allows for granular data processing and gridded maps.
(3) Big Data are often available real-time	Currently Big Data are made available ad hoc. Time-lags between data sources that are linked may jeopardize the accuracy and reliability of the analysis.	Increased frequency will allow for: Regular updates Trend analysis Dynamic models
(4) Big Data often merge different sources	Merging different data sources poses technical challenges. The compatibility of various data sources needs to be considered carefully.	Opportunities to match different data types and sources including various Big Data sources, traditional data sources or both, are manifold.
(5) and the full name of Big Data is Big Analytics	Analysis merely involves correlations and no causality which bares the risk of limited knowledge generation. Risk of overfitting. Contextual differences. Need for large storage and processing resources and capabilities.	Opportunities to match different data types and sources including various Big Data sources, traditional data sources or both, are manifold.

Secondly, the abundance of cheaply created data may give the impression that human development can be engineered and as such risks to estrange human development priorities and understandings from local realities and from academic insights which might add historical or political context (Taylor & Broeders, 2015). The review furthermore reveals that all publications mainly involved researchers from the North analysing development of the global South, with China being the exception. Additionally, when the data is made available in the context of a contest such as the D4D challenge for instance, one might expect further alienation between the researcher and the researched. Only one study included fieldwork into their research to gain local knowledge (Watmough et al., 2013).

Thirdly, conventional data collection is usually supported by a sound theory of change. Big Data in general require interpretation after data collection rather than upfront; together with limited contextual information and knowledge, this can easily lead to misinterpretations. This does not fully apply to the studies under review as for all of them the selection of a particular data source is derived from an underlying theory of change. However, these theories and assumptions can be challenging which is illustrated by the ongoing debate regarding the use of night-light as a proxy for development. Over time, different authors have showcased that stronger correlations exist between night-light and population density (Ebener et al., 2005; Levin & Duke, 2012; Mellander et al., 2015; Xu et al., 2015), between night-light and road length (Levin & Duke, 2012; Proville et al., 2017) and between night-light and the availability of electricity which in turn depends on population density (Doll & Pachauri, 2010; Proville et al., 2017). Doll et al. (2006) also reflect on the possible influence of globalisation and investments between countries as light detected in one location may not fully contribute to the economy of that location. More generally speaking, the identification of natural and physical capital assets important to livelihood coupled with the ways how these might be reflected in alternative data sources remains a challenge (Watmough et al., 2013).

Reversely, the peculiar origin of Big Data also offers many opportunities amongst which the production at limited financial and labour cost (e.g. Wu et al., 2018; Smith et al., 2013; Hall et al., 2001). This uplifts the cost barrier to collect poverty statistics and delivers a major opportunity to act upon. Furthermore, and contradictory to official statistics, Big Data are surrounded by an appearance of consistency, independence and objectivity. For the first time ever, data offer the opportunity for a consistent assessment of socio-economic development over time and across countries and regions (Doll et al., 2006; C D Elvidge et al., 2012; Christopher D Elvidge et al., 2009; Ghosh et al., 2009; Ghosh, Powell, Elvidge, et al., 2010; Noor et al., 2008; Sutton & Costanza, 2002; Wu et al., 2018). Moreover, Big Data are not affected by inflation nor regional price differences and therefore believed to better reflect real differences (Lessmann & Seidel, 2017; Zhou et al., 2015).

Having said that, techniques such as crowd sourcing can be used to decode Big Data (Boyd et al., 2018) or collect ground truth data (Jean et al., 2016). The data generated this way are no longer a by-product but collected with specific purposes and as such may support a theory of change. One may also combine Big Data with small data collection: while Big Data can for instance help to identify locations of interest based on certain features, traditional data collection methods can be added later on to generate local knowledge.

4.1.2. Big Data do not use random sampling but uses mostly all there is

The non-use of random sampling poses an undeniably challenge when working with data that does not represent the total population of interest as is the case for CDR data. Mobile phone ownership after all, is often still biased towards the more affluent members of society (Bamberger, Raftree, & Olazabal, 2016). Just three papers though mention this limitation explicitly (Frias-Martinez & Virseda, 2012; Njuguna & McSharry, 2017; Chris Smith-Clarke & Capra, 2016).

On the other hand, sensing data covers the total population with maximum accuracy. Data with broad spatial coverage are created while at the same time delivering information on finer spatial scales. This allows for levels of analysis on spatial scales that traditional analyses cannot cover including global coverage (e.g. Elvidge et al., 2009; Ghosh, Powell, Elvidge, et al., 2010; Sutton & Costanza, 2002), small area estimates (e.g. Kholodilin & Siliverstovs, 2012;

Wu et al., 2018; Xu et al., 2015) and estimates at the pixel level (Doll et al., 2006). In the near future, improved high-resolution data with exhaustive coverage will allow to process these data granularly, down to fine spatial levels that were never possible before. In turn, this creates an opportunity to design a new class of gridded poverty maps. These maps can be analysed to whichever spatial unit one is interested in, and are no longer limited to any administrative or other predefined boundaries (Doll et al., 2006; Ghosh, Powell, Elvidge, et al., 2010).

4.1.3. Big Data are often available real-time

The high actuality of Big Data is frequently mentioned as one of its major strengths. In reality, thus far, Big Data are currently only available ad-hoc. CDR data are made available under control of private corporations who are reluctant to provide information out of privacy and commercial concerns. DMSP/OLS data is available on a yearly basis for the period 1992- 2013; since 2013, the more recent VIIRS Day/Night Band Nighttime Lights data is published monthly, with a time lag of two months (NOAA, n.d.). Satellite day imagery comes from many sources amongst which the Landscan data is frequently used; to date (February 2019) the Landscan data for 2017 can be retrieved (LandScan, n.d.). As a consequence, time-lags frequently occur between the collection of Big Data and official statistics (Christopher Smith-Clarke et al., 2014; Smith et al., 2013; Sultan et al., 2015), baring the risk to further jeopardize the accuracy and reliability of the analysis.

On the other hand, if high-resolution satellite image data are more frequently released, dynamic models can be created which capture changes. These may inform trend analysis and offer regular updates allowing policy makers and stakeholders to monitor and evaluate effects of policies and projects within reasonable time frames (Hall et al., 2001; Henderson et al., 2012; Jean et al., 2016; Kholodilin & Siliverstovs, 2012; Mao et al., 2015; Nischal et al., 2015; Njuguna & McSharry, 2017; Christopher Smith-Clarke et al., 2014; Steele et al., 2017; Sutton & Costanza, 2002). Longitudinal analysis adds possibilities to assess on a regular basis the validity of analytical models (Duque et al., 2015).

4.1.4. Big Data often merge different sources

Merging different data sources requires extensive preproduction treatment as each source might have different characteristics, including for instance different units of analysis, which might affect comparability over time and space (Marchetti et al., 2015). Further considerations may involve questions about the origin of the different data sources. In their study on formal and informal economic activity in Mexico, Ghosh et al. (2009) highlight that population statistics could improve their model. However, as the latter are to a large extent based on official administrative data, they simultaneously point out that this would affect the independence of their estimates.

However, if careful consideration is given, Big Data analytics undoubtedly offer many possibilities to match different data types and sources including various Big Data sources, traditional small data or both. This can help to increase the predictive power of the models constructed which may further increase learning and feedback to policy makers (Dai et al., 2017; Doll et al., 2006; Ebener et al., 2005; Jean et al., 2016; Nischal et al., 2015; Proville et al., 2017; Steele et al., 2017; Sutton, 2007; Wu et al., 2018; Zhao et al., 2017).

4.1.5. The full name of Big Data is Big Analytics

Typically, in Big Data analytics the volume of data drives the choice for an analytical model: more sophisticated theoretical models are preferred for small datasets, simpler analytical models work better with higher data volumes (Hilbert, 2015). Amongst the studies reviewed, some authors recognize that their findings merely involve correlations and are not able to detect causal relationships. This in turn poses challenges for sound knowledge generation (Eagle et al., 2010; Lessmann & Seidel, 2017; Nischal et al., 2015; Zhou et al., 2015). In addition, if the research only focusses on increasing the predictive power of a model, the risk of overfitting is real. The estimates resulting from algorithmic adjustments will be of limited value when the adjustments are driving the improvements of the estimate rather than the underlying data (Mellander et al., 2015).

Other challenges revealed relate to contextual and capacity differences. As the poverty symptoms vary between contexts the possibilities for most models to "travel" is limited (Biggs et al., 2015; Watmough et al., 2016). Big Data also require large storage and processing resources and, not to mention, specialised skills which may incur unforeseen costs (Njuguna & McSharry, 2017).

Besides these challenges, Big Data analytics simultaneously offer many promising opportunities now and for the near future. Strong correlations, despite their limited ability to uncover causality, highlight the potential benefits for targeted policies (Njuguna & McSharry, 2017). Furthermore, this is the first time ever that this kind and volume of data illustrating human activity is available which allows to study social phenomena that are widely accepted in social sciences at population level (Eagle et al., 2010). As discussed in Hilbert (2015, p. 6), high volumes of high-quality data may help to "detect spurious confounding variables and to isolate potential causation mechanisms better than ever before". Coupled with real-time availability, relations can be observed over time, as such allowing to investigate the causal mechanisms underlying the observed correspondence. Combining Big Data and theory-driven model simulation offers additional opportunities to calibrate models, investigate causation and explore scenarios that never existed (Hilbert, 2015).

What's more, the growing availability and capacity of spatial analysis and GIS software allow the spatial properties of socio-economic data to be included in the analysis (Doll et al., 2006; Christopher D Elvidge et al., 2009; Ghosh et al., 2009; Ghosh, Powell, Elvidge, et al., 2010; Njuguna & McSharry, 2017; Steele et al., 2017). GIS software easily combines spatial information from multiple sources for solving complex location-oriented problems potentially spurring new insights about what is occurring where and when in the world. A fuller picture of the spatial distribution of poverty might also provide the foundation for evidence-based strategies to eradicate poverty (Steele et al., 2017). GIS software might as well facilitate accounting for geomorphic features in the analysis when linked to high-resolution sensing data, improving in that way data quality; night-light data for example, is notably impacted by geomorphic features (Dai et al., 2017).

Lastly, as discussed in Hackl (2016) and Elvidge et al. (2012), the exploration of alternative data sources can lead to new and interesting metrics to capture complex socio-economic phenomena and may open up new avenues for research within the socioeconomic research community. Elvidge et al. (2012) for instance argue night-light is a spatially explicit measure of development although capturing other dimensions of development compared to traditional statistics.

4.2. Implications for monitoring and evaluation

The lack of high-quality ground truth data remains a major bottleneck to both training statistical models as well as assessing the quality of the predictions. Jerven (2013) encourages scholars to as critically scrutinize national statistics as research findings. Yet most researchers covered in our review utilised official development statistics within their analysis without further questioning. The use of erroneous data sources to train statistical models or explore relations potentially jeopardises the findings and conclusions of many of the studies reviewed as it carries the risk of both false positives or negatives (type I and II errors) (Jerven, 2013). It puts into perspective the validity of each and every analysis built on official statistics as they are a key factor affecting the analysis (Doll & Pachauri, 2010; Elvidge et al., 2009; Ghosh et al., 2009; Gutierrez et al., 2013; Shi et al., 2014; Chris Smith-Clarke & Capra, 2016; Wang et al., 2012). Put differently, as long as poverty rates or GDP statistics are unreliable, it is difficult to conclude anything from models that only rely on these sources (Gutierrez et al., 2013).

Secondly, as discussed earlier, the papers under review only utilized a limited amount of potential Big Data types. This can be explained by the private nature of many Big Data sources, with commercial and privacy concerns currently leading to anonymized and ad hoc data availability. As long as no protocols are identified that enable data analysis without compromising the privacy of individuals, research will be limited to ad hoc explorations (Blumenstock et al., 2015; Letouzé, 2015). Also for publicly available data sources the ethical, moral, and legal implications are profound (Sutton, 2007). Increased image resolution for example clearly increases the myriad of opportunities for research, but how far can we go without compromising people's privacy? Who owns the data and related information? Big Data can and should be used to empower people. This could start with giving people rights over their own data (Letouzé, 2015). With increased transparency comes increased accountability though; not all parties will welcome such a major power shift by default. Large corporations have their own logic, but also governments and citizens worry about transparency. Measuring economic activity, for instance, can be a sensitive area for governments in rentier states depending on for example oil resources. On the other hand, when there is a lack of mutual trust citizens can fear transparency too (Hilbert, 2015).

Thirdly, the studies reviewed mainly assess the value of Big Data for 'accountability' measuring and monitoring; so far learning is not the main goal. Such a measurement approach supports the assumption "that one can only manage what one can measure, and that measuring goes a long way towards impacting" (Letouzé, 2015, p.2). For knowledge generation however, more adventurous data exploration and modelling are necessary. Notwithstanding, the SDGs are created because of the wide consensus that monitoring has a causal although indirect effect on what is measured (Letouzé, 2015). Big Data clearly offer opportunities to help monitoring the SDGs.

Fourthly, capacity building is key. Integrating Big Data and social sciences does not only require the fusion of data but also of different scientific traditions (Rindfuss & Stern, as cited in Taubenböck et al., 2009, p.1). In a developmental context, ideally a holistic approach is applied which involves all stakeholders and blends academic, institutional and local knowledge. Data science with the aim to inform development policies necessitates an understanding of the local policy context. This signals the importance of contextualisation, offering insights into what the data do, and maybe more importantly, do not tell us. Without these understandings, research risks to only inform the field of data science. The challenge is not only in gaining more data but in gaining data that is relevant to country priorities (Taylor & Schroeder 2014). In addition, as Big Data may alienate researchers from reality there is an emerging need for sound ethical guidance.

Lastly, a quality framework and quality criteria need to be designed for the various Big Data types and their analytics. This will allow the user to assess the relevance of alternative statistics, their accuracy, reliability and comparability (Hackl, 2016). Amongst the authors reviewed, there is little agreement on the minimum quality requirements needed to include Big Data estimates in the regular toolbox for policy making. Some authors are convinced time has come to accept Big Data as a reliable and accurate source for estimating socio-economic indicators (Njuguna & McSharry, 2017) and argue that increasing Big Data quality will be less challenging than improving national statistics (Ebener et al., 2005). Others insist that official, but reliable, statistics are still needed even if it is only to help building more accurate models and assessing their quality (Gutierrez et al., 2013). Regardless of their position however, all authors agree on the opportunities Big Data can offer where no data are available or when the quality is highly disputable. While it is highly unlikely that Big Data are the panacea to solve the "poor data" issue of development statistics, it can play an important role to improve accuracy, timeliness, and relevance of socio-economic statistics in a way that is more cost and labour efficient than expanding traditional data collection (Steve Landefeld, 2014).

5. Conclusion

Development indicators are essential to grasp an accurate picture of a country's development status and are core to monitor progress towards specific development goals including the SDGs. To be of value, such data must be accurate, timely, disaggregated and widely available. This is not the case in many countries of the global south, rather the contrary: they often suffer from poor statistics with data of poor quality produced infrequently or not in time. These poor-quality numbers feed into decision models and have direct impacts on policy; a significant part of development knowledge is generated by poor numbers and histories of development are written based on them.

This scoping review explored the current and potential use of Big Data sources to help solving poor numbers. The review therefore focussed on Big Data in relationship to poverty and economic development (SDG1).

The relevance of the field so far is driven by data availability and appeals to researchers of many different fields as the data types and analytical approaches stem from various research domains. The main sources of Big Data consulted are CDR and satellite imagery data; night-light data is commonly used to capture economic development. The choice for certain data types is based on the hypothesis that the manifestations of poverty and development leave traces that are captured by Big Data sources. With this in mind, the challenges and opportunities of Big Data for development statistics and their implications for monitoring and evaluation are identified. For now, it is highly unlikely that Big Data statistics will replace traditionally generated development statistics any time soon as some major technical challenges and sustainability issues as well as institutional and legal aspects have to be overcome first. In cases though where no data are available or where the quality of the existing data is highly disputable, Big Data can play a role to improve accuracy, timeliness and relevance of socio-economic indicators in a cost and labour efficient way. On a final note, experimentation should be done with care and needs sound ethical guidance avoiding that countries of the global south turn into an experimental kindergarten feeding mainly, northern, data science needs and knowledge building. A holistic approach blending academic, institutional and local knowledge is required to guarantee the best possible outcomes.

References

Anderson, B. C. (2008). The End of Theory : The Data Deluge Makes the Scientific Method Obsolete The End of Theory : The Data Deluge Makes the Scientific Method Obsolete The End of Theory : The Data Deluge Makes the Scientific Method Obsolete, 2–3. Retrieved from https://www.wired.com/2008/06/pb-theory/

Bamberger, M., Raftree, L., & Olazabal, V. (2016). The role of new information and communication technologies in equity-focused evaluation : Opportunities and challenges. *Evaluation, 22*(2), 228–244. https://doi.org/10.1177/1356389016638598

Biggs, T. W., Anderson, W. G., & Pombo, O. A. (2015). Concrete and Poverty, Vegetation and Wealth? A Counterexample from Remote Sensing of Socioeconomic Indicators on the U.S.–Mexico Border. *Professional Geographer, 67*(2), 166–179. https://doi.org/10.1080/00330124.2014.905161

Blazquez, D., & Domenech, J. (2018). (BD1) Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change, 130*(July 2017), 99–113. https://doi.org/10.1016/j.techfore.2017.07.027

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science, 350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Boyd, D. S., Jackson, B., Wardlaw, J., Foody, G. M., Marsh, S., & Bales, K. (2018). Slavery from Space: Demonstrating the role for satellite remote sensing to inform evidence-based action related to UN SDG number 8. *ISPRS Journal of Photogrammetry and Remote Sensing, 142*, 380–388. https://doi.org/10.1016/j.isprsjprs.2018.02.012

Chen, H., & Storey, V. C. (2012). (12) B Usiness I Ntelligence and a Nalytics : F Rom B Ig D Ata To B Ig I Mpact. *Mis Quarterly, 36*(4), 1165–1188. https://doi.org/10.1145/2463676.2463712

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences, 108*(21), 8589–8594. https://doi.org/10.1073/pnas.1017031108

Dai, Z., Hu, Y., & Zhao, G. (2017). The Suitability of Different Nighttime Light Data for GDP Estimation at Different Spatial Scales and Regional Levels. *Sustainability, 9*(2), 305. https://doi.org/10.3390/su9020305

data for development. (n.d.). Retrieved August 15, 2018, from http://www.d4d.orange.com/en/Accueil

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review, 65*(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061

Doll, C. N. H., Muller, J.-P., & Morley, J. G. (2006). Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics, 57*(1), 75–92. https://doi.org/10.1016/j.ecolecon.2005.03.007

Doll, C. N. H., & Pachauri, S. (2010). Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery. *Energy Policy, 38*(10), 5661–5670. https://doi.org/10.1016/j.enpol.2010.05.014

Duque, J. C., Patino, J. E., Ruiz, L. A., & Pardo-Pascual, J. E. (2015). Measuring intra-urban poverty using land cover and texture metrics derived from remote sensing data. *Landscape and Urban Planning, 135*, 11–21. https://doi.org/10.1016/j.landurbplan.2014.11.009

Eagle, N., Macy, M., & Claxton, R. (2010). Network Diversity and Economic Development. *Science, 328*(5981), 1029–1031. https://doi.org/10.1126/science.1186605

Ebener, S., Murray, C., Tandon, A., & Elvidge, C. C. (2005). From wealth to health: Modelling the distribution of income per capita at the sub-national level using night-time light imagery. *International Journal of Health Geographics, 4*. https://doi.org/10.1186/1476-072X-4-5

Elvidge, C. D., Baugh, K. E., Anderson, S. J., Sutton, P. C., & Ghosh, T. (2012). The Night Light Development Index (NLDI): A spatially explicit measure of human development from satellite data. *Social Geography*, *7*(1), 23–35. https://doi.org/10.5194/sg-7-23-2012

Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., Davis, E. R., & Davis, C. W. (1997). Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing*, *18*(6), 1373–1379. https://doi.org/10.1080/014311697218485

Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B., & Bright, E. (2009). A global poverty map derived from satellite data. *Computers and Geosciences*, *35*(8), 1652–1660. https://doi.org/10.1016/j.cageo.2009.01.009

Frias-Martinez, V., & Virseda, J. (2012). On the relationship between socio-economic factors and cell phone usage. *ACM International Conference Proceeding Series*, 76–84. https://doi.org/10.1145/2160673.2160684

Ghosh, T., Anderson, S., Powell, R. L., Sutton, P. C., & Elvidge, C. D. (2009). Estimation of Mexico's informal economy and remittances using nighttime imagery. *Remote Sensing*, *1*(3), 418–444. https://doi.org/10.3390/rs1030418

Ghosh, T., Powell, R. L., Anderson, S., Sutton, P. C., & Elvidge, C. D. (2010). Informal Economy and Remittance Estimates of India Using Nighttime Imagery. *International Journal of Ecological Economics & Statistics (IJEES) Spring Int. J. Ecol. Econ. Stat*, *17*(P10), 973–1385. Retrieved from www.ceser.res.in/ijees.htmlhttp://www.ceserp.com/cp-jour

Ghosh, T., Powell, R. L., Elvidge, C. D., Baugh, K. E., Sutton, P. C., & Anderson, S. (2010). Shedding Light on the Global Distribution of Economic Activity. *The Open Geography Journal*, *3*, 148–161. https://doi.org/10.2174/1874923201003010147

Gutierrez, T., Krings, G., & Blondel, V. D. (2013). Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *ArXiv:1309.4496 [Physics]*, 1–6. Retrieved from https://arxiv.org/pdf/1309.4496.pdf

Hackl, P. (2016). Big Data: What can official statistics expect? *Statistical Journal of the IAOS*, *32*(1), 42–52. https://doi.org/10.3233/SJI-160965

Hall, G. B., Malcolm, N. W., & Piwowar, J. M. (2001). Integration of remote sensing and GIS to detect pockets of urban poverty: The case of Rosario, Argentina. *Transactions in GIS*, *5*(3), 235–253. https://doi.org/10.1111/1467-9671.00080

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*. https://doi.org/10.1257/aer.102.2.994

Hilbert, M. (2015). Big Data for Development: A Review of Promises and Challenges. *Ssrn*, *34*(1), 135–174. https://doi.org/10.1111/dpr.12142

Home | LandScan™. (n.d.). Retrieved August 21, 2018, from https://landscan.ornl.gov/

Huang, Q., Yang, X., Gao, B., Yang, Y., & Zhao, Y. (2014). (S1) Application of DMSP/OLS nighttime light images: A meta-analysis and a systematic literature review. *Remote Sensing*, *6*(8), 6844–6866. https://doi.org/10.3390/rs6086844

Imran, M., Stein, A., & Zurita-Milla, R. (2014). Investigating rural poverty and marginality in Burkina Faso using remote sensing-based products. *International Journal of Applied Earth Observation and Geoinformation*, *26*(1), 322–334. https://doi.org/10.1016/j.jag.2013.08.012

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

Jerven, M. (2013). Poor numbers : how we are misled by African development statistics and what to do about it. Claremont, South Africa: UCT Press.

Kholodilin, K. A., & Siliverstovs, B. (2012). Measuring regional inequality by internet car price advertisements: Evidence for Germany. *Economics Letters*, *116*(3), 414–417. https://doi.org/10.1016/j.econlet.2012.04.039

Kimijiama, S., & Nagai, M. (2014). Study for urbanization corresponding to socio-economic activities in Savannaket, Laos using satellite remote sensing. In *IOP Conference Series: Earth and Environmental Science* (Vol. 20). https://doi.org/10.1088/1755-1315/20/1/012005

Kuffer, M., Pfeffer, K., & Sliuzas, R. (2016). (14) Slums from space-15 years of slum mapping using remote sensing. *Remote Sensing*, *8*(6). https://doi.org/10.3390/rs8060455

Lessmann, C., & Seidel, A. (2017). Regional inequality, convergence, and its determinants – A view from outer space. *European Economic Review*, *92*, 110–132. https://doi.org/10.1016/j.euroecorev.2016.11.009

Letouzé, E. (2015). Thoughts on Big Data and the SDGs. Retrieved from https://sustainabledevelopment.un.org/content/documents/7798BigData - Data-Pop Alliance - Emmanuel Letouze.pdf

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). (2) Scoping studies: Advancing the methodology. *Implementation Science*, *5*(1). https://doi.org/10.1186/1748-5908-5-69

Levin, N., & Duke, Y. (2012). High spatial resolution night-time light images for demographic and socio-economic studies. *Remote Sensing of Environment*, *119*, 1–10. https://doi.org/10.1016/j.rse.2011.12.005

Li, X., Xu, H., Chen, X., & Li, C. (2013). Potential of NPP-VIIRS nighttime light imagery for modeling the regional economy of China. *Remote Sensing*, *5*(6), 3057–3081. https://doi.org/10.3390/rs5063057

Mao, H., Shuai, X., Ahn, Y. Y., & Bollen, J. (2015). Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to Côte d'Ivoire. *EPJ Data Science*, *4*(1), 1–16. https://doi.org/10.1140/epjds/s13688-015-0053-1

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., … Gabrielli, L. (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics*, *31*(2), 263–281. https://doi.org/10.1515/JOS-2015-0017

Mellander, C., Lobo, J., Stolarick, K., & Matheson, Z. (2015). Night-time light data: A good proxy measure for economic activity? *PLoS ONE*, *10*(10). https://doi.org/10.1371/journal.pone.0139779

Morikawa, R. (2014). Remote sensing tools for evaluating poverty alleviation projects: A case study in Tanzania. In *Procedia Engineering* (Vol. 78, pp. 178–187). https://doi.org/10.1016/j.proeng.2014.07.055

Nischal, K. N., Radhakrishnan, R., Mehta, S., & Chandani, S. (2015). Correlating night-time satellite images with poverty and other census data of India and estimating future trends. In *Proceedings of the Second ACM IKDD Conference on Data Sciences - CoDS '15* (pp. 75–79). https://doi.org/10.1145/2732587.2732597

Njuguna, C., & McSharry, P. (2017). Constructing spatiotemporal poverty indices from big data. *Journal of Business Research*, *70*, 318–327. https://doi.org/10.1016/j.jbusres.2016.08.005

Noor, A. M., Alegana, V. A., Gething, P. W., Tatem, A. J., & Snow, R. W. (2008). Using remotely sensed night-time light as a proxy for poverty in Africa. *Population Health Metrics*, 6. https://doi.org/10.1186/1478-7954-6-5

Pham, M. T., Rajić, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., & Mcewen, S. A. (2014). (1) A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. *Research Synthesis Methods*, *5*(4), 371–385. https://doi.org/10.1002/jrsm.1123

Pokhriyal, N., Dong, W., & Govindaraju, V. (2015). Virtual Networks and Poverty Analysis in Senegal. *ArXiv:1506.03401v1 [Cs.CY]*. Retrieved from https://www.cia.gov/library/

Proville, J., Zavala-Araiza, D., & Wagner, G.

(2017). Night-time lights: A global, long term look at links to socio-economic trends. *PLoS ONE, 12*(3). https://doi.org/10.1371/journal.pone.0174610

Ruggeri Laderchi, C., Saith, R., & Stewart, F. (2003). Does it matter that we do not agree on the definition of poverty? A comparison of four approaches. Oxford Development Studies (Vol. 31). https://doi.org/10.1080/1360081032000111698

Šćepanović, S., Mishkovski, I., Hui, P., Nurminen, J. K., & Ylä-Jääski, A. (2015). Mobile phone call data as a regional socio-economic proxy indicator. *PLoS ONE, 10*(4). https://doi.org/10.1371/journal.pone.0124160

Sedda, L., Tatema, A. J., Morley, D. W., Atkinson, P. M., Wardrop, N. A., Pezzulo, C., … Rogers, D. J. (2015). Poverty, health and satellite-derived vegetation indices: Their inter-spatial relationship in West Africa. *International Health*, *7*(2), 99–106. https://doi.org/10.1093/inthealth/ihv005

Shi, K., Yu, B., Huang, Y., Hu, Y., Yin, B., Chen, Z., … Wu, J. (2014). Evaluating the ability of NPP-VIIRS nighttime light data to estimate the gross domestic product and the electric power consumption of China at multiple scales: A comparison with DMSP-OLS data. *Remote Sensing*, *6*(2), 1705–1724. https://doi.org/10.3390/rs6021705

Smith-Clarke, C., & Capra, L. (2016). Beyond the Baseline. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16* (pp. 425–434). https://doi.org/10.1145/2872427.2883076

Smith-Clarke, C., Mashhadi, A., & Capra, L. (2014). Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 511–520). https://doi.org/10.1145/2556288.2557358

Smith, C., Mashhadi, A., & Capra, L. (2013). Ubiquitous sensing for mapping poverty in developing countries. In *Paper presented at 3rd International Conference on the Analysis of Mobile Phone Datasets*. https://doi.org/http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.408.9095

Sonka, S. T. (2016). Big Data: Fueling the Next Evolution of Agricultural Innovation. *Journal of Innovation Management*, *4*(1), 114–136. https://doi.org/hdl.handle.net/10216/83250

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., … Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface*, *14*(127). https://doi.org/10.1098/rsif.2016.0690

Steve Landefeld. (2014). Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues, (October 2014), 20. Retrieved from https://unstats.un.org/unsd/trade/events/2014/beijing/Steve Landefeld - Uses of Big Data for official statistics.pdf

Sultan, S. F., Humayun, H., Nadeem, U., Bhatti, Z. K., Khan, S., & Ali, S. B. (2015). Mobile Phone Price as a Proxy for Socio-Economic Indicators. In *ICTD'15 Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*. https://doi.org/10.1145/2737856.2737892

Sutton, P. C. (2007). Estimation of Gross Domestic Product at Sub-National Scales Using Nighttime Satellite Imagery GEO Global Urban Observation and Information (GI-17): Monitoring Urban Assets in Support of Sustainable Cities View project Nighttime lights and economic activit. *International Journal of Ecological Economics & Statistics (IJEES) Spring Int. J. Ecol. Econ. Stat*. Retrieved from https://www.researchgate.net/publication/242254394

Sutton, P. C., & Costanza, R. (2002). Global estimates of market and non-market values derived from nighttime satellite imagery, land cover, and ecosystem service valuation. *Ecological Economics*, *41*(3), 509–527. https://

doi.org/10.1016/S0921-8009(02)00097-6

Taubenböck, H., Wurm, M., Setiadi, N., Gebert, N., Roth, A., Strunz, G., … Dech, S. (2009). Integrating remote sensing and social science: The correlation of urban morphology with socioeconomic parameters. In *2009 Joint Urban Remote Sensing Event*. https://doi.org/10.1109/URS.2009.5137506

Taylor, L., & Broeders, D. (2015). In the name of Development: Power, profit and the datafication of the global South. *Geoforum*, *64*, 229–237. https://doi.org/10.1016/j.geoforum.2015.07.002

Taylor, L., & Schroeder, R. (2015). Is bigger better? The emergence of big data as a tool for international development policy. *GeoJournal*, *80*(4), 503–518. https://doi.org/10.1007/s10708-014-9603-5

Walasek, L., & Brown, G. D. A. (2016). Income Inequality, Income, and Internet Searches for Status Goods: A Cross-National Study of the Association Between Inequality and Well-Being. *Social Indicators Research*, *129*(3), 1001–1014. https://doi.org/10.1007/s11205-015-1158-4

Wang, W., Cheng, H., & Zhang, L. (2012). Poverty assessment using DMSP/OLS night-time light satellite imagery at a provincial scale in China. *Advances in Space Research*, *49*(8), 1253–1264. https://doi.org/10.1016/j.asr.2012.01.025

Watmough, G. R., Atkinson, P. M., & Hutton, C. W. (2013). Predicting socioeconomic conditions from satellite sensor data in rural developing countries: A case study using female literacy in Assam, India. *Applied Geography*, *44*, 192–200. https://doi.org/10.1016/j.apgeog.2013.07.023

Watmough, G. R., Atkinson, P. M., Saikia, A., & Hutton, C. W. (2016). Understanding the Evidence Base for Poverty-Environment Relationships using Remotely Sensed Satellite Data: An Example from Assam, India. *World Development*, *78*, 188–203. https://doi.org/10.1016/j.worlddev.2015.10.031

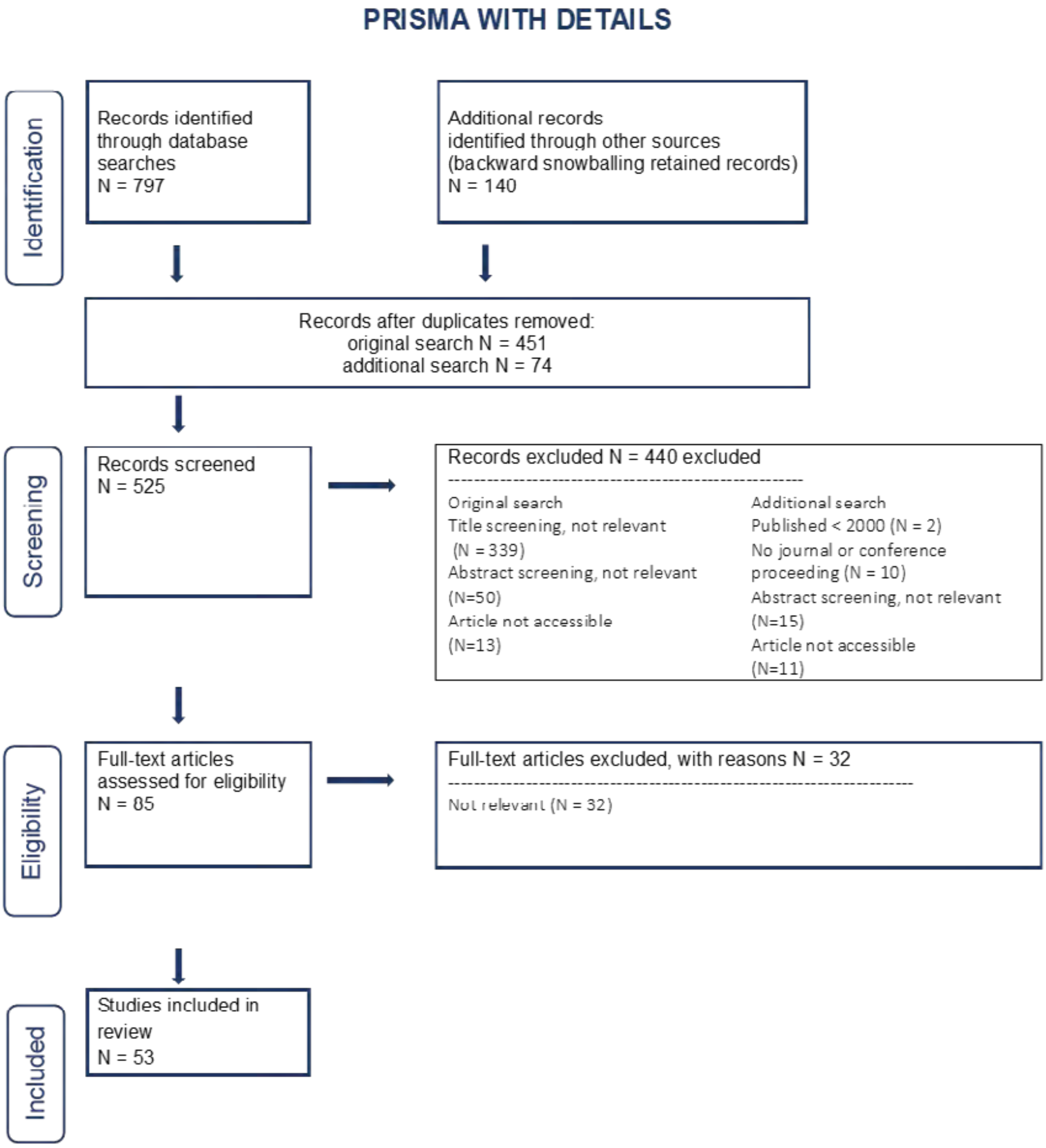
Wu, R., Yang, D., Dong, J., Zhang, L., & Xia, F. (2018). Regional inequality in China based on NPP-VIIRS night-time light imagery. *Remote Sensing*, *10*(2). https://doi.org/10.3390/rs10020240

Xu, H., Yang, H., Li, X., Jin, H., & Li, D. (2015). Multi-scale measurement of regional inequality in Mainland China during 2005-2010 using DMSP/OLS night light imagery and population density grid data. *Sustainability (Switzerland)*, *7*(10), 13469–13499. https://doi.org/10.3390/su71013469

Zhao, M., Cheng, W., Zhou, C., Li, M., Wang, N., & Liu, Q. (2017). GDP spatialization and economic differences in South China based on NPP-VIIRS nighttime light imagery. *Remote Sensing*, *9*(7). https://doi.org/10.3390/rs9070673

Zhou, Y., Ma, T., Zhou, C., & Xu, T. (2015). Nighttime light derived assessment of regional inequality of socioeconomic development in China. *Remote Sensing*, *7*(2), 1242–1262. https://doi.org/10.3390/rs70201242

Appendix A: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram


PRISMA WITH DETAILS
Identification
Records identified through database searches N = 797
Additional records identified through other sources (backward snowballing retained records) N = 140
Records after duplicates removed: original search N = 451 additional search N = 74
Screening
Records screened N = 525
Records excluded N = 440 excluded
Original search
Title screening, not relevant (N = 339)
Abstract screening, not relevant (N=50)
Article not accessible (N=13)
Additional search
Published < 2000 (N = 2)
No journal or conference proceeding (N = 10)
Abstract screening, not relevant (N=15)
Article not accessible (N=11)
Eligibility
Full-text articles assessed for eligibility N = 85
Full-text articles excluded, with reasons N = 32
Not relevant (N = 32)
Included
Studies included in review N = 53

Annex B: Table: List of references employing big data for measuring concepts related to SDG1

Reference	Analytical methodology	Geography	Concept	Primary Big Data source	Abailibility primary Big Data	Other data	Big-data feature	Ground thruth data	Ground thruth data used for training?	Quality of the analytical model	Quality measured against ground thruth data?	Alternative for ground truth data
(Boyd et al., 2018)	Crowd computing	Region (Brick Belt, India)	Poverty	Google earth (2014,2016)	private - free available		Kilns seen, manually indicated by crowdsourcing	Expert visualisation	not trained	Expert visualisation 99.6% accuracy compared to crowd	no	Collect ground-truth
(Njuguna & McSharry, 2017)	General linear model (GLM)	Country (Rwanda)	Poverty	1) MNO (mobile network operator) 2) NPP-VIIRS (Dec 2014)	1) private 2) public	Landscan	1) Call activity or consumption, ownership 2) Night-light, light intensity 3) Population density	Oxford Poverty & Human Development Initiative (2004-2010); multi-dimensional poverty index (2013, data collected 2012)	trained	R² =76%	yes	
(Watmough et al., 2016)	Geostatistical model (randomised forest)	Regional (Assam, India)	Rural poverty	1) Landsat (2001-2002) 2) MODIS (2001)	public		1) Environmental morphology 2) NDVI	Census (2001); poverty index	trained	Realtively good results for poorest and richest: Accuracy of 61% of for lowes income classes and 57% for highes income classes Weak accuracy for middle income classes Overall accuray of 36%	yes	
(Wang et al., 2012)	General linear model (GLM)	Country (China)	Poverty	DMSP-OLS (2007-2009)	public		Light intensity	National Bureau of Statistics of China (2007-2009), socio-economic data	trained	R² =85%	yes	
(Duque et al., 2015)	GLM / spatial properties	Regional (Medellin, Colombia)	Urban poverty	Quickbird, VHR imagery (2008)	public		Physical urban morphology	Quality of Life Survey of Medellin; slum Index (2007)	trained	GLM: R² =0.59 GLM, including spatial properties: R² =0.62	yes	
(Imran et al., 2014)	GLM / spatial properties	Country (Burkina Faso)	Rural poverty	1) Spot vegetation NDVI (2009) 2) TAMSAT (2009) 3) FAO (2012) and HYDRO1k data sets of (USGS, 2012).	public		1) Normalized difference vegetation index (NDVI) 2) Climatic stress 3) Other agro-ecological gridded data as potential stressors on food production	AGRISTAT (2010); asset index	trained	RMSE (root mean square error) = 0.17	yes	
(Morikawa, 2014)	Trend analysis	Regional (Pangani, Tanzania)	Rural poverty	MODIS (2003-2013)	public		NDVI	1) Census (2002) 2) GPS waypoints and field reports (Floresta Tanzania)	not trained	Poor results, directional only	no	report NDVI as outcome
(Levin & Duke, 2012)	Correlation analysis, GLM	Country (Israel, including Palestinian Authority)	Socio-economic development	1) DMSP-OLS (2003) 2) Argentinean SAC-C (2007) 3) ISS (2003) 4) Landsat (2003)	public		1), 2) and 3) Night-light, , feature not specified 4) Environmental morphology	Isrelian Central Bureau of Statistics Palestinian Central Bureau of Statistics	not trained	Strong indication NTL may inform on economic activity at the local scale, and not only on regional and national scale as done so far.	yes	

Reference	Analytical methodology	Geography	Concept	Primary Big Data source	Abailibility primary Big Data	Other data	Big-data feature	Ground thruth data	Ground thruth data used for training?	Quality of the analytical model	Quality measured against ground thruth data?	Alternative for ground truth data
(Doll & Pachauri, 2010)	Index calculation	Global	Rural poverty	DMSP-OLS (1992-2000)	public	Population of the World (GPW) (1990,2000) Global Rural Urban Mapping Project (GRUMP) (2000, 2004)	Light present or not	World Energy Outlook 2002 (IEA, 2002) for energy estimation	not trained	Poor results, considerable overestimation of population without electicity	yes	
(Elvidge et al., 2009)	General linear model (GLM)	Global	Poverty	DMSP-OLS (2003)	public	Landscan (2004)	Unspecified or unclear	World Development Indicators (2006), poverty index	trained	Indices correlate strongly with other independently derived measures of poverty on national and sub-national level (not specified).	yes	
(Mao et al., 2015)	Correlation analysis	Country (Yvory Coast)	Socio-economic development	Orange - Data for Development (D4D) Challenge	private		Call activity or consumption, Introversion, Network advantage	IMF (2009), Country report 09/156, total average annual per capita income and poverty rate, Gini index and the ratio of average income in urban areas to average income in rural areas (i.e. U/R ratio)	not trained	R=0.80 intitiating calls and income. R=0.83 intitiating calls and poverty rate. Computed 'callrank' reflects economic importance of a region.	yes	
(Steele et al., 2017)	GLM / spatial properties	Country (Bangladesh)	Poverty	1) CDR GP subscribers (2013-2014) 2) 25 raster and vector datasets from existing sources, daylight and night-light (details not provided)	1) private 2) public		1) Basic available CDR features, not specified further 2) Environmental and physical metrics	Bangladesh DHS (2011), FII survey (2014), national household surveys conducted by Telenor Group (2013-2014)	trained	Best results for predicting asset-based wealth index: R^2 =78% for urban poverty R^2 =66% for rural poverty R^2 =76% at the national level Consumption-based and income-based poverty proved more elusive.	no	in study data collection for predictor
(Blumenstock et al., 2015)	Machine learning	Country (Rwanda)	Poverty	Primary mobile phone operator in Rwanda	private		Feature generation several thousand metrics whereafter elimation through "elastic net" regularization	Follow-up phone surveys, wealth index calculated (PCA)	trained	R^2=0.68 on individual level R^2=0.79 on village level R^2=92 on district level	no	in study data collection for predictor
(Jean et al., 2016)	General linear model (GLM)	Multi-country (Africa)	Poverty	1) Google Static Maps API (year not specified) 2) DMSP-OLS (2010)	public		1) Environmental morphology 2) Night-light, feature not specified	World Bank's Living Standards Measurement Study (LSMS) (2000-2010)	trained	R^2 =37%-55% for household consumption R^2 =55%-75% for household asset wealth	yes	
(Šćepanović et al., 2015)	Correlation analysis	Country (Yvory Coast)	Socio-economic development	Orange - Data for Development (D4D) Challenge	private		Mobility patterns	Multidimensional Poverty Index (MPI)	not trained	R=-0.77 commuting patterns R=-0.74 number of calls	yes	

Reference	Analytical methodology	Geography	Concept	Primary Big Data source	Abailibility primary Big Data	Other data	Big-data feature	Ground thruth data	Ground thruth data used for training?	Quality of the analytical model	Quality measured against ground thruth data?	Alternative for ground truth data
(Sedda et al., 2015)	Geostatistical model	Multi-country (West-Africa)	Rural poverty	MODIS (2001-2010)	public		NDVI, day- and night-time land surface temperature, elevation	Oxford Poverty & Human Development Initiative (2004-2010); intensity of poverty	trained	Strongest correlation with NDVI: R=-0.62 for intensity of poverty R=-0.59 for MPI R=-0.54 for poverty headcount Elevation $R < 0.1$ Rainfall $R < 0.5$	yes	
(Hall et al., 2001)	Index calculation (supervised classification)	Regional (Rosario, Argentina)	Urban poverty	1) Radarsat-1 (1997) 2) Landsat TM (1984, 1995)	public		Physical urban morphology	Census (1991); poverty index	trained	Generally encouraging (not specified)	yes	
(Noor et al., 2008)	Correlation analysis, GLM	Multi-country (Africa)	Poverty	DMSP-OLS (2000)	public	Global Rural Urban Mapping Project (GRUMP)	Area (un)lit (%), Light Intensity, Distance to nearest night-time light	Multiple Indicators Cluster Surveys (MICS) and Demographic and Health Surveys (DHS) (surveys closest to 2000 selected), asset based wealth index	not trained	Light intensity strongest correlation (R=0.64), Area (un)lit (%) R=0.63 and Distance to nearest night-time light (R=-0.61)	yes	
(Proville et al., 2017)	Correlation analysis, GLM	Global	Socio-economic development	DMSP-OLS (1992-2013)	public	United Nations, Eurostat and national census Energy Information Administration, CDIAC, EDGAR	Area (un)lit (%)	Worldbank data for GDP and poverty headcount	trained	Maximum correlation of 0.93 with GDP and electricity, to r=-057 with poverty headcount. Multiple amount of correlations with a collection of socio-economic indicators.	yes	
(Biggs et al., 2015)	GLM / spatial properties	Regional (Tijuana, Mexico)	Socio-economic development	Landsat TM (2003)	public		Physical urban morphology, land cover	Census (2000); socioeconomic marginality index (MI)	trained	R^2 =0.37	yes	
(Kimijiama & Nagai, 2014)	Correlation analysis, GLM	Country (Laos)	Socio-economic development	Landsat (1973, 2000, 2001, 2002, 2003, 2005, 2006, 2007, 2009, 2010, 2012, 2013)	public		Degree of urbanisation	Asian Development Bank; socio-economic indicators (years available differ by indicator, all between 1960-2011)	not trained	Very high correlation for school enrolment (R=0.97) to very low for literacy rate amongst youth (R=0.01)	yes	

Reference	Analytical methodology	Geography	Concept	Primary Big Data source	Abailibility primary Big Data	Other data	Big-data feature	Ground thruth data	Ground thruth data used for training?	Quality of the analytical model	Quality measured against ground thruth data?	Alternative for ground truth data
(Chris Smith-Clarke & Capra, 2016)	General linear model (GLM)	Multi Country (Yvory Coast, Senegal)	Poverty	Orange - Data for Development (D4D) Challenge	private		Call activity or consumption, Introversion, Network advantage, Gravity residuals	Demographic and Health Surveys (DHS), wealth and poverty index	trained	Correlation with best model: Senegal Correlation with wealth: no CDR R: 0.75-0.80, CDR only R: 0.75-0.80, CDR + population + time lag R: +0.80 Correlation with poverty: no CDR R: 0.75-0.80, CDR only R: 0.70-075, CDR + population + time lag R: 0.75-0.80 (but better than no CDR) Yvory Coast Correlation with wealth: no CDR R: 0.70-0.75, CDR only R: 0.65-0.70, CDR + population + time lag R: +0.75 Correlation with poverty: no CDR R: 0.70-0.75, CDR only R: 0.65-070, CDR + population + time lag R: 0.70-0.75 (but better than no CDR)	yes	
(Sultan et al., 2015)	Correlation analysis	Country (Pakistan)	Socio-economic development	Mobilink (mobile network operator)	private	Scraping data; price of mobile phone	Phone model	Socio-economic variables from Mauza Census (2008), Population Census (1998)	not trained	Highest correlation reported = 0.62	yes	
(Nischal et al., 2015)	Correlation analysis, GLM, Trend analysis	Country (India)	Socio-economic development	DMSP-OLS (2000-2012)	public	Census	Light intensity	Census	trained	Big differences between the regions, adjusted R² ranging between 0.25 - 0.69.	yes	
(Christopher Smith-Clarke et al., 2014)	Correlation analysis, General linear model (GLM)	Country (Yvory Coast)	Poverty	Orange - Data for Development (D4D) Challenge	private		Call activity or consumption, Introversion, Network advantage, Gravity residuals	Poverty rate, IMF (2008)	not trained	All covariates correlate with poverty, rang R=-0.43 - -0.83	yes	
(Walasek & Brown, 2016)	Correlation analysis	Global Country (USA)	Inequality	Google trends (2009-2014)	public		Searches for status goods	International Database of Income Inequality (GINI, 2009) World Development Indicators (Income, 2009) U.S. Census Bureau (to compute GINI 2010, 2012)	not trained	Internet search terms related to positional goods are relatively more frequent in regions with higher levels of income inequality (multiple correlations reported)	yes	

Reference	**Analytical methodology**	**Geography**	**Concept**	**Primary Big Data source**	**Abailibility primary Big Data**	**Other data**	**Big-data feature**	**Ground thruth data**	**Ground thruth data used for training?**	**Quality of the analytical model**	**Quality measured against ground thruth data?**	**Alternative for ground truth data**
(Kholodilin & Siliverstovs, 2012)	Correlation analysis	Country (Germany)	Inequality	Scraping data (http://www.mobile.de) (May 2010)	public		Car prices	Statistisches Bundesamt (GINI, 2010)	not trained	R=-0.77 commuting patterns R=-0.74 number of calls	yes	
(Wu et al., 2018)	General linear model (GLM)	Country (China)	Economic development, Inequality	NPP-VIIRS (2014-2017)	public		Unspecified or unclear	China Statistical Yearbook for regional economy (2013–2014), GDP & population Regional Statistical Yearbooks (2015-2016), GDP & population	trained	R^2 =85% provincial R^2 =86% prefecture	yes	
(Sutton, 2007)	General linear model (GLM)	Country (China, India, Turkey, US)	Economic development	DMSP-OLS (1992-1993,2000)	public	Landscan (2000)	Area (un)lit (%), Sum of lights	GDP of nations (source not specified)	trained	China: R^2 =94% simple model - R^2 =96% urban model India: R^2 =70% simple model - R^2 =84% urban model Turkey: R^2 =58% simple model - R^2 =94% urban model USA: R^2 =70% simple model - R^2 =72% urban model	yes	
(Ghosh et al., 2009)	General linear model (GLM)	Country (Mexico)	Informal economic activity	DMSP-OLS (2000-2001)	public	Landscan (2000)	Area (un)lit (%), Population in lit area	U.S. Bureau of Economic Analysis (2000), GDP US World Development Report (2002), GDP and GNI US INEGI (2000), GDP and GNI Mexico	not trained	Model predicting state GDP in USA transferred to Mexico for determining the magnitude of underestimation of informal economy and remittances in the official estimates of GNI of Mexico. % residual by American state ranges from -205 (maximum deviation) to 4 (minimum deviation).	no	transfer model based on US
(Gutierrez et al., 2013)	Index calculation	Country (Yvory Coast)	Socio-economic development	A mobile phone operator in Yvory Coast	private		Call activity or consumption	No ground thruth data	trained	No model, not evaluated by comparing with ground thruth data	no	use wealth index computed on CDR
(Shi et al., 2014)	General linear model (GLM)	Country (China)	Economic development	1) DMSP-OLS (2012) 2) NPP-VIIRS (2012)	public		Sum of lights	GDP (2011)	trained	Provincial level R^2 =87% for VIIRS, R^2 =73% OLS 61% of provinces high accuracy for GDP estimate (max error 30%)	yes	

Reference	Analytical methodology	Geography	Concept	Primary Big Data source	Abailibility primary Big Data	Other data	Big-data feature	Ground thruth data	Ground thruth data used for training?	Quality of the analytical model	Quality measured against ground thruth data?	Alternative for ground truth data
(Ebener et al., 2005)	General linear model (GLM)	Global	Economic development	DMSP-OLS (1994-1995)	public		Multiple features (logs)	Country specific administrative data (1991-1997)	trained	Different models by groups of countries. Grouping on % of GDP from agriculture: below 5% Adj. R²=0.58 between 5-10% Adj. R²=0.63 between 10-25% Adj. R²=0.55 above 25% Adj. R²=55.	yes	
(Zhao et al., 2017)	General linear model (GLM)	Region (South-China)	Economic development	1) NPP-VIIRS (2014) 2) DMSP-OLS (2012)	public	Landsat	Sum of lights, Light intensity, Night time light index	Regional Statistical Yearbooks (2015), GDP	trained	R² =89% prefecture R² =92% county	yes	
(Sutton & Costanza, 2002)	Index calculation	Global	Economic development; Market and non-market valued	1) DMSP-OLS (night-light) (1995-1996); for GDP estimate 2) IGBP global land-cover dataset (Belward, 1996) and corresponding ecosystem service value (Costanza etal., 1997); for ESP estimate	public		1) Night-light, sum of lights 2) Land coverage	NA	trained	NA; output is a map and country ranking not checked against ground thruth data	yes	
(Ghosh, Powell, Anderson, et al., 2010)	General linear model (GLM)	Country (India)	Informal economic activity	DMSP-OLS (2000-2001)	public	Landscan (2000)	Area (un)lit (%), Population in lit area	U.S. Bureau of Economic Analysis (2000), GDP US World Development Report (2002), GDP and GNI US CSO (2000), GDP and GNI India	not trained	Model predicting state GDP in USA transferred to Mexico for determining the magnitude of underestimation of informal economy and remittances in the official estimates of GNI of Mexico. % residual by American state ranges from -205 (maximum deviation) to 4 (minimum deviation).	no	transfer model based on US
(Taubenböck et al., 2009)	Correlation analysis	Regional (Padang, Indonesia)	Socio-economic development	Ikonos (year not specified)	public		Physical urban morphology	In study data collection	trained	Encouraging, clear but low direct correlation (R=0.21), adding qualitative information increases comprehensiveness of the analysis	no	in study data collection for predictor
(Doll et al., 2006)	GLM / spatial properties	Multi-country (USA, Europe)	Economic development	DMSP-OLS (1996-1997)	public		Area (un)lit (%)	European System of Integrated Economic Accounting (ESA) US Bureau of Economic Analysis (BEA), GDP (2002-2003)	trained	Different models by country. R² range from 0.99 for Portugal to 0.75 for France. Additional analysis of geographically disaggregated GRP estimates provided.	yes	

Reference	Analytical methodology	Geography	Concept	Primary Big Data source	Abailibility primary Big Data	Other data	Big-data feature	Ground thruth data	Ground thruth data used for training?	Quality of the analytical model	Quality measured against ground thruth data?	Alternative for ground truth data
(Henderson et al., 2012)	General linear model (GLM)	Global	Formal + Informal economic activity (growth)	DMSP-OLS (1992-2008)	public		Area (un)lit (%), Light intensity, Frequency saturated pixels	World Development Indicators (1992-2006)	trained	For countries with high-quality national accounts data, the information contained in lights growth is of little value in improving income growth measures. For countries with low-quality national accounts data, however, the optimal composite estimate puts roughly equal weight on lights growth and national accounts data. For these countries, the estimated set of income growth numbers for the years 1992/3–2005/6 differ from measured WDI real GDP growth numbers by up to 3.2 percent per year.	no	focus on growth
(Xu et al., 2015)	Index calculation	Country (China)	Inequality	DMSP-OLS (2004.2006,2010)	public	Global Change Research Data Publisher & Repository (2005,2010)	Sum of lights	NA	not trained	NA, no benchmark with ground thruth data	no	compute new indicator and explore
(Eagle et al., 2010)	Correlation analysis	Country (UK)	Socio-economic development	The most complete record of a national communication network covering 90% of mobile phones and greater than 99% of residential and business landlines	private		Call activity or consumption, Social network variables, including diversity	UK government's Index ofMul- tiple Deprivation (IMD) (2004)	not trained	R=0.73 with social diversity R=0.58 with spatial network diversity	yes	
(Mellander et al., 2015)	Correlation analysis, GLM / spatial properties	Country (Sweden)	Economic development	DMSP-OLS (2006)	public		Light Intensity	Statistics Sweden demographic data (year not specified)	trained	NTL captures density better than total count values (R approximately 0.7) and relationship between NTL and local economic activity is not as strong as with population and population density. Results clearly show that the saturation problem is real, which can lead to underestimating the activity associated with larger, brighter cities and regions.	yes	

Reference	Analytical methodology	Geography	Concept	Primary Big Data source	Abailibility primary Big Data	Other data	Big-data feature	Ground thruth data	Ground thruth data used for training?	Quality of the analytical model	Quality measured against ground thruth data?	Alternative for ground truth data
(Zhou et al., 2015)	General linear model (GLM), Index calculation	Country (China)	Inequality	NPP-VIIRS (2012)	public	Census (2010)	Light intensity	Statistical yearbook of urban areas GDP (2012)	trained	R^2 =86%	yes	
(Frias-Martinez & Virseda, 2012)	Correlation analysis, General linear model (GLM)	Multi Country (Latin-America)	Socio-economic development	5 months of cell phone calls and SMSs from over 10, 000, 000 pre-paid and contract sub- scribers across twelve large- and middle-sized cities	private		Call activity or consumption, Social network variables, including diversity, Mobility patterns	Socio-economic level from Census (National Statistical Institute)	trained	R^2=0.83	yes	
(Li et al., 2013)	General linear model (GLM)	Regional (31 provincial regions of mainland China)	Economic development	1) NPP-VIIRS (2012) 2) DMSP-OLS (2009-2010)	public		Sum of lights	China Statistical Yearbook for Regional Economy and Urban Statistical Yearbook of China, GRP (2010)	trained	NPP-VIIRS data more predictive for GRP than DMSP-OLS. For 2010, at the provincial level R^2= 0.87 for NPP-VIIRS and R^2= 0.71 for DMSP-OLS. For 2010, at the county level R^2= 0.85 for NPP-VIIRS and R^2= 0.67 for DMSP-OLS.	yes	
(Watmough et al., 2013)	GLM / spatial properties	Regional (Assam, India)	Rural poverty	Landsat (2002)	public		Environmental morphology	Indian National Census; female literacy (2001)	trained	Poor results in predicting a single social-economic condition (female literacy)	yes	
(Lessmann & Seidel, 2017)	GLM / spatial properties	Global	Inequality	DMSP-OLS (1992-2012)	public	Global Administrative Areas (GADM), Gridded Population of the World (GPW) v.3	Area (un)lit (%), Light intensity, Frequency saturated pixels	Dataset provided by Gennaioli et al. (2014) (regional income)	trained	Complex modelling resulting in maximum R^2 of 76% within regions and 85% between regions. Contribution (Beta) for night-light decreases to 0.102 (10% increase in income with 10% increase in light) when taking all other available variables into account.	yes	
(Ghosh, Powell, Elvidge, et al., 2010)	GLM / spatial properties	Global	Formal + Informal economic activity	DMSP-OLS (2006)	public	Landscan (2006)	Sum of lights	World Development Report (2008) and CIA World Factbook for GDP. GSP statistical organisations respective countries. DYMIMIC developed by Schneider for informal economy estimates.	trained	Regression in groups based on similar ratio of light to GDP. High R^2 > 0.9 for regression groups.	yes	

Reference	Analytical methodology	Geography	Concept	Primary Big Data source	Abailibility primary Big Data	Other data	Big-data feature	Ground thruth data	Ground thruth data used for training?	Quality of the analytical model	Quality measured against ground thruth data?	Alternative for ground truth data
(Elvidge et al., 2012)	General linear model (GLM), Index calculation	Global	Inequality	DMSP-OLS (2006)	public	Landscan (2006)	Area (un)lit (%)	World Bank Gini Coefficient Human Development Index (HDI) Series of other variables (not specified)	not trained	Night Light Development Index (NLDI) calculated following same reasoning as Gini. NLDI not correlated with income Gini, inversely correlated with electrification rates (R^2=0.69) and HDI (R^2=0.71), positive correlated with poverty rate (R^2=0.59) and multidimensional poverty index (R^2=0.64). More results available.	yes	
(Dai et al., 2017)	General linear model (GLM)	Country (China)	Economic development	1) NPP-VIIRS (2014) 2) DMSP-OLS (2013)	public		Sum of lights	China Statistical Yearbook, GDP (2014)	trained	NPP/VIIRS data more predictive for GDP than DMSP/OLS. Highest fits are: at the provincial level R= 0.93 for NPP-VIIRS and R= 0.87 for DMSP-OLS. at the city level R= 0.93 for NPP-VIIRS and R^2= 0.84 for DMSP-OLS.	yes	
(Smith et al., 2013)	General linear model (GLM)	Country (Yvory Coast)	Poverty	Orange - Data for Development (D4D) Challenge	private		Call activity or consumption, Introversion, Gravity residuals, Social network variables, including diversity	Multidimensional Poverty Index (MPI) (2005)	trained	All covariates correlate with poverty, range R= -0.85 - 0.80	yes	
(X. Chen & Nordhaus, 2011)	Other	Global	Poverty	DMSP-OLS (1992-2008)	public		Sum of lights	G-Econ dataset (1990-2005), gridded GDP Worldbank (1992-2008), country GDP	trained	Luminosity data may be a useful supplement to current economic indicators in countries and regions with very poor quality or missing data. Minimal added value for other countries due to high measurement error in NTL data.	yes	
(Pokhriyal et al., 2015)	General linear model (GLM)	Country (Senegal)	Poverty	Orange (mobile network operator)	private		Call activity or consumption, Introversion, Network advantage, Gravity residuals	Oxford Poverty & Human Development Initiative; multi-dimensional poverty index (2011)	trained	Highest correlation reported = 0.92	yes	

IOB
Institute of Development Policy
University of Antwerp