Assessing Progress towards Development Goals: Data Source Tradeoffs and Complementarities

Dr. Fumiko Kasuga
Senior Fellow, Center for Global Commons
University of Tokyo, Japan

Abstract

Data sources available to researchers and decision-makers are growing, and the potential savings and uses of new data to provide insights into development is the subject of considerable research[1]. This paper provides a framework for evaluating utility and quality tradeoffs when choosing data sources, and applies the framework to a case study comparing survey data and satellite-based estimates on consumption in Malawi and Ethiopia. The case study illustrates that while satellite data provides some relatively cost-effective insights into consumption, it still depends on (validating) survey data and may be less accurate when used to study some demographic sub-populations mentioned in the Sustainable Development Goals.

Keywords – data for development; SDGs; data quality; satellite data; survey data

1 Introduction

Policy-makers increasingly face multiple data options to track and evaluate progress towards development goals. High dimensional data – from satellite imagery to social media – provide a massive and diverse array of sources from which numbers can be generated. Example applications range from using Twitter data as an early warning system for spiking food prices in Indonesia (UN Global Pulse, 2012) to using nightlights and satellite imagery to predict poverty across the globe (Jean et al., 2016). Such novel data sources can also be coupled with novel computational tools such as machine learning, which automates the process by which data can be analyzed for patterns in order to make predictions previously available only through survey and census data. These advancements offer an opportunity to track statistics that previously relied on costly or infrequently conducted surveys.

Yet choosing which data sources and methodologies to use involves making tradeoffs on numerous dimensions[1]. Our goal is to support decision-makers thinking across data options by developing a data dimension framework that illustrates those tradeoffs and complementarities with a case example. Our framework is based on reviewing the literature, distinguishes between novel and traditional data sources, and the data dimensions of utility and quality. We then apply this framework to a case study rooted in the Sustainable Development Goals (SDGs) to illustrate a tradeoff that is most relevant to decision-making in low-resource settings. The approach we use builds on Jean et al.'s (2016) study of satellite imagery as a predictor of per-capita consumption at the community level.

We compare the use of publicly available and low-cost but also indirect assessment using satellite data to higher-cost but direct assessment using survey data from the 2016 Malawi and Ethiopia LSMS-ISA survey to construct indicators for sub-populations within the UN Sustainable Development Goals (SDGs). This particular example illustrates that while satellite data hold great potential as a means of measuring consumption, they remain complex to use well, predict less accurately in rural areas common to development goals, and still require quality survey data to train and validate.

2 Categorizing Data Sources and Dimensions

We begin with a review of the data for development literature, resulting in two outputs.

We first construct a typology to distinguish between traditional and novel data sources commonly used in the development literature, as shown in Figure 1. Traditional

[1] See FAO & ITU (2019), MacFeely (2019), UN Global Pulse (2012), and Ziesche (2017)

data sources include those that have commonly been used in development, such as household surveys, administrative data such as government records, and census data. Novel data sources include a diverse and growing range of data types from recent technological advancements and trends. This includes, for example, sensor data – such as that from satellite imagery or scientific sensors such as weather or pollution stations. While satellite imagery has been in existence for some time, recent gains in precision, affordability, and access have resulted in its increasing utility to development actors (Ziesche, 2017). This typology thus provides a helpful framework that can be used to identify and examine differences between traditional and novel data sources.

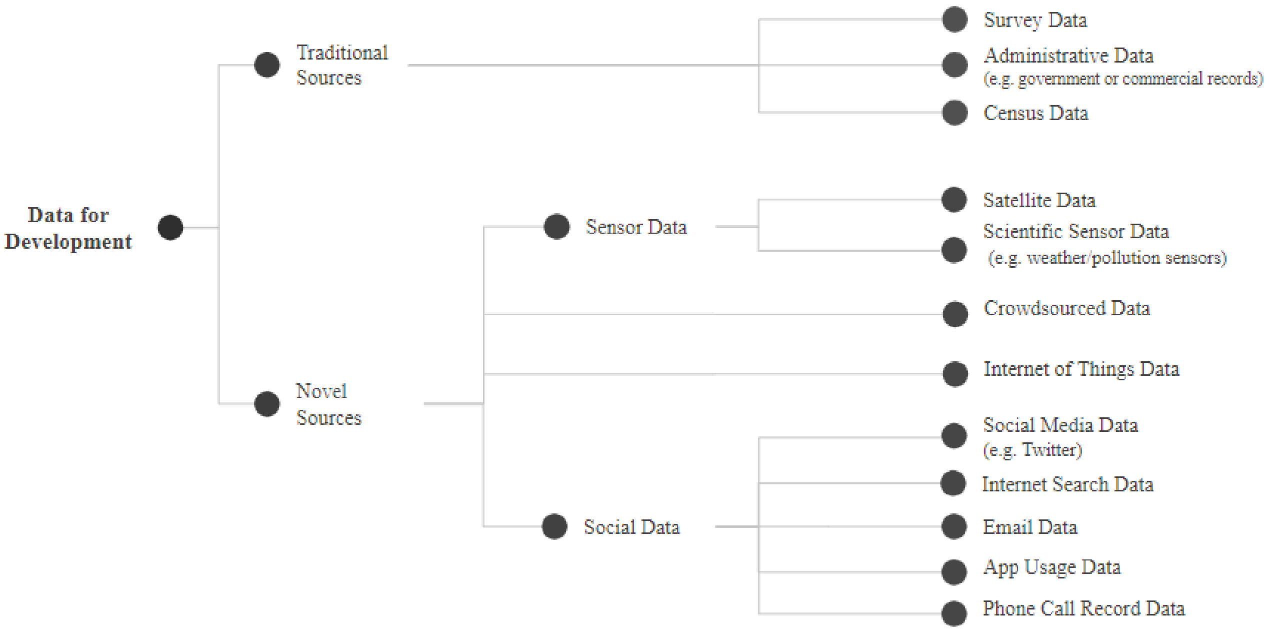

Figure 1. Development data source typology.

Second, we identify dimensions that can be used to assess each of the data sources shown above according to their ability to measure progress towards development goals. Data utility refers to the practicality and usefulness of how and when the data are collected for a given purpose – their public accessibility, spatial coverage, and temporal periodicity. Data quality refers more to the actual observation, how accurately and consistently it represents what is trying to be measured. Table 1 below lists these dimensions of data utility and data quality, as well as their definitions based on key literature sources.

Table 1. Definitions of data source dimensions.

	Dimension	**Definition**	**Key Source(s)**
Data Utility	Access & Availability	The ability access and use data; ease of use	IMF (Laliberté, Grünewald, & Probst, 2004)
	Cost	The financial costs associated with collecting, but not using, data (e.g. equipment, training, time, and materials)	SDSN (2015)
	Timeliness	Data are collected when needed and frequently, and the timeliness between data collection to dissemination	USAID (2009) UN DESA (2019)
	Coverage	Available for the geographies and subpopulation of interest	SDSN (2015)
Data Quality	Accuracy & Validity	The data capture the truth, measure what is intended, and are appropriate for inferences made; the absence of bias	OECD (2011) UN DESA (2019)
	Precision	The level of detail is sufficient for the purpose	USAID (2009)

Consistency & Reliability	The absence of measurement error – i.e., differences due to data issues rather than true changes	UN DESA (2019)
Completeness	Representativeness and the proportion of missing values	DAMA UK (2013)
Integrity	No manipulation or transcription error between collection and use	USAID (2009)
Uniqueness	The data are not duplicative	DAMA UK (2013)

3 Applying the Framework

We next apply this framework to a case study relevant to the SDGs, to compare survey and satellite data and assess for what purposes each is advantageous. Increasing the incomes of women, smallholder farmers, youth, and migrant workers are targets of several SDGs. We focus on the accuracy and validity of these data sources by assessing their ability to accurately measure consumption for and classify these demographic sub-populations of interest.

Jean et al. (2016) found that aggregate satellite-derived estimates are similar to survey-based estimates for the same model parameters. Consequently, this work provides an experimental system to analyze potential validity issues when using satellite data to study demographic groups.

3.1 Hypothesized Differences

A priori, one can assess differences in data utility and quality across satellite and survey data using the criteria described in Table 1, as summarized in Table 2 below.

Table 2. Data source dimension hypotheses.

	Data Utility				Data Quality
					Accuracy & Validity
	Access & Availability	Cost	Timeliness	Coverage	Absence of bias for demographic groups
Survey Data	+	-	-	-	?
Satellite Data	-	+	+	+	?

*Relative strengths are denoted by a '+' symbol, and relative weaknesses by a '-' symbol

Our assessment indicates that satellite data are typically cheaper (Jean et al., 2016; Jerven, 2014), have broader coverage globally (Jean et al., 2016; Ziesche, 2017; Donaldson & Storeygard, 2016), and are updated more frequently or even continuously (Jean et al., 2016; Donaldson & Storeygard, 2016). However, survey data may have an advantage over satellite data in terms of access and availability, as satellite data vary between publicly available and proprietary while also requiring challenging and technically complex analytical techniques due to its unstructured nature (Jean et al., 2016).

There are also potential tradeoffs for dimensions of data quality. Crucially, in terms of accuracy and validity, the current precision with which most satellite information is captured means that larger, more easily identifiable objects are often used as proxies for less visible metrics (roofing material used to proxy wealth, or a school to proxy education). Sub-populations can be obscured – though we do note that satellite data may allow for coverage of sub-populations that place-based surveys miss, such as mobile pastoralists. The potential of novel data sources such as satellite data to misclassify certain sub-populations that are of interest in the SDGs poses a potential threat to validity, but remains understudied. We thus focus our analysis on shedding light on this particular dimension of data quality.

3.2 Approach

To evaluate satellite data, we use a model developed by Jean et al (2016) to predict per capita daily consumption from variables visible to satellites. The authors used ridge regression to fit satellite-visible features (e.g. road materials, roof materials, and weather conditions) at the census tract (EA) in the Living Standards Measurement Survey Integrated Surveys on Agriculture (LSMS-ISA) merged with nightlight intensity data from NOAA. The LSMS-ISA survey is project conducted by the World Bank in collaboration with national statistical offices, and conducts multi-topic nationally-representative surveys with a focus on agriculture in eight countries in Sub-Saharan Africa that are publicly available and open access (World Bank, 2016). We use LSMS-ISA survey data from Malawi and Ethiopia, from surveys conducted in 2016. The NOAA data takes raw data from NASA's Virtual Imaging Infrared Radiometer Suite (VIIRS), and creates estimates of average nightlight intensity. NOAA data on average nightlight intensity is available both monthly and yearly (Earth Observation Group 2016); we use yearly average numbers for the analysis in this paper.

3.3 Findings

We explore these tradeoffs by adapting the ridge regression model from Jean et al.'s (2016) methodology as replicated in Mathur (2020) for using satellite data for predicting poverty at the census-tract level in Malawi and Ethiopia using satellite data. Using only attributes that can be seen by satellites (i.e. road material, nightlights), they explain 50% of the variation in consumption between EAs and correctly classify 73% of neighborhoods as poor or not poor (using the international definition of consumption less than $1.90/capita/day) using ridge regression. To examine possible bias in the classifications, we used t-test analyses to compare rates of misclassification in special populations to the whole-sample classification rates.

3.3.1 Variable Selection and Alternative Models

The LSMS-ISA are multi-topic, nationally representative household surveys that collect information, including on metrics such as household roof material. Given the level of multicollinearity and variability in accuracy inherent in these metrics, identifying a parsimonious set of predictors represents a significant challenge. Furthermore, responses can be conceived as continuous (mean consumption value) or binary (mean community consumption is/is not above the poverty line). Interpreting the data as continuous is better for assessing the overall accuracy of satellite estimates of income. However, we will also evaluate the accuracy of satellite models in classifying areas as above or below the poverty line, as the poverty line is an important threshold used in assessing progress in combating poverty (Beegle & Christiaensen, 2019). Also, a binary dependent variable can smooth over extreme values and identify broader trends that may be discrete outcomes of achieving a certain level of affluence (e.g., road paving or new construction with more durable roofing materials).

For Malawi, the models produced by Jean et al. (2016) and Mathur (2020) identified nightlights, roof materials, and road materials as the most significant variables. To see if variable selection would change if the model was run using a non-parametric dependent variable, we fit a logit model using lasso variable selection to the same data set, with EA's categorized as above or below the $1.90/day consumption threshold for poverty. In the logit model, nightlights, roof materials, and road materials were also the most significant factors associated with poverty status, indicating that they are important indicators not only for consumption estimation, but also for poverty prevalence assessment.

3.3.2 Rural Versus Urban Household Consumption

83.6% of Malawi's population lived rurally in 2016 (World Bank, 2016). The validation sample for Malawi was 780 enumeration areas (EAs), of which 78 (10%) were located in a major urban center. The average consumption of the urban center EAs was higher than the sample as a whole ($4.60 PPP/capita/day for urban, $2.50 PPP/capita/day for all); thus, odds of misclassification were lower largely due to fewer urban EAs falling close to the poverty line – where misclassification rates are expected to be higher – than rural ones.

In Ethiopia, the rural-urban divide, both in numbers and income, is smaller, with about 30% of the sample coming from EAs classified as non-rural. Rural EAs were still poorer than urban ones on average ($1.73 PPP/capita/day versus $2.97 PPP/capita/day).

In practice, this greater level of households and census tracts close to the $1.90 poverty threshold in rural areas means that the certainty of estimates based on satellite data is reduced significantly. The model was 9-22% ($p<0.05$) more likely to indicate that a poor neighborhood (according to the household survey) is not poor (when using the satellite data) if the neighborhood is predominantly agricultural.

Furthermore, the ridge regression used to predict EA consumption based on variables derived from satellite images has an R^2 value of 0.55 when using all EAs, but only 0.30 when using only rural EAs, further demonstrating difficulty of estimating rural poverty using satellite data. What is visible to a satellite does not explain the variation in consumption in rural areas as well as urban ones. This likely to be due to the fact that rural areas in both countries are not electrified. In Malawi, 11% of households had electricity in 2016. In Ethiopia, it is 33% (World Bank, 2016). Thus, nightlight intensity, a significant indicator of income, cannot be used in most rural areas to predict income.

3.3.3 Demographic Groups of Interest

In addition to comparing overall model accuracy and the effect of the urban-rural divide, we used LSMS-ISA data to identify subpopulations that might experience different levels of classification accuracy due to the possibility that they (a) live in less affluent, rural areas or (b) may lack sufficient representation in the population as a whole to significantly influence the model, leading the model to fail to take into account factors that may indicate higher or lower incomes among these populations. From data in the LSMS-ISA, we determine what percentage of household heads in each census tract were women or youth (age 10-24), and whether the mean daily per-capita consumption of each census tract fell within the bottom forty percentiles its country. In addition, the LSMS-ISA survey asks a representative from each census tract whether they "come during certain times of the year to look for work" (World Bank, 2016). We used this question to determine in which census tracts migrant workers live and work.

When a rural dummy is used, we find that none of these demographic groups are associated with larger errors or increased likelihood of being misclassified as poor or not poor. However, some demographic groups are more likely to live in rural areas, and thus satellite models will likely have more trouble accurately estimating their incomes.

We used t-tests tests to determine which demographic groups were more likely to live in rural areas, with the results shown in Table 3 below.

Table 3. Accuracy of satellite estimates of consumption and demographic factors.

Table 3a. Malawi.

Demographic Indicator	T-Test Result	Probability
% Female-Headed Households in EA	0.534	0.000
% Youth living in EA	-0.254	0.274
EA has seasonal migrant workers	-0.132	0.000
EA is in bottom 40 income percentiles	0.300	0.000

Table 3b. Ethiopia.

Demographic Indicator	T-Test Result	Probability
% Female-Headed Households in EA	-1.01	0.000
% Youth living in EA	2.31	0.000
EA has seasonal migrant workers	0.447	0.000
EA is in bottom 40 income percentiles	0.373	0.000

Female-Headed Households

EAs with higher percentages of female heads-of-household are significantly more likely to be found in rural areas in Malawi, but in urban areas in Ethiopia. Thus, while we cannot say that satellites will inevitably have trouble estimating the incomes of female headed households, this result underscores the importance of understanding the demographic layout of a country and bias and validity issues when selecting data sources.

Bottom 40% Income

EAs in the bottom 40% of income percentiles were significantly more likely to be rural than other areas. Thus, we expect satellite models to have more trouble estimating incomes accurately among these populations.

Migrant Workers

EAs where respondents indicate that people come to seek work during certain times of the year were significantly more likely to be found in urban areas in both Ethiopia and Malawi. Thus, we have no reason to believe satellite estimates will have any particular trouble estimating consumption in these areas, and the ability of satellites to provide new data in each season could be advantageous to researchers studying seasonal workers.

Youth

We calculate the percentage of people living in each census tract who are between the ages of 10 and 24, the definition the UN uses in the SDGs (UN, 2020). We find no statistically significant relationship between the percentage of youth in a census tract and whether it is in a rural or urban area in Malawi, but youth are more likely to be found in rural areas in Ethiopia. Again, it seems that the validity of satellite estimates of youth income and consumption depends on the country being studied.

3.4 Discussion

This analysis helps to illustrate the tradeoffs present when determining what data source to use. In the application to poverty analysis, and as shown in Jean et al. (2016), satellite models are more effective at predicting income and the prevalence of poverty in urban areas, with electricity and nightlights. For rural areas where poverty is less detectable via observable proxies, satellite-based models will likely perform less accurately.

Thus, returning to the question of validity from Table 2, our results show that the validity of satellite data depends very much on the question it is being used to answer. To a researcher interested in changes in income or consumption between seasons within a year in towns and cities, satellite data can provide a reasonably accurate answer, and satellites can provide more timely estimates than surveys can. Such a researcher might be willing to trade a certain amount of accuracy for the ability to see seasonal dynamics in consumption changes. However, a researcher evaluating progress toward SDG 2.3 (which aims to double the production and incomes of small-scale producers) over the last decade would be especially concerned about the problems with accuracy in rural areas that satellite models have, would not be as interested in within-year income changes, and would thus likely prefer to use survey data.

We acknowledge a few shortcomings of our methodology. The first is the absence of any infallible data on neighborhood-level poverty in Malawi and Ethiopia and the assumption that survey data are correct, though they contain their own well-documented biases (Tasciotti & Wagner, 2017; Carletto & Gourlay, 2019). Second, some of the errors in the satellite estimates may be due to the methodology, which relies on regression and LSMS-ISA reported data on satellite visible factors, rather than on shortcomings inherent with using satellite data, which generally rely on neural networks and image data (and, as a result, do not provide explanations on how they make decisions). It is also important to note that regression is one of a multitude of approaches to processing CNN-identified satellite features and shortcomings in the regression approach are not the same as shortcomings in CNN-based feature identification. This difference is of particular note because regression has constant coefficients for each independent variable, while neural networks are capable of understanding conditional relationships (i.e. that the relationship between roof material and consumption might be different in rural areas than urban ones).

Simultaneously, it is important to note that one major reason for the relatively poor performance of the model in rural areas is the overall lack of information; for example, some rural areas did not have *any* nightlights in these areas making the possibility of an alternative relationship moot. It is likely any satellite-based model will struggle to differentiate between the incomes of rural populations, because there are simply fewer satellite-visible factors that correlate with consumption for these populations than their richer or more urban counterparts. Thus, the results obtained using the Jean et al. (2016) approach likely hold true for other models with other approaches, as they stem from trends in the underlying data.

3.5 Further Research

We propose a framework for evaluating data sources for use in answering various questions of interest in development, and use it to compare survey and satellite-based estimates of daily consumption for use in evaluating progress for the SDGs. Satellite data is far from the only new data source being used to construct development indicators, and it will be important to evaluate what tradeoffs are present for other novel sources of data, such as social media data and cell phone data, to evaluate progress in development.

4 Concluding Remarks

Our results tend to indicate that, since the strongest satellite visible indicators of income, such as nightlights, tend to be less reliable in rural and poor areas, satellite-based estimates of income are less accurate in rural areas. Certain demographic groups, female household heads, smallholder farmers, and the poorest 40% of the population, are overrepresented in rural areas. While there are many tradeoffs between satellite and survey data, researchers interested in these groups should be wary of the validity of satellite data for their purposes. More optimistically, satellite data has greater accuracy in estimating income and poverty in more urban areas, and can provide updated data at a greater frequency than surveys can, something that will be of interest to researchers studying changes in income and consumption within years.

Another consideration is access and availability of satellite data and its analysis. A tradeoff to consider when selecting data sources is that it takes training in computer programming to create or even use satellite models for poverty prediction. Thus, when including people from areas where access to computers or higher education is an important part of research, survey models provide a more inclusive way of obtaining and disseminating data.

Finally, models using satellite data currently must be trained with ground-truthed data so that they can learn to make estimates. The variation we see in the relationship between income and poverty, and factors such as electrification and urbanization, underscore that there is no universal relationship between satellite-visible factors and income. Satellite models will always rely on survey data in order to identify relationships. Thus, these two data sources can complement each other, as well.

This application of the data source typology and dimension definitions highlights that there is no “best” data source for all questions. We find that while there are tradeoffs between traditional and novel data sources, there is also the potential for complementarities to better measure progress toward development goals. But a more detailed and transparent understanding of data source tradeoffs is important to informing more effective development policy.

Acknowledgements

This work was supported by the Bill & Melinda Gates Foundation through a grant (grant #636303) to the Evans School Policy Analysis & Research Group (EPAR). The findings and conclusions presented here are those of the authors and do not necessarily reflect positions or policies of the foundation.

References

Beegle, K. & Christiaensen, L. (2019) Accelerating poverty reduction in Africa. World Bank, 10.1596/978-1-4648-1232-3, https://elibrary.worldbank.org/doi/abs/10.1596/978-1-4648-1232-3

Carletto, C., & Gourlay, S. (2019). A thing of the past? Household surveys in a rapidly evolving (Agricultural) data landscape: Insights from the LSMS-ISA. Agricultural Economics, 50(S1), 51–62. https://doi.org/10.1111/agec.12532

DAMA UK (2013). The six primary dimensions for data quality assessment: Defining data quality dimensions. Retrieved from https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

Donaldson, D., & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. Journal of Economic Perspectives, 30(4), 171–198. https://doi.org/10.1257/jep.30.4.171

Earth Observation Group (2016). Night Band Nighttime Lights. NOAA National Centers for Environmental Information. https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html

FAO (2020). Factsheets on the 21 SDG indicators under FAO custodianship. A highlight of the main indicators with the greatest gaps in country reporting. Rome. https://doi.org/10.4060/ca8958en

FAO & ITU (2019). E-agriculture in Action: Big Data for Agriculture. Bangkok. Retrieved from http://www.fao.org/3/ca5427en/ca5427en.pdf

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. https://doi.org/10.1126/science.aaf7894

Jerven, M. (2014). *Benefits and costs of the data for development targets for the post-2015 development agenda.* Copenhagen Consensus Working Paper. Retrieved from https://www.copenhagenconsensus.com/sites/default/files/data_assessment_-_jerven.pdf

Laliberté, L., Grünewald, W., & Probst, L. (2004). Data quality: A comparison of IMF's data quality assessment framework (DFAQ) and Eurostat's Quality Definition. Retrieved from https://www.imf.org/external/pubs/ft/bop/2003/dataq.pdf

Mathur, J. (2020). Jmather625/predicting-poverty-replication [Jupyter Notebook]. https://github.com/jmather625/predicting-poverty-replication (Original work published 2019)

MacFeely, S. (2019). The big (Data) bang: Opportunities and challenges for compiling SDG indicators. *Global Policy, 10*(S1), 121–133. https://doi.org/10.1111/1758-5899.12595

OECD (2011). Quality dimensions, core values for OECD statistics and procedures for planning and evaluation statistical activities. Retrieved from oecd.org/sdd/21687665.pdf

Sustainable Development Solutions Network (SDSN). (2015). Data for development: A needs assessment. Retrieved from https://sustainabledevelopment.un.org/content/documents/2017Data-for-Development-Full-Report.pdf

Tasciotti, L., & Wagner, N. (2018). How much should we trust micro-data? A comparison of the socio-demographic profile of Malawian households using census, LSMS and DHS data. The European Journal of Development Research, 30(4), 588–612. https://doi.org/10.1057/s41287-017-0083-6

UN General Assembly (2015). Transforming our world: the 2030 Agenda for Sustainable Development. Resolution A/RES/70/1. Retrieved from http://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/70/1

UN DESA (2019). United Nations national quality assurance frameworks manual for official statistics. Retrieved from https://unstats.un.org/unsd/methodology/dataquality/references/1902216-UNNQAFManual-WEB.pdf

UN Global Pulse (2012) Big Data for Development: Challenges and Opportunities. Retrieved from https://beta.unglobalpulse.org/wp-content/uploads/2012/05/BigDataforDevelopment-UNGlobalPulseMay2012.pdf

UN Sustainable Development Goals (2020). Youth and the SDGs. https://www.un.org/sustainabledevelopment/youth/

USAID (2009). TIPS data quality standards. Retrieved from https://www.fsnnetwork.org/sites/default/files/tips-dataqualitystandards.pdf

World Bank (2016). LSMS-ISA. https://www.worldbank.org/en/programs/lsms/initiatives/lsms-ISA

Wold Bank (2016). Access to Electricity. Sustainable Energy for All database, https://data.worldbank.org/indicator/EG.ELC.ACCS.RU.ZS

Ziesche, S. (2017). Innovative Big Data Approaches for Capturing and Analyzing Data to Monitor and Achieve the SDGs. Report of the United Nations Economic and Social Commission for Asia and the Pacific: Subregional Office for East and North-East Asia (ESCAP-ENEA). Retrieved from https://reliefweb.int/sites/reliefweb.int/files/resources/Innovative%20Big%20Data%20Approaches%20for%20Capturing%20and%20Analyzing%20Data%20to%20Monitor%20and%20Achieve%20the%20SDGs.pdf