Data disaggregation for the SDGs

2021

UNITED NATIONS
AZERBAIJAN

SUSTAINABLE DEVELOPMENT GOALS

A GUIDEBOOK FOR PRACTITIONERS

Azərbaycan Respublikasının
Dövlət Statistika Komitəsi

SUSTAINABLE
DEVELOPMENT
GOALS

This Guidebook identifies current challenges to data disaggregation in the context of Azerbaijan and offers attainable solutions based on best practices. It supports a partnership that has been set up between the United Nations Country Team (UNCT) and the State Statistics Committee of Azerbaijan (SSC) to bring SDG data collection and analysis methods on a par with international standards. To this end, the UNCT, in collaboration with the SSC, organized a workshop in November 2020 where local and international experts exchanged their experiences on and approaches to data disaggregation in various country contexts. The parties agreed to develop a guidebook for practitioners, reflecting some of the lessons learnt during the seminars

The Guidebook aims at strengthening expertise within the institutions that produce SDG data. It also addresses Development Partners in Azerbaijan that work with these actors and others to produce and analyze SDG data. Practitioners and decision-makers in the Europe and Central Asia region and beyond may also find inspiration in this Guidebook to guide their thinking on SDG data collection and analysis.

CONTENTS

CONTEXT AND RATIONALE

LEAVE NO ONE BEHIND

Leave No One Behind (LNOB) is the central promise and fundamental tenet of the Agenda 2030. This principle suggests that no Sustainable Development Goal (SDG) should be met unless it is met for everyone. In other words, governments should target the most vulnerable populations to achieve the SDGs. Ensuring that the commitments outlined in the Agenda 2030 are translated into effective action requires better understanding of the nature and scope of vulnerability in society.

Disaggregated, timely and quality statistical data is imperative to identify people who are left behind in the development process. Aggregates and averages are not enough to assess the extent to which the needs of marginalized groups have been met, as they typically capture national averages. On the other hand, disaggregation of data allows to unpack and assess inequalities in various groups. The availability of data at finer level allows for a comparison of different population groups that helps achieve a clearer measure of inequality and identify patterns that would otherwise go unnoticed. For example, a country can achieve target related to aggregate employment, while availability of disaggregated data may help uncover the incidence of unemployment across vulnerable groups (i.e. youth, women, migrants etc.). As such, data disaggregation is not only important for the identification of vulnerable populations but also for the design of inclusive policies that benefit the target groups. Data-driven decision-making is essential for evidence-based policies which allow governments to effectively use limited resources. The COVID-19 pandemic is a good case in point to highlight the importance of data disaggregation for an optimal allocation of limited health resources. Evidence suggests that hospitalization and mortality rates for COVID-19 are higher for elderly people, people with chronic medical conditions and health professionals. Thus, disaggregated data is imperative for determining the relative size of each risk group and design appropriate policy responses.

The global SDG indicator framework has an overarching principle of data disaggregation: "Sustainable Development Goal indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics, in accordance with the fundamental principles of official statistics ".

This Guidebook was produced at a time Azerbaijan has started designing recovery plans from the second wave of the COVID-19 pandemic (autumn of 2020) and the recovery plans in the liberated territories after the flareup of hostilities in and around Nagorno-Karabakh. From this perspective, adequate level of data disaggregation can help assess the social, physical and psychological impact of the pandemic, as well as the military conflict on disadvantaged people. For example, thousands of Azerbaijanis were displaced and wounded, while almost 3,000 Azerbaijani people lost their lives as a result of the war. Similarly, thousands of families were directly or indirectly impacted by the pandemic. As such, to ensure no one is left behind, availability of disaggregated data and rapid identification of the vulnerable families are very crucial.

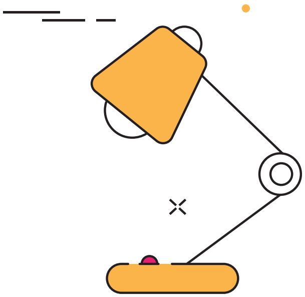
Additional incentives for disaggregation

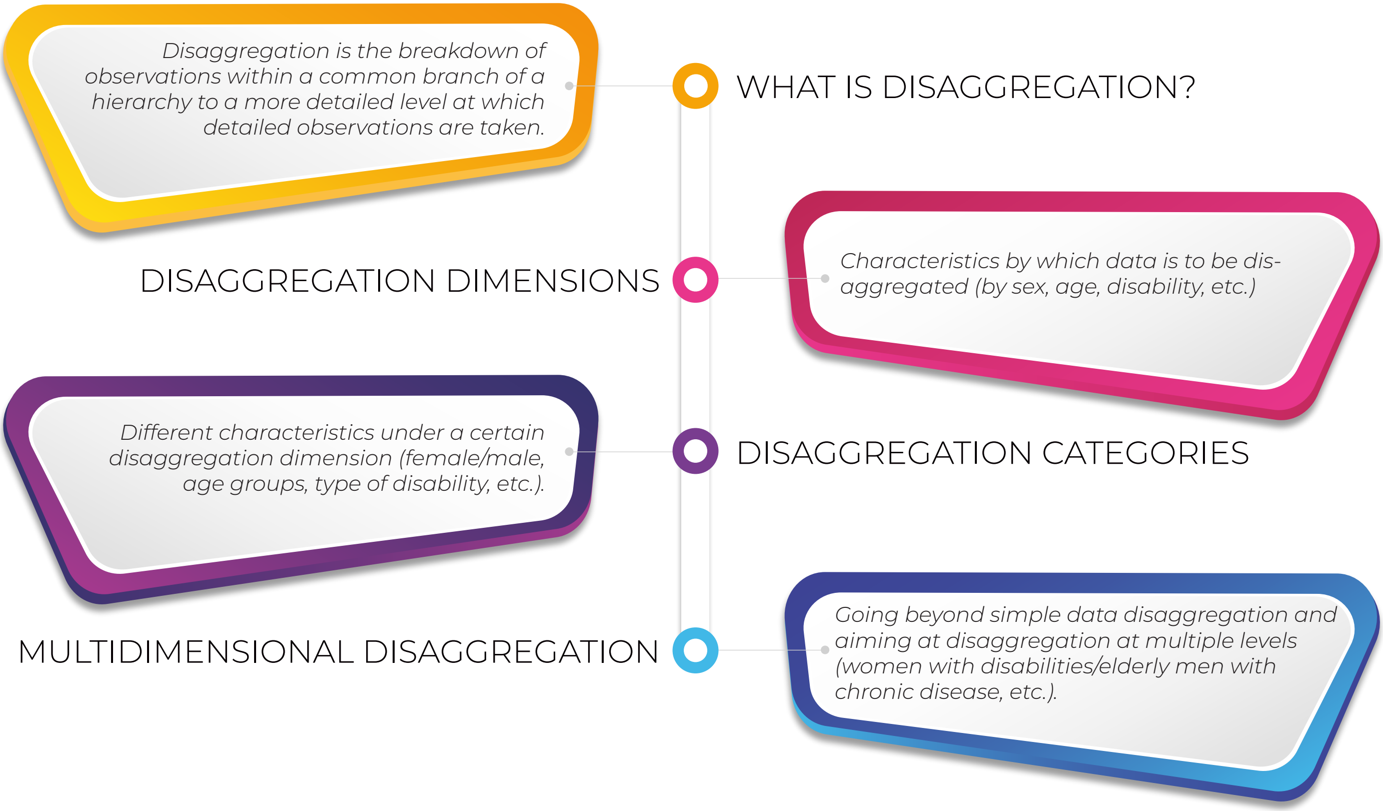
The five-year Sustainable Development Cooperation Framework signed by the Government of Azerbaijan with the United Nations for the period 2021-2025, as well as the decree by the President of the Republic of Azerbaijan signed on February 2, 2021 (No. 2469) – "Azerbaijan 2030: National Priorities for Socio-Economic Development" create additional impetus for availability of disaggregated data. The Guidebook is therefore forward-looking, focusing on methodologies that can improve the evidence to support these strategic planning efforts.

The next section of the Guidebook introduces basic concepts and discusses the importance of data availability for measuring multidimensional poverty. The Guidebook then briefly discusses the state of data disaggregation for SDG monitoring in Azerbaijan. It provides an overview of the availability of granular data in the context of Azerbaijan and discusses the plans. Next, the Guidebook proceeds with a discussion of methods to improve data disaggregation in the context of Azerbaijan. Practical steps towards implementation of Small Area Estimation (SAE) are discussed in detail. Challenges and opportunities to harness possibilities of the big data are discussed next. Use of mobile phone data for data disaggregation purposes is also discussed.

MAJOR CONCEPTS

What is disaggregation?





WHAT IS DISAGGREGATION?

Disaggregation is the breakdown of observations within a common branch of a hierarchy to a more detailed level at which detailed observations are taken.

DISAGGREGATION DIMENSIONS

Characteristics by which data is to be disaggregated (by sex, age, disability, etc.)

DISAGGREGATION CATEGORIES

Different characteristics under a certain disaggregation dimension (female/male, age groups, type of disability, etc.).

MULTIDIMENSIONAL DISAGGREGATION

Going beyond simple data disaggregation and aiming at disaggregation at multiple levels (women with disabilities/elderly men with chronic disease, etc.).

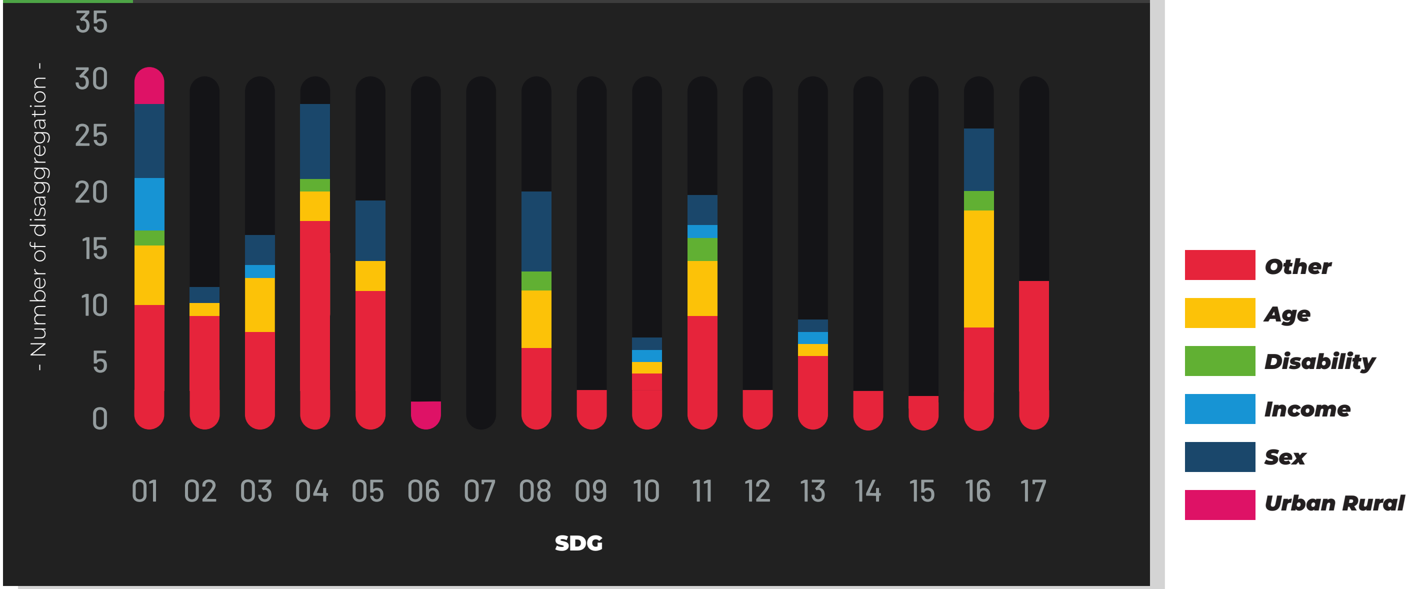
WHY IS DATA DISAGGREGATION IMPORTANT?

Data disaggregation allows to :

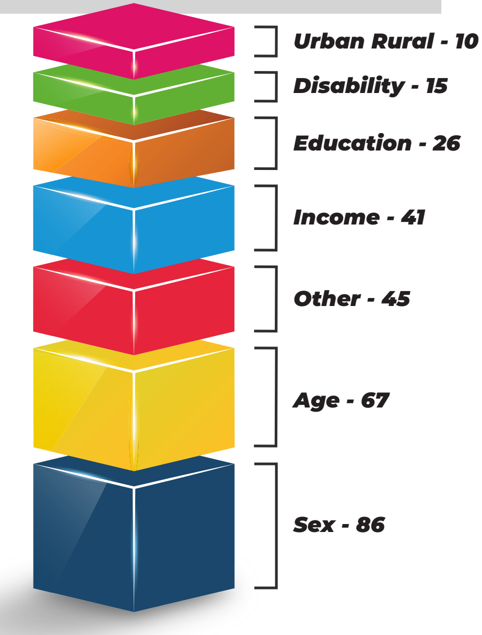
- assess progress towards the Sustainable Development Goals
- put the ‘Leave No One Behind’ promise into practice
- develop effective preventive measures
- design inclusive development policies
- assess the impact of policies on populations at risk of deprivation and exclusion
- identify inequalities between and within different groups

Overall, more than 200 indicators throughout 17 SDGs require disaggregation. The most common are by sex, age, income or wealth, and education. Other critical types of disaggregation such as disability, race and ethnicity, urban or rural location, employment, citizenship, and indigenous status are rarely required.

Minimum disaggregation requirements by Goals



Frequency of disaggregation categories



Multidimensional disaggregation

While single level disaggregation analysis is an important first step towards understanding vulnerabilities of particular population groups, it fails to adequately identify those who face deprivations across multiple dimensions. The evidence suggests that deprivations tend to cluster and reinforce each other. Intersection of different forms of discrimination for sub-populations requires assessing exclusions through a multidimensional lens. Only policies that are tailored to address root causes can be effective to ensure these groups are not excluded.

Consider a hypothetical example in a country where poverty line stands at USD 150. A 45-year old woman from little village moves to the capital city and rents an apartment with her child. She is employed non-formally with a low monthly salary of USD 310. Also, she suffers from a respiratory disease. Despite low monthly wage and illness, she is not identified as vulnerable by existing policies since her salary is above per capita poverty line (USD 310/2=USD 155) and has no underlying conditions such as disability. Growing medical bills amplify the financial struggle. In such a case, while she is not classified a poor by a single indicator, she must be considered vulnerable when several characteristics of age, sex, health, wealth, and location are put together simultaneously.

Example

Monitoring intentional homicides is necessary to better assess their causes, drivers, and consequences and, in the longer term, to develop effective preventive measures. If data are properly disaggregated, the indicator can identify the different types of violence associated with homicide: inter-personal violence including partner and family-related violence, crime including organized crime and other forms of criminal activities and socio-political violence including terrorism and hate crime. Each of these results would require different policy responses with different target beneficiaries.

Counting the invisible: who could be left behind?

Hard to reach: *Subgroup of population who are difficult to target for a variety of reasons such as being small or having specific characteristics.*

Hidden population: *When public acknowledgement of the population is potentially threatening for the members of the subgroup. Size of the population is often unknown due to privacy and security concerns (e.g. LGBT persons, irregular migrants, person living with HIV)*

Excluded, marginalized, and discriminated: *Representatives of this group are 'known', but 'ignored' in one way or another (e.g. Roma populations, ethnic minorities)*

Vulnerable sub-population groups: *A sub-group of population that is potentially in a disadvantaged position due to its socio-economic situation (e.g. uninsured, low income, informal employees, elderly persons)*

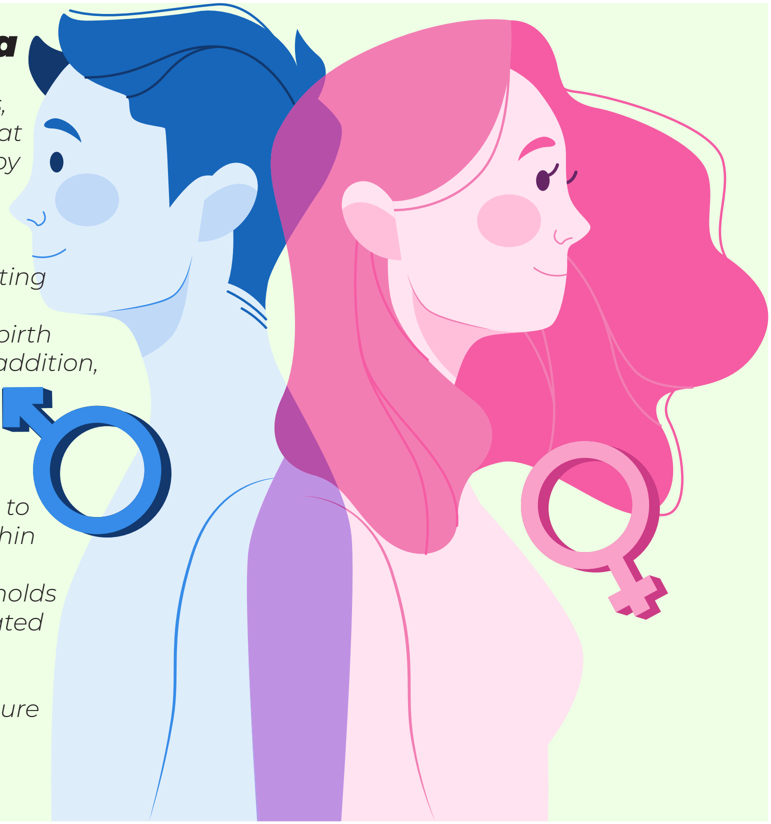
Geographically disadvantaged: *sub-populations that live in geographically unfavorable location either because of harsh climate, remoteness, poor infrastructure or hard to access.*

As the hypothetical scenario above depicts, an invisible share of a population may face multiple deprivations simultaneously. Several studies have shown that vulnerabilities tend reinforce each other. For example, while people with disabilities face discrimination in terms of job opportunities, children with disabilities face further exclusions (stigma, exclusion from education process, healthcare challenges). Similarly, the discrimination that a young single woman working in the private sector in a large city faces (for example, a lower pay compared to her male peers) is much different to challenges faced by a single mother of three residing in a remote village. Women in each category require different anti-discrimination response actions.

An important challenge in identifying people who face multiple deprivations is they are simply hard to reach. This is especially true for poverty, which has multiple dimensions beyond income. Meeting the requirements of LNOB principle requires a closer insight into the main challenges confronting the people. This approach moves away from measures of material deprivation and instead suggests addressing the root causes and structural drivers of poverty, including dimensions such as differences in access to education, judicial challenges to claim property rights, and entrenched cultural stereotypes. Thus, improving both the collection and dissemination of raw data to assess multidimensional poverty allows for a more targeted reach and policy response for this population.

Disaggregation of gender data

UN Women has identified that out of 232 indicators, 54 are gender-specific, meaning they are targeted at women and girls, explicitly call for disaggregation by sex or refer to gender equality as the underlying objective. When the SDG metadata considered, sex disaggregation is required for 73 indicators, with 13 additional indicators applying only to women, resulting in a total of 86 gender-relevant indicators. Data on violence against women or on assistance of skilled birth attendant at delivery are examples of the latter. In addition, there are indicators that are gender-related, meaning they monitor areas that indirectly affect women and girls. For example, SDG indicator 6.1.1 (proportion of population using safely managed drinking water services), monitors change in access to an improved water source located on premises (within the dwelling, yard or plot). As women and girls are responsible for water collection in 8 out of 10 households with water off-premises, the indicator is gender-related (UN Women, 2018).
To this end, disaggregation of all gender-relevant indicators (beyond indicator title) is desirable to ensure full analysis of vulnerabilities that women face.

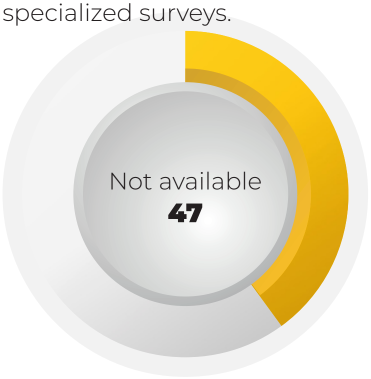


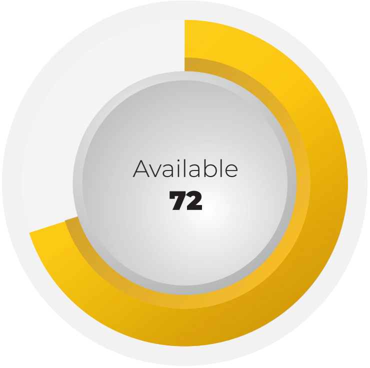
STATE OF DATA DISAGGREGATION IN AZERBAIJAN

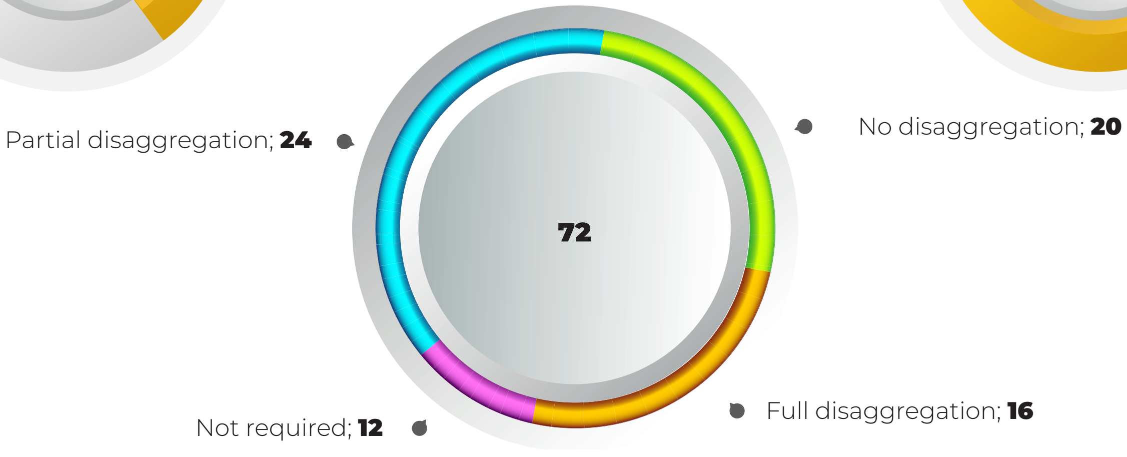
Azerbaijan nationalized SDGs and prioritized 17 SDG goals, 88 targets, and 119 indicators covering economic, social, and environmental aspects of sustainable development. The SSC is a key national agency responsible for processing and maintaining an effective and responsive database to measure progress in the achievement of nationalized SDGs.

Although collection of SDG data has started a few years ago, significant progress towards availability of disaggregated SDG data has been made. Based on the information provided on the SDG portal run by the SSC, of 119 prioritized indicators, data for 72 are available, while data on remaining 47 indicators are not available or in the process of collection. Most data for the SDGs are derived from official statistical bulletins, surveys, the census, and administrative records provided by other agencies. Household survey and analysis of employment data are key sources of socio-economic data in addition to specialized surveys such as Multiple Indicator Cluster Survey (MICS). The Multiple Indicator Cluster Survey (MICS) planned for 2022 will be used to complement the existing data sources, in addition to specialized surveys.

Availability of disaggregated data for prioritized indicators

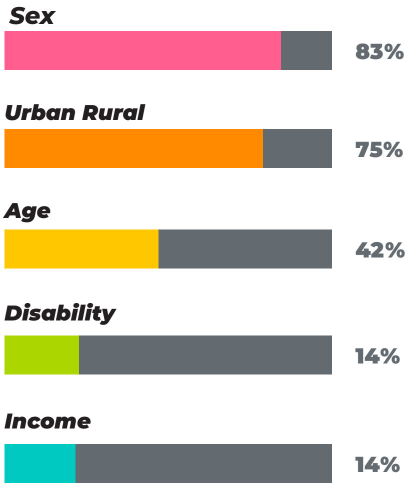






Of available 72 indicators, 58 are produced by the SSC, while remaining 14 are produced by other agencies (Ministry of Agriculture-1, State Committee for Family, Women and Children Affairs- 2, State Migration Service- 1, Ministry of Ecology and Natural Resources- 2, Ministry of Internal Affairs- 1, Ministry of Justice- 2, Prosecutor's Office- 1, Ministry of Foreign Affairs- 3, Azersu OSC- 1). Disaggregation dimension of 12 indicators has been expanded in 2020.

Moreover, 12 indicators do not require disaggregation. Only 16 indicators have been fully disaggregated as per minimum disaggregation requirement developed by Inter-agency Expert Group on SDGs. Partial disaggregation is available for 24 indicators, while disaggregated data is not available for 20 prioritized indicators (although aggregate data exists).



Priority in Azerbaijan is the SDG data on indicators level of disaggregation, %

The SSC works in close cooperation with other state agencies to improve availability of granular SDG data. As a result of this collaboration, disaggregated data for 13 indicators is expected by the end of 2021. In addition, data for nine SDG indicators is expected from MICS (to be held in 2022), while data for four indicators will come from statistical surveys planned as part of the State Program on Improvement of the Official Statistics in the Republic of Azerbaijan in 2018-2023. Last, data for one indicator will be sourced from the 2019 Census.

Bias coming from inadequate cover

Institutional residents (prisons, elderly care, military establishments etc.) and transient groups (homeless people, street children) are often omitted from household surveys. This tendency is explained by several economic, logistical, and administrative difficulties. However, failure to cover these groups may result in biased results and lead to ill-informed policy choices. For example, a study based on OECD countries shows that although general health of elderly population has improved over time, health of elderly population residing within institutions have deteriorated.

USING EXISTING DATA SOURCE : FOCUS ON HOUSEHOLD SURVEYS

Along with being a rich source of data for evidence-based policy, household surveys are among central pillars of data ecosystem for monitoring progress towards SDGs. Globally, over 80 SDG indicators, spread across 13 Goals, can be disaggregated using household surveys. They are an important instrument for identifying those who are left behind. With the emergence of innovative data sources, several disadvantages associated with household surveys have raised concerns about their usefulness. Nevertheless, surveys are likely to remain one of the most important sources of SDG data for the foreseeable future.

Adequate sample size for policy making

The level of available geographical disaggregation from household surveys may not be optimal for policymaking, requiring more granular data. For example, poverty statistics in Azerbaijan is available for urban/rural split. To ensure targeted policies, the government would like to have poverty data for all 65 regions or ideally, at village level. However, the sample size of the household survey is too small to provide precise estimates of poverty at such a granular level. Yet innovative data techniques, such as use of non-traditional data sources such as satellite imagery and mobile phone data, allow for combination of various data sources to produce data at required level.

Household surveys allow for collecting wealth of information related to different characteristics and behavior of households with wide range of disaggregation options. Majority of the SDG indicators calculated from household survey data have a requirement of at least one level of disaggregation.

While high-quality disaggregated survey data allows for in-depth analysis of various population groups, sample size requirements have restricted exploiting full potential of household survey data. Surveys are generally well suited to accommodate disaggregation by age, sex, and living place, however as the disaggregation level becomes finer, minimum required sample size increases. This means that existing survey methods are not able to accommodate data requirements for disaggregation beyond certain point. An attempt to estimate indicators at more granular level may lead to lower quality and reliability of the estimate. As such, existing capacities need to be amplified to ensure minimum sample size.

This, however, leads to another drawback associated with household surveys- high implementation costs. Household surveys are generally costly and resource intensive. Ensuring additional disaggregation dimension would lead to larger sample size requirement and significantly increase survey budget.

To improve the coverage, as well as the disaggregation potential of household surveys follow steps are recommended:

• PLAN AHEAD

Planning ahead is a critical step towards achieving adequate disaggregation. Identifying all stakeholders that are involved in collection and dissemination of data and ensuring their active participation in the process is a crucial process. Taking stock of missing dimensions and mapping causes for missing data should be followed by a study of resource requirements for expanding the scope of surveys. Statisticians should determine relative costs and benefits of collecting additional data, noting high resource requirements and tight budget. Alternative sources of data should be considered along with improving existing data collection methods. If all of the desired outcomes cannot be achieved, prioritization of outcomes and options will be required.

1 - ***Identify and Involve Stakeholders***
a.Who are major stakeholders and how to involve them?

2 - ***Assess existing data systems***
a.What data collections and systems related to the specific indicator are currently in use?
b.Can the existing capacity be sufficient for data disaggregation?
c.If not, what needs to be done?

3 - ***Estimate costs***
a.What are costs?
b.What are non-cost requirements (e.g. technical capacity, political support etc.)?

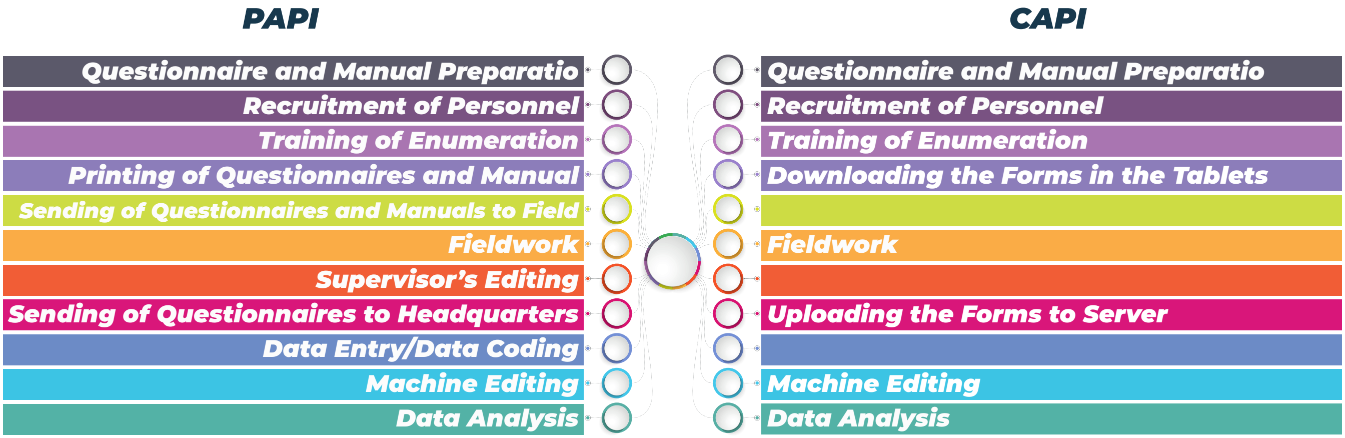
4 - ***Develop a plan***
a.Identify timelines and roles and responsibilities

5 - ***Update data systems***
a.What needs design, development, and testing?
b.How will data be collected and shared?
c.Identify questions, sample parameters

6 - ***Review data quality***

• Exploit full potential of existing resources

Household surveys are tried and tested, yet underexploited sources of disaggregated data. One way of harnessing more from surveys is to expand its scope (i.e. increasing sample size; adding new modules) keeping existing data collection practices. However, this is a resource-hungry endeavor as was discussed above. An alternative option is to find efficient solutions by leveraging advances in information and communication technologies.



PAPI	CAPI
Questionnaire and Manual Preparatio	Questionnaire and Manual Preparatio
Recruitment of Personnel	Recruitment of Personnel
Training of Enumeration	Training of Enumeration
Printing of Questionnaires and Manual	Downloading the Forms in the Tablets
Sending of Questionnaires and Manuals to Field	
Fieldwork	Fieldwork
Supervisor's Editing	
Sending of Questionnaires to Headquarters	Uploading the Forms to Server
Data Entry/Data Coding	
Machine Editing	Machine Editing
Data Analysis	Data Analysis

Modernization of the data collection: **Hungarian case**

Hungarian Household Budget Survey is an annual survey covering about 8,000 households. Paper diary (PAPI) was the standard data collection method but since 2015 e-Diary (CAWI) has been offered as an option for respondents. The online diary uses a special IT infrastructure developed by a contractor for the purpose. The tool offers predictive data entry, user friendly structure and tailor-made modular system.

The web-entry option considerably speeds up the data collection and increases the quality of COICOP coding process. It shortens the evaluation period and improves the quality and consistency of the recorded data. The inclusion of the online monitoring system provides up-to-date information on the evolution of the achieved sample over time.

Despite efforts, the share of e-Diary is very small in total sample. However, with increasing digital literacy, this share is likely to increase going forward. The authorities offer incentives (offering lottery, tablets) for household to switch to online version.

For example, advancing the introduction of computer aided personal interviewing (CAPI) and Computer-Aided Web Interviews (CAWI) modes of data collection could lead to substantial cost savings and open additional opportunities for increased scope. Computer-based solutions are cheaper to implement and results could be processed much faster. In addition, software-based data collection method would reduce the probability of enumerators making errors during interviews [1].

Another way of reaping full benefits of existing survey data is to combine it with other data sources, such as census (discussed in more detail below).

> *Overall, it is estimated that worldwide, some 300 to 350 million people may be missing from survey sampling frames, either by design (at least 45%), i.e., omitted altogether, or in practice, because they are likely to be undercounted.*

• Advocate for data transparency and flexibility

Household surveys and administrative data are likely tobe major sources of data for analysis and so need to be publicly available in a timely manner. Enabling and promoting data transparency would allow disaggregation at multiple layers and pave the way for analysis of multidimensional deprivations. To this end, a solid foundation of evidence exists about the benefits of open data.

• Redress exclusions

Including all populations is important as all people have a right to be counted. The exclusion of the most vulnerable groups suggests the sampling framework of household surveys become incomplete and may lead to biased survey results. One option is to include non-household groups in censuses and surveys. An alternative approach is to consider complementing existing surveys with targeted supplementary surveys for excluded groups (i.e. surveys among institutionalized population). The results then could be integrated to the regular household survey or presented alongside wider results. Use of innovative data sources could also be considered to identify excluded population.

[1] *The experience of other countries has shown that this transition may pose challenges of its own if not managed appropriately.*

SMALL AREA ESTIMATION

Despite immense value for disaggregation, survey data is either not available or is unreliable for subnational areas due to small sample size. On the other hand, census, which technically should include all citizens, covers only limited amount of information about respondent's socio-economic background. Borrowing strength of these two data sources, Small Area Estimation (SAE) provides an analytical framework for improving the level of granularity without necessarily collecting additional data in the field.

It refers to a set of statistical techniques involving the estimation of parameters for smaller sub-populations than those for which original survey was designed. These sub-populations often have too few sample observations for producing direct estimates with sufficient reliability. For example, if household survey data was collected to reflect urban/rural split with no reliable further disaggregation, SAE can be used to produce these estimates at town level. SAE combines the survey data for the indicator with correlated auxiliary variable data and estimates their relationship. For example, household income in regional level may be regressed on village-level values of related variable (for example, level of educational attainment) to describe the relationship between these two.

What is Small Area?

A small area could be either a geographical unit or a subpopulation group that is not adequately represented in the underlying survey because it is finer than the pre-specified survey domain and because of limitations in sample size.

Step 1: Determine the plan and requirements

In many instances, requirements need to be articulated right from the start at the planning period. Granular estimates should be generated only when there is no alternative proxy indicator and there is a clear policy rationale. For example, does the government need poverty data at village level? If yes, would alternative data be used as a proxy? If potential uses of any statistical activity are not clear at the outset, mobilizing budget may prove to be a very difficult task.

Step 2: Identify dependent variable

The process should proceed with identifying statistics to be reported and related requirements (for example, software requirements). The level of disaggregation and data requirements need to be considered at this stage. This is important as different dependent variable may require different model assumptions and auxiliary variables. For example, it is more challenging to model the indicator if dependent variable is binary than continuous.

Step 3: Identify auxiliary variables

The key requirement for successful implementation of SAE is finding good auxiliary variables. As such, care needs to be taken while choosing these variables. The explanatory variables should be a good predictor of the outcome variable. For example, since education attainment (from census) is typically a good predictor of income, it is a good candidate for including in the SAE model to estimate poverty for sub-population.

Definitions, concepts, classifications, and other elements of the methodology need to align well with the modelled variable. For example, if the quality of data source is poor, this may reflect in the quality of SAE estimates.

Testing the strength and significance of the relationship between the auxiliary data and the variable(s) of interest is imperative. This could be done through simple scatter plots, correlation, and simple models.

Almost any SAE method will work well if the auxiliary variables are closely related with the outcome variable of interest.

Although typical sources of auxiliary data are census, surveys and administrative data, with the evolution of Big Data, innovative data sources are utilized more often in the context of SAE model. This include data from social media and private sector.

Bear in mind

Although quality auxiliary variable is a key requirement for SAE method, it is important to note too many auxiliary variables may cause the model estimation to fail to converge, and the accumulation of errors in the auxiliary variables add uncertainty to the small domain predictions. Therefore, when many auxiliary variables are in consideration, some method of reducing the set of variables is required.

Qualities of good proxy indicator

- *Be related to outcome variable*
- *High quality*
- *Available for all areas of population*
- *Minimum missing values*
- *Timely*
- *Consistently defined*

Step 4: Modelling and estimation

SAE-based estimators are based on an assumed relationship between the outcome variable of interest and the auxiliary data. The functional form between variables, informed by earlier steps (what is the outcome variable, what level of disaggregation is required etc.), provides the foundation for the structure of the estimation model.

The next step is to choose estimation procedure. There are wide range of SAE estimation techniques with advantages and disadvantages of their own. The choice of modelling techniques varies from the most straightforward direct survey estimation model to more complex regression-based models. Depending on the specifications of outcome variable (for example, if it is a categorical variable that takes on value of 0 or 1), binary dependent regression model (for example, logistic model) may be appropriate.

Limitations

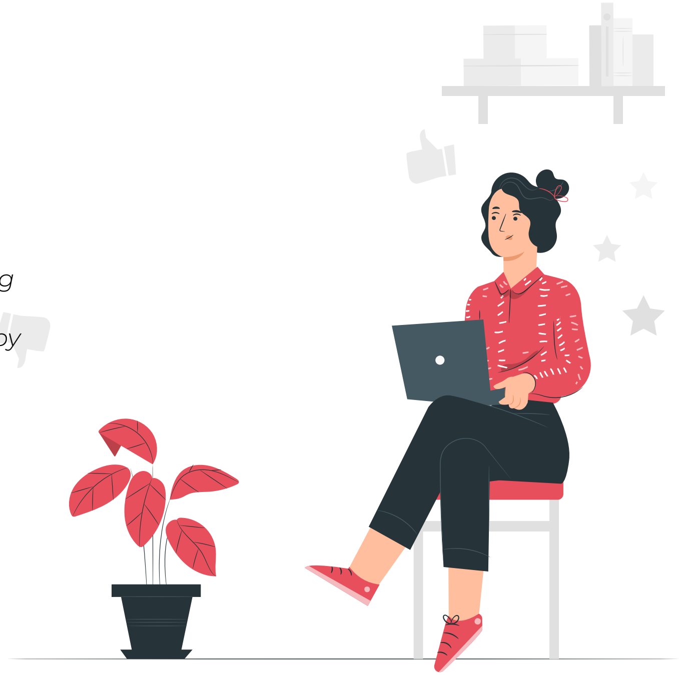
Despite obvious advantages, SAE technique has several limitations:

- *Different time periods for combined data sources*
- *Reliability of estimates*
- *Missing data for entire areas*

Step 5: Evaluation and validation

After the fitted model has been estimated and thoroughly checked, it is wise to evaluate the resulting small domain estimates and validate them against other sources of data where possible. For starters, the small domain estimates should closely align with the direct survey estimates in domains with larger sample sizes. For example, if the results from higher disaggregation level (i.e. urban/rural) for a variable is not consistent with more granular estimates (i.e. by regions) coming from SAE, it is likely that the model contains error. Furthermore, the root mean squared errors of the small domain estimates should be better than those of the direct survey estimates, or else there is no point to doing SAE.

Sometimes a separate survey or census provides estimates that can be used to cross-check the estimates produced from the SAE. For example, the American Community Survey (ACS) generates estimates of many outcome variables which can be useful for validating the SAE results.



Assessing quality of estimates

- *Are the SAE estimates in line with survey results using larger domain?*
- *Are the coefficients generated from SAE supported by underlying theory?*
- *Is the model parsimonious?*

Harnessing benefits of SAE using conventional method

To estimate poverty at the provincial and municipal level, the Philippines Statistical Authority applied SAE technique integrating Family Income and Expenditure Survey, Labor Force Survey and Census of Population and Housing. A single model supplemented by urban/rural effects within each region was found to be an adequate model for predicting household income and expenditure at provincial level, allowing for estimation of poverty. The estimates were broadly in line with results of official survey estimates (at larger geographical domain) with relatively acceptably standard errors. The results of the study brought considerable benefits for the purposes of targeted social assistance programs. For example, target beneficiaries were identified following the identification of poor municipalities or those living in pockets of poverty.

Recognizing the growing benefits of unconventional data sources, ADB and Philippines Statistics Authority took on a pilot initiative to integrate Big Data into a Small Area Estimation Framework (discussed below)

USING UNCONVENTIONAL DATA SOURCES

At present, data generation activities are going beyond the traditional data sources such as censuses, surveys and administrative records. The surge in new data source from digital technology opens enormous opportunities for collecting immense amount of quality, disaggregated and timely data at a fraction of the cost required by traditional methods. The integration of new data sources, such as satellite imagery, social media, mobile telephony and citizen-generated data, as well as the use of innovative techniques and technologies such as Big Data analysis and Artificial Intelligence, can help address some data gaps to fulfil the promise of leaving no one behind.

Adapting to this paradigm shift in data generation, national statistical systems around the world are gradually exploring the smart use of existing and new sources of data to support evidence-based policy making. UN led Big Data Group survey in 2016 showed that statistical community is accepting Big Data as a new strategic priority. Few years later, in 2021, many statistical offices have already started work on integrating alternative data sources into official statistics. However, the potential of Big Data has not been fully utilized by national statistical offices due to several challenges:

• Privacy, security and confidentiality challenges

Big data encompasses masses of personal information that raises civil liberty concerns. Failure to protect privacy and confidentiality, use of Big data for surveillance purposes, vulnerability to cyberattacks are among key risks that could also lead to reputational damage. Inadequate protection of fundamental human rights can raise serious ethical concerns at a time when big data, artificial intelligence and algorithms are increasingly used for policies and marketing by private sector.

As such, adequate regulatory and legislative framework for the collection, analysis, and sharing of data from innovative sources could act as a steppingstone acting as safeguards to restrict the misuse (for example, for discriminatory purposes) and mishandling of disaggregated data.

• Legislative challenges

At present, the majority of national statistical offices relies only on traditional data sources in production of official statistics. There is no legal base to reflect the multitude of data sources. The legislation is not flexible to support the use of new data sources, for example, by allowing information exchange between the national statistical offices and public and private data holders. As such, integration of non-traditional data into the official statistics requires legislative changes.

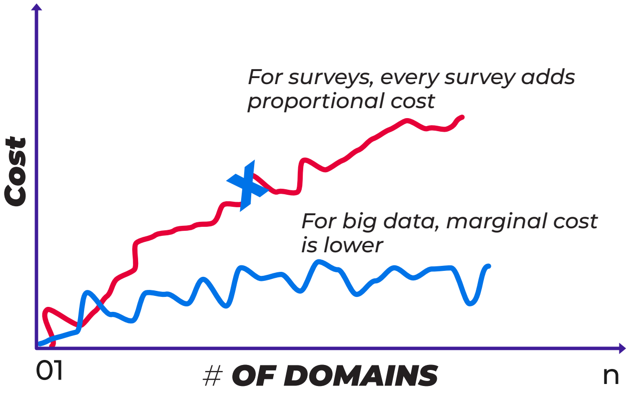
• Data quality

There are quality concerns for indicators produced by alternative data sources. Most of the data coming from Big Data is unstructured and proper data manipulation techniques should be applied to transform them into time series data. If the process does not meet minimum data quality requirements, unconventional data sources could pose a threat to credibility of official statistics. It is therefore essential for the producers of data to expand collaborations with international organisations to adopt international practices at national level. These actions would address the quality concerns related to data production and allow for integrating alternative data into official statistics.

• Lack of financial resources

Although producing data using alternative sources may be substantially cheaper compared to traditional methods, establishing adequate infrastructure can be costly. In the case of geospatial data, for example, the provision of maps from satellite imagery relies on having spatial data infrastructure, which is costly. Apart from funds required to build modern data infrastructure, the process of data generation (for example, process of anonymization of disaggregated data prior to publication) may incur significant costs.

Many statistical offices around the world lack financial resources to invest in Big Data. Continuing business as usual, along with transitioning to new data sources may be challenging in this fiscally constrained environment. As such, any move towards exploring the use of alternative data sources should be preceded by adequate cost-benefit analysis.



• Technical capacity

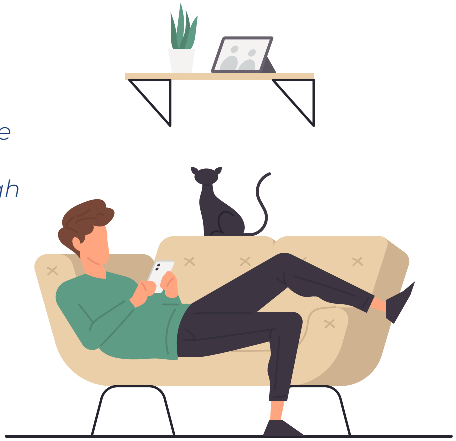
Big Data pose entirely unprecedented challenges in IT infrastructure and data analysis. Many statistical offices lack technical and organizational capacities to exploit the full benefits of new data sources. Upgrading technical capacities of the staff to generate and carefully process and analyze big data is crucial.

USE OF MOBILE PHONE DATA

All challenges discussed above are very relevant to the SDG data disaggregation contexts. Integration of data generated by private sector or civil society into official statistics is important, yet complex process. National statistics offices should consider seizing these new opportunities that rapid technological advances create. The experience of countries with similar socio-economic background can be used for inspiration. This section focuses on the potential of incorporating data generated by mobile phones, which would allow to collect quality, timely and disaggregated data for SDG monitoring.

Benefits of using mobile phone data

- ***Better coverage of population allowing disaggregation:*** *High mobile phone penetration allows a better coverage of population. This in turn allows for disaggregation at more granular levels compared to data collected through traditional methods.*
- ***Validating survey statistics:*** *In both Indonesia and Georgia mobile phone data can be used to validate and support findings from survey-based tourism statistics, particularly domestic tourism.*



- ***Improving accuracy:*** *Accuracy is also a recognized driver for using big data, including mobile phone data. The accuracy of traditional data collection methods, particularly sample surveys, needs to be considered when assessing various data sources for official statistics.*
- ***Reduced cost:*** *Cost is a recognized driver by many for using big data, including mobile phone data. The cost of traditional data collection methods, such as household surveys, is an increasing reason why national statistical offices are looking at alternative data sources. Using mobile phone data may not replace the need for household surveys, but may complement the existing surveys.*
- ***Timely:*** *Data derived from mobile phones is more timely compared to those based on traditional methods.*
- ***Alleviate respondent burden:*** *Passive data collection puts no burden on respondents in contrast to responding to a traditional survey.*

According to the SSC data, there were 104 mobile phone subscriptions in Azerbaijan per 100 people in 2020. Moreover, the entire population of Azerbaijan has been covered with mobile phone since 2015. In addition, the cost of internet, including mobile phone data has been consistently declining.

High level of mobile phone and internet penetration opens immense opportunities for filling data gap for SDG monitoring in Azerbaijan. Due to its ubiquity, cell phones are arising as one of the main sensors of human behavior that might help explore various socioeconomic characteristics of population. High population coverage and tendency to carry mobile phones on the person opens rich source of geolocation and time information that can inform on population densities and mobility. Apart from being an additional data source, mobile phone and internet data can help improve the granularity of information which is much higher than what can be obtained through traditional surveys. In addition, mobile phone data would allow to cut significantly time required for data collection.

PRACTICAL STEPS TOWARDS USE OF MOBILE PHONE DATA

Step 1: Preparatory works

The first step should be exploring a range of questions related to the objective of and preparedness towards use of mobile phone data for official statistics. The former implies outlining the ultimate goals of using Big Data (i.e. what data will be produced as an outcome?). The latter, implies carefully considering existing challenges, including legal barriers, privacy and confidentiality concerns. The feasibility study should also consider technological readiness (software and hardware requirements, technical capacities of staff) for using Big Data. Once these potential benefits and costs are clearly identified, a cost-benefit analysis should be developed to make a strong case for using mobile phone data for official statistics.

Step 2:
Advocacy and policy dialogue

Exploiting Big Data without cooperation from the owners of the data- mobile operators and other related stakeholders is nearly impossible. Advocacy work and policy dialogue with these stakeholders is a crucial step for SSC to enable access to mobile phone data. This step should aim at ensuring cooperation at three levels:

Data philanthropy *describes a form of collaboration in which private sector companies share data for public benefit.*

• *Mobile Network Operators (MNO)*

Partnership with MNOs is the only feasible way to integrate mobile phone data into official statistics. Convincing the MNOs to contribute to this Data Philanthropy initiative is a key step towards successful outcome. Even in the absence of legislative requirements mandating data sharing, operators interested in sustainable development may opt to partner as part of their corporate social responsibility strategy.

• *Government and Parliament*

There must be appropriate legal and institutional foundations for network providers to give access to data in a manner that is secure and respects the privacy rights of individuals. Ideally, this means that legislative access arrangements allowing (or potentially mandating) private sector providers to share data be adopted and enforced. The absence of adequate legal and institutional safeguards makes the use of mobile phone data for official statistics riskier. This, along with failure to ensure cybersecurity, could lead to resistance from telecommunications providers to cooperate.

• *Citizens*

Ensuring public acceptance of the use of mobile phone data for official statistics is equally important. As it was mentioned above, important advocacy work is required for convincing the public about adequate use of data to eliminate concerns about privacy and confidentiality.

Several countries are useful case studies of regulation allowing access to data held by private entities on other respondents. In Croatia, for example, recent change of the Statistical Law enables access to data held by private legal persons containing information on other statistical units for statistical purposes. In Latvia, the amendments to the legislation stipulate private entities to provide data from its administrative data sources, including restricted access information needed for production of official statistics. Currently, based on these articles, Latvia has received individual data from private entities for:

- ***Data on electricity consumption from the maintainer and developer of the electricity network in Latvia***
- ***Data on individual students from private universities***
- ***Individual data from personalized e-tickets from the provider public transport systems, motor transportation and parking services***

Similarly, legislative amendments stipulate that for the purpose of rational implementation of the activities of the national statistics, the Statistical Office of Slovenia may use data from various official and other administrative data collections of the public and private sectors (records, registers, databases, etc.). In compliance with law, register holders must, free of charge, submit to the Office and to authorized producers all the requested information.

Step 3:
Implementation

Once all administrative and legal procedures are in place allowing integration of mobile phone data into national statistics, the national statistics authority should start the implementation phase in cooperation with the MNOs. Following steps should be agreed for successful outcome:

• Agreeing the principles for data processing and dissemination

The process entails several aspects related to data generation process including preparation, anonymization, encryption, transmission and archiving. Depending on the agreement between MNOs and national statistics office, these steps can be implemented by either parties. There are certain risks and opportunities associated by either approach. Regardless of where the processing happens, the method for data transfer should also be agreed on in advance. In addition, authorities and MNOs should agree on the methodology to ensure data meets minimum quality standard.

• Upgrading physical infrastructure and technical capabilities

The national statistical offices should ensure there is adequate computing infrastructure in place to handle Big Data. In addition, the staff should have adequate training to handle enormous amount of data. Even if the earlier stages of data processing take place outside the national statistical offices, they will bear ultimate responsibility for quality assurance and dissemination. Providing anonymized data at granular level using mobile phone data will be at the national statistical offices' domain, which will require enhancing in-house big data analytics capacity. To this end, it will be important to identify staff who will have ongoing contact with MNOs and provide regular trainings to build domain-specific knowledge of mobile phone data.

Partnership
Advocacy
Skills
Funding
Technology

- *Data preparation: The data preparation step typically involves creating a data extraction script to extract the necessary data from the storage unit. It should be done by the MNO, most likely employing a database communication language such as SQL.*

- *Data anonymization: The purpose of the data anonymization process is privacy protection. During this process, a subscriber's personal identity code can be modified, or data can be aggregated to give anonymity to the subjects.*

- *Data encryption: Data encryption is a small but important step that is needed when data processing takes place outside an MNO's premises. Its purpose is to ensure that unauthorized parties are not able to read the data in the case of a security breach during data transmission.*

- *Data transmission: During transmission, data will be made available for the NSI either for in-house processing or processing outside an MNO's premises (centralized). In the first case, the data is generally transported into a database dedicated for this purpose or made available in the form of a data file on a server that the receiving end is able to access.*

- *Data archiving: In the case of an accidental data loss due to technical problems or any other unforeseeable cause, it is important that historical data could be reacquired and used for recalculation. For the MNOs, this simply means the re-extraction of historical (archived) data.*

POTENTIAL USE OF MOBILE PHONE DATA FOR BETTER DISAGGREGATION

Poverty

Mobile data could complement existing poverty assessment methodologies not only verifying the existing estimates, but also allowing for forecasting poverty at more granular level. The methodology is based on identifying relationship between mobile phone data and socio-economic indicators and forecast poverty based on this relationship. The experience of pioneering countries show that alternative sources of data could be used to provide predictions of factors of well-being and socio-economic status and allow for filling gaps in SDG data. For example, a study focusing on England and Ivory Coast has shown that mobile phone data can be used to calculate poverty indicators at a very granular level. Similarly, this approach was undertaken in Colombia, where cell phone records of 500,000 citizens were used to forecast poverty at local level. The results suggest that socio-economic variables produced by the National Statistics Office highly correlate with the data collected from mobile phones (number of calls, the average distance covered whilst making calls and the distance calls are made to or received from).

Quality of education

Research has established strong correlation between access to mobile phones and educational outcomes. In addition, mobile phones can also be used to track the performance related to indicators related to education. For example, the ratio of SMSs to voice calls tend to correlate with literacy rates. Predictions based on this tendency can help map literacy rate across various ethnicities and geographies and lead to localized interventions. This method was applied to mobile phone data to estimate literacy rates for each commune in Senegal.

Food insecurity

Big Data also allows for identifying communities that are most vulnerable to food insecurity. Mobile phone data provides a proxy for assessing progress towards SDG 2 indicators in real-time and at high geographic granularity. This approach is based on the finding from academic literature, which shows high correlations (over 80 percent) between mobile phone data derived indicators and several relevant food security variables such as expenditure on food or vegetable consumption. Some countries have already piloted projects relying on this relationship. For example, a Uganda-based NGO assessed food insecurity in various parts of the country to identify communities at risk.

APPENDIX 1: BIG DATA CLASSIFICATION BY UNECE

Data Source*	Data Type	Statistical Domains	Additional Statistical Domains**
Social Networks	1100. Social Networks: Facebook, Twitter, LinkedIn 1200. Blogs and comments 1600. Internet searches and search engines (Google) 1700. Mobile data content: text messages, Call Detail Record, Data Detail Record, Location update, Radio coverage updates, Online news	National accounts External sector statistics Financial statistics Price statistics Government finance statistics (public debt statistics)	Sentiment indices (investor, consumer) Social statistics Labour statistics Migration statistics Tourism statistics Population statistics Household consumption statistics SDG indicators Early-warning indicators Urban statistics
Traditional Business Systems	**Data produced by public agencies** Administrative data	Government finance statistics National accounts Price statistics External sector statistics	SDG indicators
	Data produced by businesses 2210. Commercial transactions 2220. Banking/stock records 2230. E-commerce 2240. Credit cards Business websites Scanner data	National accounts Price statistics External sector statistics Financial statistics	Social statistics Business registers Employment statistics Household consumption statistics Transport statistics SDG indicators
Internet of Things (machine-generated data)	**Data from sensors** **311. Fixed sensors** 3111. Home automation 3112. Weather/pollution sensors 3113. Traffic sensors/webcam 3114. Scientific sensors 3115. Security/surveillance videos/ images **312. Mobile sensors (tracking)** 3121. Mobile phone location 3122. Cars 3123. Satellite images	National accounts Satellite accounts External sector statistics Government finance statistics Price statistics	Traffic/transport statistics Energy statistics Land use statistics Agricultural statistics Environment statistics Transport and emission statistics Air emission statistics SDG indicators
*Based on adapted UN big data classification			**Based on European Statistical System Committee (2014)

1 NO POVERTY

2 ZERO HUNGER

3 GOOD HEALTH AND WELL-BEING

4 QUALITY EDUCATION

5 GENDER EQUALITY

6 CLEAN WATER AND SANITATION

7 AFFORDABLE AND CLEAN ENERGY

8 DECENT WORK AND ECONOMIC GROWTH

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

10 REDUCED INEQUALITIES

11 SUSTAINABLE CITIES AND COMMUNITIES

12 RESPONSIBLE CONSUMPTION AND PRODUCTION

13 CLIMATE ACTION

14 LIFE BELOW WATER

15 LIFE ON LAND

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

17 PARTNERSHIPS FOR THE GOALS

SUSTAINABLE
DEVELOPMENT GOALS

Baku 2021