I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

POLICY PAPER SERIES

IZA Policy Paper No. 215

Better Tracking SDG Progress with Fewer Resources?
A Call for More Innovative Data Uses

Hai-Anh Dang
Calogero Carletto
Dean Jolliffe

DECEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

POLICY PAPER SERIES

IZA Policy Paper No. 215

Better Tracking SDG Progress with Fewer Resources? A Call for More Innovative Data Uses

Hai-Anh Dang
World Bank, IZA, Indiana University and LSE

Calogero Carletto
World Bank

Dean Jolliffe
IZA and World Bank

DECEMBER 2024

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

IZA Policy Paper No. 215 DECEMBER 2024

ABSTRACT

Better Tracking SDG Progress with Fewer Resources? A Call for More Innovative Data Uses[*]

Existing data are severely insufficient for monitoring progress on the Sustainable Development Goals (SDGs), particularly for poorer countries. While we should continue efforts to produce new, high-quality data, this approach seems not feasible for all poorer countries. We call for a more systematic use of recent innovations with techniques such as data imputation to address existing data challenges. Given some resistance to utilizing new methods for filling data gaps, efforts aiming at changing the current perception and employing a mix of new data collection and data imputation can be useful. We also note that the best and most cost-effective approach would be highly context-specific and depends on various factors such as available budget, logistical capacity, and timeline.

JEL Classification: C15, I32, O15

Keywords: poverty, imputation, Sustainable Development Goals, developing countries

Corresponding author:
Hai-Anh Dang
World Bank Development Data Group
1818 H Street
Washington, DC
USA
E-mail: hdang@worldbank.org

[*] We would like to thank the UK Foreign Commonwealth and Development Office (FCDO) for funding assistance through various Knowledge for Change programs (P169210) and the Data and Evidence for Tackling Extreme Poverty (DEEP) Research Program (P175686), and the United States Agency for International Development (USAID) for additional funding support (P172751). We would like to thank the editor Roy Maconachie and various colleagues at and outside the World Bank for helpful discussion.

Existing data challenges

Tracking timely progress with the Sustainable Development Goals (SDGs) poses data and measurement challenges for countries, both rich and poor. These challenges are commonly characterized as either a lack of data (e.g., unavailable or infrequently updated data) or low-quality data (e.g., unusable or incomparable data across countries, or even within the same country over time). Concerns have been raised that the current SDG data framework will not be successfully populated if we rely on the existing approaches and mechanisms (MacFeely & Nastav, 2019; Dang & Serajuddin, 2020). As such, new tools, such as citizen science or artificial intelligence, should be employed to address data gaps (Vinuesa *et al.*, 2020; Fraisl *et al.*, 2022). Indeed, a recent review of the United Nations' SDG database suggests that data coverage was just less than 10% of the required data for monitoring poverty trends for the first SDG; for monitoring SDG progress, the corresponding figure was generally less than 20%. More worrisomely, data gaps for other SDGs, such as environment quality, were even larger (Dang & Serajuddin, 2020).

Unfortunately, these challenges are especially severe for poorer countries, where a stronger data need paradoxically exists to improve welfare. For example, more than half of under-five-year-old children in Sub-Saharan Africa are not registered at birth (World Bank, 2021). Reviewing 46 SDG-related health indicators from population surveys in 47 poorest countries during 2015‑2020, Zhao *et al.* (2022) find that just 19 countries collected data on half or more of these indicators, while distressingly, nine countries collected no data on the SDGs. Data gaps are particularly acute for vulnerable population groups, including women and girls (UN Women, 2023). Other vulnerable groups, such as individuals with disabilities, migrant workers, and refugees, simply do not appear in SDG progress reports (Sachs *et al.*, 2023).

While the traditional, standard solution to data gaps is to collect more data, this approach seems not feasible for all poorer countries. For example, SDG number 1 calls for reporting on poverty on an annual basis by all countries. This is a daunting and practically impossible undertaking. In practice, most low and middle-income countries can only implement a new household (consumption or income) survey *every few years*, which forms the basis for official poverty estimates. Unsurprisingly, poorer countries typically spend much fewer resources on collecting data, and they also have lower capacity to frequently collect and timely process data (Dang *et al.*, 2023).

Alternative approaches

Alternative approaches to generating new data exist. Data imputation has been widely used to fill data gaps in various disciplines, such as agricultural and health sciences, and is a common tool for official statistical agencies such as the U.S. Census Bureau. But here we focus on tracking progress with poverty reduction (the first SDG) for clearer illustration.

One early application of data imputation was to produce “poverty maps”—which offer poverty estimates at a lower administrative level than those typically allowed by the available household consumption surveys (Elbers *et al.*, 2003). Specifically, the key idea is to use a household consumption survey (typically with a sample adequate for measuring poverty only at higher administrative units, such as provinces) to build an imputation model using certain predictor variables. This imputation model is subsequently applied to a larger data set that is representative at lower administrative units (such as communes), which contains the same predictor variables but lacks consumption data (such as a population census). The end results are imputed poverty estimates that are statistically representative at a lower level than is possible with just the household survey.

Building on Elbers *et al.*’s (2003) method, imputation has also been applied from older household consumption surveys to more recent, non-consumption surveys (such as labor force surveys or health surveys) to generate updated poverty estimates for many poorer countries in different regions ranging from Sub-Saharan African countries to Bangladesh, India, and Vietnam (Dang & Lanjouw, 2023). Recent imputation studies have introduced new data sources and the latest advances in statistics or machine learning. These studies use administrative data from humanitarian organizations to predict poverty for Syrian refugees in the Middle East or for poor refugees in Chad (Altindag *et al.*, 2021; Beltramo *et al.*, 2024) or satellite data to target poor Nigerian households at more disaggregated geographical levels (Smythe & Blumenstock, 2022). In areas recently struck by unexpected natural disasters, where it may be impossible to implement a new survey, imputation can also be employed if other auxiliary data sources, such as satellite data, are available and can be combined with existing survey data (Dang *et al.*, 2024).

Pros and cons

Given the advantages of imputation, why has it not been adopted more often and systematically to monitor SDG progress? Before addressing this question, it is useful to briefly compare the advantages and disadvantages of obtaining poverty estimates by implementing a new household survey versus using imputation methods. Table 1 shows that a household survey generally offers more accurate estimates than imputation methods, since the former obtains estimates using actual data (direct estimates), while the latter must rely on modeling assumptions to provide estimates. (Yet, while this is the general rule, it is a well-known statistical result that model-based estimates can have smaller standard errors than those of direct estimates, provided the modeling assumptions are correct or the imputation samples are larger). Analyzing household surveys requires less analytical capacity than implementing imputation, but fielding household surveys entails greater logistical efforts, more personnel, and longer data collection time. All in all, implementing a survey is much more expensive.

However, household surveys do not typically capture certain vulnerable population groups such as refugees, nor are they representative at lower administrative levels. But imputation can help address these data gaps to some extent (as discussed with the examples above). In fact, for certain situations, such as collecting data in conflict (or politically sensitive) environments, imputation may be the only viable option. As an example, since the latest poverty data for South Sudan were collected 15 years ago, the World Bank imputes poverty for this country based on more recent household survey data—which contain less information than needed to measure poverty directly but include the relevant poverty predictor variables. Another recent example is India, where household consumption survey data have not been released for over 12 years. Imputed estimates based on a recent survey conducted by a private firm offer some preliminary evidence that poverty has declined in India, but not by as much as previously expected (Roy & van der Weide, 2024).

Coming back to the question above, a key reason can be due to more trust in household surveys, since these data sources represent the traditional, dominant, and much more familiar mode of data collection. Put differently, most directors of national statistical offices (NSOs) would see fewer political risks with implementing a new survey—despite the logistic, personnel, and cost challenges discussed above—than implementing imputation. Another impediment is that imputation requires stronger analytical capacity, which is not readily available with many poorer countries' NSOs. Finally, some resistance to imputation methods might also come from (older) statistical staffs, who may fear losing their jobs if there is less demand for data collection.

Reflections on the ways forward

We do not propose that imputation should replace actual data collection; rather, collecting data will remain a crucial input for measuring SDG progress and, more broadly, for monitoring global wellbeing. But combining actual data with imputation methods can fill gaps with both data availability and data quality, resulting in improved timeliness and spatial refinement for our indicators of progress. The ability of imputation models to fill data gaps, however, will always require some form of up-to-date, high-quality data to train (and re-train) the imputation models. Old training data will typically fail to account for critical changes of behavior, and outdated outcome-predictor relationships can form the basis for erroneous imputation.

While it is widely recognized that more data are needed to monitor the SDGs, there is still significant resistance to using imputation methods for filling data gaps. Efforts to change the perception toward alternative ways to generating data such as imputation would go a long way toward addressing data challenges. Specifically, at the global level, more coordinated efforts to systematically employ imputation can lead to clearer, feasible plans for monitoring SDG progress. Such plans can explicitly require, for instance, conducting a new survey *every few years* and conducting imputation for *the intervening years*.

This can help produce better results in a more transparent manner. This also offers a viable alternative to the unrealistic goal of requesting all the countries to implement new surveys *every single year* to track poverty in all its forms everywhere (as currently interpreted with the first SDG). The challenge of analytical capacity can be addressed if experts from universities or international organizations can be matched up with local staffs to improve local analytical capacity (and this can just cost a fraction of a new household survey). The World Bank's recent addition of imputed poverty estimates to its global poverty database (with clear data notes for imputed numbers) can present a step in the right direction.

Some caveats are in order. First, the best and most cost-effective approach would be highly context-specific and depends on numerous factors such as available budget, logistical capacity, and timeline. There is no one-size-fits-all solution. In some contexts, we need to collect more data. In other contexts, employing imputation is better. In fact, a hybrid approach exists that combines data collection with imputation. This involves a survey with two components: a smaller sample with full consumption data and a larger sample without consumption data. The data from the smaller sample can be used to build an imputation model that is subsequently applied to the larger sample without consumption data. This hybrid approach could offer cost savings and less logistic burden (compared to a full consumption survey) and more accuracy (compared to a pure imputation study).

Second, imputation modeling assumptions should be properly vetted, and clear documentation of the estimation process and standard errors should be provided. Since imputation-based estimates present our second-best option in the absence of actual survey data, they should also be updated with direct estimates using a household survey whenever the latter is available. After all, imputation models must be built on good data inputs. Such inputs would require the establishment of a robust data system that periodically produces new data on a longer-term, sustainable basis.

Finally, our proposal to use imputation to address data gaps aligns with recent calls to utilize promising new tools, such as citizen science or artificial intelligence, to tackle these gaps. However, unlike these tools—where various concerns ranging from ethics to regulations and information bias should perhaps be carefully considered before their applications (Vinuesa *et al.*, 2020; Fraisl *et al.*, 2022)—imputation is different. It is based on statistical methods that allow for application at scale. Against the existing framework that unrealistically requires annual poverty data that hardly any poorer country can afford, a judicious use of this tool could help us make crucial, even if incremental, progress with measuring SDG performance.

References

Altındağ, O., O'Connell, S. D., Şaşmaz, A., Balcıoğlu, Z., Cadoni, P., Jerneck, M., & Foong, A. K. (2021). Targeting humanitarian aid using administrative data: Model design and validation. *Journal of Development Economics*, *148*, 102564.

Beltramo, T., Dang, H. A., Sarr, I., & Verme, P. (2024). Estimating poverty among refugee populations: a cross-survey imputation exercise for Chad. *Oxford Development Studies*, 52(1): 94-113.

Dang, H. A., & Lanjouw, P. (2023) "Regression-based Imputation for Poverty Measurement in Data Scarce Settings". In Jacques Silber. (Eds.). *Handbook of Research on Measuring Poverty and Deprivation*. Edward Elgar Press.

Dang, H. A., & Serajuddin, U. (2020). Tracking the sustainable development goals: Emerging measurement challenges and further reflections. *World Development*, *127*, 104570.

Dang, H. A., Hallegatte, S., & Trinh, T. A. (2024). Does Global Warming Worsen Poverty and Inequality? *Journal of Economic Surveys*, 38(5), 1873-1905.

Dang, H. A., Pullinger, J., Serajuddin, U., & Stacy, B. (2023). Statistical performance indicators and index—a new tool to measure country statistical capacity. *Scientific Data*, *10*(1), 146.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, *71*(1), 355-364.

Fraisl, D., Hager, G., Bedessem, B., Gold, M., Hsing, P. Y., Danielsen, F., ... & Haklay, M. (2022). Citizen science in environmental and ecological sciences. *Nature Reviews Methods Primers*, *2*(1), 64.

MacFeely, S., & Nastav, B. (2019). "You say you want a [data] revolution": A proposal to use unofficial statistics for the SDG Global Indicator Framework. *Statistical Journal of the IAOS*, *35*(3), 309-327.

Roy, S. S., & van der Weide, R. (2024). Poverty in India Has Declined over the Last Decade But Not As Much As Previously Thought. *Journal of Development Economics*, 103386.

Sachs, J.D., Lafortune, G., Fuller, G., Drumm, E. (2023). *Implementing the SDG Stimulus. Sustainable Development Report 2023*. Dublin: Dublin University Press.

Smythe, I. S., & Blumenstock, J. E. (2022). Geographic microtargeting of social assistance with high-resolution poverty maps. *Proceedings of the National Academy of Sciences*, *119*(32), e2120025119.

UN Women. (2023). *Progress on the Sustainable Development Goals: The gender snapshot 2023*. https://www.unwomen.org/sites/default/files/2023-09/progress-on-the-sustainable-development-goals-the-gender-snapshot-2023-en.pdf

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., ... & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, *11*(1), 1-10.

World Bank. (2021). *World Development Report 2021: Data for Better Lives*. World Bank. https://wdr2021.worldbank.org/

Zhao, L., Cao, B., Borghi, E., Chatterji, S., Garcia-Saiso, S., Rashidian, A., ... & Asma, S. (2022). Data gaps towards health development goals, 47 low-and middle-income countries. *Bulletin of the World Health Organization*, *100*(1), 40.

Table 1: Pros and cons of implementing a household survey vs. data imputation

No	Characteristics	Household survey	Imputation
1	Time in use	Traditional data generation approach	New data generation approach
2	Measurement accuracy	More accurate	Less accurate
3	Analytical capacity	Less demanding	More demanding
4	Logistics/ survey capacity	More demanding	Less demanding
5	Costs	More expensive	Less expensive
6	Timeline	Longer time to complete	Shorter time to complete
7	Current perception	More favorable	Less favorable
8	Vulnerable population groups (e.g., refugees)	Do not typically capture	Can provide estimates, subject to certain data conditions
9	Estimates at lower administrative levels (e.g., poverty maps)	Not representative at lower levels	Can provide estimates, subject to certain data conditions
10	Feasibility in challenging contexts (e.g., political sensitive situations, recent disaster-struck areas)	More difficult or impossible to implement	Possible to implement, subject to certain data conditions

Note: Traditional data collection methods typically involve implementing a household survey. New data generation methods include combination of existing data sources to generate new imputation-based estimates.