International Union for the Scientific Study of Population ■ IUSSP
Union internationale pour l'étude scientifique de la population ■ UIESP

Defining and successfully accomplishing the Data Revolution: The perspective of Demographers

International Union for the Scientific Study of Population
15 October 2014

1. Introduction

The International Union for the Scientific Study of Population (IUSSP) convened a meeting of 22 prominent demographers from both developing and developed worlds in Paris from 9-10 October 2014 to discuss how demographers, and demographic skills, could contribute to the Data Revolution.

In framing our deliberations, we recognised that the Data Revolution is not an end in itself. It is a mechanism to improve the lives of people over the next decade and beyond. We believe that the likely success of the post-2015 development agenda will be enhanced through plans and strategies that are evidence-led and informed by better data, and in turn lead to better policy options.

Demography has a distinguished history. The special skill of demographers lies in our ability to understand the systemic linkages between human population stocks and flows across space and time. This deep understanding puts demographers in a strong position to evaluate what is feasible and realistic with data collected on human populations, the limitations of those data, and the validity of the results. The assessment and evaluation of data quality, the capacity to link and process data from a multiplicity of disparate sources, and the ability to see the data as part of a larger systemic framework, are central aspects of this skill.

With specific reference to the Data Revolution, demographers are well-placed to appraise indicators – such as the population-related indicators for the Sustainable Development Goals, SDGs – that are proposed to track progress towards the final desired outcomes, and to ensure that they are coherent, valid, and operationalisable. The numerators and/or denominators of many of the other SDG indicators will also reflect population data, and the suite of tools that demographers routinely use to produce more reliable estimates and projections of population variables from limited, deficient and defective data can readily be pressed into service of the SDGs.

Understanding **data quality** and **interoperability** is a crucial aspect of the process of producing, analysing, and interpreting data to meet the Data Revolution's end goal. This briefing identifies the possible unique contributions to the Data Revolution of the global community of demographers and population scientists. We draw particular attention to two linkages. The first is the role that demographers could play in ensuring that the SDGs are measurable, valid and useful. The second relates to insights into broader dimensions of the Data Revolution from a demographer’s perspective.

We urge those organisations involved in framing the Data Revolution and the SDGs to:

- identify points of intervention where demographers can contribute to the post-2015 development agenda, particularly those related to the issues raised in the sections that follow;
- recommend significant investments in methods and training; and
- consider forming oversight groups or high level panels to investigate further the issues raised here.

2. Ensuring that the SDGs are measurable, valid, and useful

The likelihood of success of the Data Revolution and the post-2015 development agenda could be improved by involving demographers in the development and evaluation of proposed indicators that will be used to measure progress against SDGs, especially in regard to the five points:

a. Point estimates and uncertainty

Demographers ask that the framers of the SDGs maintain a keen awareness of the limitations of point estimates as indicators of development. While point estimates provide a headline figure, the substantial uncertainty that often surrounds those estimates must be articulated.

Furthermore, where highly disaggregated indicators are produced (as implied by the ‘no-one left behind’ principle) and estimates are based on relatively small populations, it follows that the degree of uncertainty will be commensurately greater. Trends drawn from a series of point estimates of these indicators may be fundamentally misleading, as the observed trend may often fall within the margin of error. As such, we propose the more immediate involvement of demographers in constructing guidelines for disaggregation, as they are experienced in working with small samples of data that represent small groups.

b. Goals, indicators and trade-offs

Demographers are concerned that the technical requirements for assessing risks vis-à-vis specific indicators could outweigh the benefits of measurement. The strong emphasis placed on collecting data for a series of SDG-related indicators may effectively de-emphasize other very important information, such as the underlying determinants of those indicators or phenomena that are not readily defined in terms of goals. For example, a target of reducing maternal mortality as measured by an indicator may see the indicator being conflated with the goal. Should this indicator be disaggregated as envisaged, given the difficulties involved in obtaining adequate data, the main focus of work may be on simply measuring levels of mortality risk. With limited resources, there may be a trade-off between collecting information to *measure* indicators, and information required to *assess its causal determinants* – here, the information that is directly useful for the development of appropriate policies and strategies to avert maternal deaths. Some other phenomena of great relevance to improvements in human welfare are difficult to define in terms of goals, and they too risk being neglected; population growth and age structure (which relate directly to investments in human capital, economic development and environmental sustainability) and migration are two such phenomena.

Again we recommend a greater involvement of demographers who are aware of the difficulties of data collection around subjects like maternal mortality and aware of the financial and opportunity costs.

c. Empirical vs. model-based estimates

Every effort must be made to ensure that the SDG indicators are based on empirical data, as opposed to being overly dependent on model-based estimates. While demographic, mathematical and statistical models are certainly useful in some cases, significant risks are associated with over-dependence on model-based estimates. First, regardless of the care taken in constructing such models, there is always a sizable risk that they may be wrong, biased, or incorrectly parameterised.

Second, the results that are produced by those models reflect, to a significant degree, the assumptions used in constructing and parameterising the models. Where model-based estimates are required (e.g. of annual population estimates in intercensal years), care should be taken to ensure that they are robust as possible, and that the modelling approach and all underlying assumptions be made transparent.

Within the context of producing reliable indicators, the reliability of estimates used to produce the denominators is of paramount importance. Empirical data on population size is obtained usually only every ten years, with the conduct of a census. The issue that arises, then, is how best to project population size and composition in intercensal periods. This, too, is an area where demographers have substantial specific expertise, and this knowledge and skill should be harnessed in the measurement of SDG indicators.

d. The importance of strengthening national vital statistics and census data

If model-based estimates are to be avoided, there is a concomitant need to strengthen the national statistics systems (NSSs) that will provide much of the data used for the SDG indicators. Improving the timeliness and completeness of national vital registration data is already acknowledged as an urgent priority across the developing world.

While the community of demographers and population scientists endorse this goal entirely, we are concerned that similar attention needs to be paid to improving the quality and coverage of census data. Not only will census data provide the baseline numbers for many indicators, especially at levels of fine granularity (e.g. small areas, or when disaggregated by factors other than those collected as part of the vital registration system), these data also usually form the basis for drawing nationally representative sampling frames for a great many surveys and other data collection exercises (including handling selectivity in big data). Additionally, the need for detailed documentation and metadata on the construction of and basis for calculating the SDG indicators must not be neglected.

Engaging the community of demographers and population scientists in the design and construction of indicators will go some way to helping to ensure that the SDG indicators are coherent and consistent. Demographers may also be able to advise on the variety of methods and approaches routinely used to estimate population-level indicators from limited and defective data.

e. Defining regions

Finally, while demographers appreciate the need for measurements to be produced for global regions, we would urge a greater degree of caution in their construction and interpretation. First, a standard set of regional definitions should be used, to avoid the problem associated with different organisations classifying countries differently. Second, users of regional statistics and indicators must be alert to the composition of those regional statistics, which – typically – will be population-weighted aggregates. Thus, if a region is comprised of one country with a sizeable population and several others with proportionally smaller populations, the regional statistics will mostly reflect those of the most populous country. In such situations, the danger of applying a regional indicator equally to all countries in the region is self-evident.

3. Insights into the Data Revolution from a demographer's perspective

Beyond the specific recommendations outlined above on the SDGs and the process of developing measurable indicators, participants at the Paris meeting reflected on several broader aspects of the Data Revolution. A successful Data Revolution would engage experts to address, and find solutions for, the following issues that are central to population sciences:

a. Data quality

If the Data Revolution is to succeed, the quality of the data collected is of prime importance. Under this heading, demographers have identified the following concerns.

i. The scientific basis for data quality

A key concern among the demographers consulted was for greater emphasis to be placed on developing a scientific basis for assessing and measuring data quality. Specifically, attention should be given to the methodological difficulties of integrating 'new' forms of data (e.g. big data, administrative data, satellite data, etc.) with extant 'old' forms, including the issues of sample selectivity at lower levels of geographic disaggregation and uncertainty. Demographers would be well placed to help identify opportunities, limitations, and potential scientific backing for these data.

Specifically, it is crucial to ensure that all data are validated before use; while this is often done to some degree with 'old' forms of data, we are unconvinced that sufficient or equal attention is being paid to the same aspects of 'new' data. This validation should be both internal (i.e. ensuring that the data are internally coherent and consistent) and external (i.e. in comparison to other similar data). Further, routinely and readily available data documentation and metadata must be made available to assist in that validation. This should include detailed information on the processes employed in, and effects on the data of, data cleaning, editing and manipulation (e.g. through static or dynamic imputation).

One potential avenue by which a Data Revolution could address these concerns would be to establish a **common system of data quality scoring**, based on independent and impartial evaluation, and which takes into account the uncertainty associated with estimates derived from the data. For example, if an indicator can be estimated from several different data sets, researchers and statisticians would be able to assess its reliability, allowing its prudent use and the calculation of a more precise composite indicator if needed.

Finally, it is important to recognize that most indicators will require some understanding of population estimates, particularly for estimating denominators; and providing grounded projections is where demographers come into their own. Indeed, many of the common tools of demographers are of real value to the SDGs beyond the narrow set of strictly "demographic" variables. One example would be efforts that aim to increase school enrolment rates over time, which must be based on projections of school-aged populations and patterns of school abandonment - an issue best examined with life table methods.

ii. Representative indicators at the level of finest granularity

In abstract terms, we understand and support the necessity of having detailed and valid SDG indicators at levels of fine granularity as implied by the principle that disaggregated estimates be tracked routinely. Practically, however, we emphasise that a compromise will have to be reached between what is desired, and what is achievable without affecting the robustness of the estimates so derived. No "one-size-fits-all" approach can be determined: indicators of comparatively rare events or outcomes (for example, maternal mortality) will become unusable at finer levels of granularity than indicators of relatively common events or outcomes (for example, infant mortality).

Demographers can help in determining the levels of granularity appropriate for each indicator eventually selected.

iii. Making data available

As part of the Data Revolution, and allied with the goals of allowing data to be used to hold systems of authority to account and encouraging transparency, all data used in the determination of indicators should be released in a timely and regular manner, with clear standards specifying benchmarks for different kinds of data. Datasets should be accompanied by appropriate metadata about the data themselves, as well as about the process whereby the released data were produced; and they should be encoded and distributed in standardised format to facilitate usage (for example, the Data Documentation Initiative (DDI) metadata standards for microdata in the social, behavioural and economic fields).

b. Interoperability

A successful Data Revolution will ensure, as far as possible, adherence to principles and standards of interoperability. We have identified three issues under this theme.

i. Open and tiered access to data

In order to ensure both data openness and appropriate considerations of confidentiality, one approach to data availability is tiered access to data. Under a tiered system, different amounts of data will need to be made available in order to calculate the indicators: the more disaggregated the indicator, the greater the volume of data required. For example, in some instances, a 10 per cent public use microsample data set will be sufficient; for others, a complete data set will be required. Where legitimate concerns about confidentiality can be advanced, alternative mechanisms of gaining access to the data must be provided. For example, secure data centres located in host countries could provide access to complete data when necessary, or efforts could be made to mask the identity of individuals and households (e.g., the DHSs makes small changes to GPS readings to make it impossible to identify specific households).

ii. Data at fine levels of granularity

Data used to provide estimates of indicators, or used as the basis for policy interventions, at fine levels of granularity pose specific problems. The first is that, in many cases, the data from a single source (e.g. a survey) may not be robust or reliable at fine levels of granularity. The limitations of the data must always be borne in mind. Frequently, estimates and indicators at this level will have to be supplemented with harmonized and linked data from other sources (including perhaps 'big data').

Methods for doing so in ways that are reliable and valid still need to be developed and tested. Similarly, appropriate standards of geocoding (including decisions as to the accuracy of geocoded co- ordinates) need to be defined. Where metrics involve some measure of distance (e.g. from a tap, a clinic, or a school) high levels of accuracy are required to produce meaningful results. Having geocoded data will also facilitate the task of linking data sets together, and provide information on localities that is essential for assessing the effects of policy or interventions. In turn, this may have implications for the confidentiality of records in the data set. How to straddle this tension requires careful consideration. These are all debates to which demographers could contribute meaningfully.

iii. Calibration of 'new' data forms

Data collected using new methods and technologies, including 'big data', must be calibrated before being merged with existing data, to ensure that – as far as is possible – any biases (e.g., sample selectivity) in those data are removed prior to their incorporation. While the evaluation of those data has been covered in earlier sections, the focus of the calibration exercise should be on the sampling, sampling frames and smoothing of these data – activities for which census data are typically most appropriate. Again, new standards, methods and techniques will have to be developed to ensure that this is done as carefully as possible

4. Ensuring institutional capacity

In the discussions of the Data Revolution that we have encountered to date, the roles of the National Statistical System (NSS) and the National Statistical Offices (NSOs) have been afforded particular attention. However, if the NSS is to be regarded as a central component of a Data Revolution, it follows that appropriate skills and expertise must exist within the NSS (and especially in the NSOs) to collect, assess and analyse the data. In this regard, the IUSSP has been concerned for some time now at the erosion of demographic skills and knowledge within NSSs and NSOs. The time has come for significant interventions to upgrade these skills, capabilities and knowledge by investing in training of NSS and NSO staff (including in core demographic methods and theory) and by developing effective strategies to ensure that well-trained staff are retained within these bodies.

5. Conclusions

Of all the social sciences, demography is the discipline that is most centrally focused on issues of data quality and how to make the most with limited and imperfect data. The structured and systematic demographic approach, along with the battery of well-tested methods that demographers have developed over the years to measure and analyse the various dimensions of population stocks and flows (e.g., health and mortality, fertility and reproductive health, migration and immigrant integration, population growth and age structure, which affect economic growth prospects and the environment), are of clear value both to appraising the proposed SDG indicators and to their measurement. From our perspective, demography has much to contribute, especially in the areas of data quality and interoperability, and we hope that population scientists will be called upon to play prominent roles in the Data Revolution as activities move forward.

Acknowledgements

This briefing document is a synopsis of the outcomes of the IUSSP meeting on 'Demography and the Post-2015 Data Revolution' held in Paris, France from 9-10 October 2014. It was prepared by Tom Moultrie (University of Cape Town; tom.moultrie@uct.ac.za), Tom LeGrand (Université de Montréal; tk.legrand@umontreal.ca) and Emma Samman (Overseas Development Institute, UK; e.samman@odi.org.uk). This meeting was undertaken with the support of the William and Flora Hewlett Foundation.