Subnational Variations in the Quality of Population Health Data: A Geospatial Analysis of Household Surveys in Africa

Valentin Seidler[a], Edson C. Utazi[b], Amelia B. Finaret[c,f*], Sebastian Luckeneder[d], Gregor Zens[e] , Maksym Bodarenko[b], Abigail W. Smith[f], Sarah E.K. Bradley[g], Andrew J. Tatem[b], Patrick Webb[h]

[a]Vienna University Department of Economics and Business; Vienna, Austria; [b]University of Southampton, WorldPop, Southampton, UK; [c]University of Edinburgh, Global Academy of Agriculture and Food Systems, Edinburgh, Scotland, UK; [d]Vienna University Department of Economics and Business, Department of Socioeconomics, Vienna, Austria; [e]International Institute for Applied Systems Analysis (IIASA); [f]Allegheny College, Department of Global Health Studies, Meadville, Pennsylvania, USA; [g]Independent Public Health Demographer; [h]Tufts University, Friedman School of Nutrition Science and Policy, Boston, Massachusetts, USA; and [h]International Institute for Applied Systems Analysis (IIASA).

Corresponding Author information: *Amelia B. Finaret
Email: afinaret@exseed.ed.ac.uk

Abstract

Background

In many low- and middle-income countries, household survey data help address health and development challenges and track achievements towards national objectives including the Sustainable Development Goals (SDGs). Such data are widely used and trusted. Yet users often lack critical information about the extent of data errors where it matters most for human wellbeing – at the district level, where health interventions are usually implemented. To assess the magnitude of such data problems, this study estimates the extent and types of errors in nationally representative household survey data from 33 African countries.

Methods

We conducted a comprehensive high-resolution geospatial analysis of household survey data from the most recent surveys of 33 countries across Africa between 2006 and 2019, using publicly available data from the Demographic and Health Surveys (DHS). We first calculated the prevalence of data errors by survey locations and then employed Bayesian model-based geostatistics using spatially explicit DHS data and covariates from gridded high-resolution datasets. Our model produced 5 × 5-km gridded estimates of three widely used health data quality indicators: age heaping, incomplete age records of interviewed women and biologically implausible height-for-age (HAZ) measures.

Findings

We report two important findings. First, the distribution of errors in survey data across and within Africa was systematic. Errors increased with remoteness. Second, moving beyond the DHS survey locations, our model found substantial heterogeneity in the distribution of errors on subnational levels. For example, the share of incomplete information of women's age in Chad (national mean 66·1%) ranged from 91·8% (sd 2·5%) in southern Chad to only 6·8% (sd 2·2%) near the eastern border with Sudan.

Interpretation

This is the first study to estimate the subnational distribution of errors in household survey data at a high spatial resolution. Survey data quality degrades with increased remoteness, a phenomenon that adds to the vulnerability of remote populations. Our results illustrate the magnitude of data errors, contribute to SDG target 17.18 on reliable data availability, and promote better targeting of health interventions and data collection efforts within countries.

Funding

Austrian National Bank's Anniversary Fund and Feed the Future Food Systems for Innovation Lab, funded by the United States Agency for International Development, and the University of Edinburgh.

Role of the funding source

VS is supported by the Austrian National Bank's Anniversary Fund Grant No. 18157. PW would like to acknowledge the support of Feed the Future Food Systems for Innovation Lab, funded by the United States Agency for International Development, Cooperative Agreement No. 7200AA21LE00001. Amelia Finaret is supported by the University of Edinburgh. Funding sources had no role in the development of this study, the writing, or the decision to submit for publication. Authors were not precluded from accessing data in the study and they accept responsibility for submission.

Research in context

Evidence before this study

Lacking reliable and representative census or birth registration systems, many low-income and middle-income countries rely on multi-purpose household surveys for information on population health. Demographic and Health Survey data, the main example used by this study, are widely trusted and used, and their value critically depends on their high quality. We identified around 6,000 publications using DHS data since 2010. While these household survey datasets are widely used, their value critically depends on their quality. DHS working papers and methodological reports, as well as studies by independent researchers, have assessed specific issues relating to data quality such as inter-interviewer variability effects and questionnaire length. Other studies have estimated data quality for specific health conditions such as HIV or Malaria. We found no studies analyzing the extent and the magnitude of errors in population health data on a sub-national level across or within low-income countries.

Added value of this study

This study is the first sub-national geospatial analysis of the quality of household survey data for many low-income countries. We calculated the prevalence of data errors by survey location in the most DHS recent surveys for 33 African countries and then estimated the geographic distribution of three widely used health-related data quality indicators at a 5 x 5 km resolution. We found a striking large-scale, subnational heterogeneity in data quality and highlighted geographic locations of particularly low data quality. The unexpectedly large magnitude of data errors will raise awareness among thousands of data users who work with survey data every day. Evidence on areas of particularly low data quality is an important contribution to the continuous improvement efforts of household survey programs in Africa. This is the first study that explores whether the data errors are distributed randomly or whether there is a systematic bias in the distribution of household survey data quality in Africa. We found that geographically remote populations in rural Africa appeared to be particularly underserved and underrepresented by household-level survey data.

Implications of all the available evidence

The DHS program counts as gold standard among periodic household surveys and is widely trusted. While survey experts or people involved in field enumeration may have an intuitive understanding that, for example, age-related information is hard to collect among remote populations, the exact magnitude of the problem is unknown. A large community of data users, researchers, and health professionals lack the specific knowledge of experts involved in enumeration but still rely on household survey data for conducting research or planning health interventions. Our study aims to raise awareness among this group. For this reason, we offer an easily accessible online tool providing high-resolution information on the predicted data quality on subnational levels. Our results allow local, national, and global decision makers to better understand the extent of data quality issues in household surveys. Because illness and malnutrition are defined as measures lying outside a defined healthy range, even modest inaccuracies result in misreported prevalence of conditions such as child stunting or age at first pregnancy for women and girls. More research and resources are needed to inform policies and health interventions in geographic areas where health-related household survey data are less reliable.

Introduction

Lacking reliable and representative census, public health information, or birth registration systems, many low-income countries use multi-purpose household surveys for information on population well-being.[1–3] The Demographic and Health Surveys (DHS), the Multiple Indicator Cluster Surveys (MICS), and the Living Standards Measurement Surveys (LSMS) form the backbone of information on age, fertility, mortality, health, wealth, education, and nutrition and are most commonly accessed by practitioners and the research community.[4] DHS data, the main example of this study, have been used in close to 6,000 publications since 2010 and this is a lower bound estimate of data utilization.[5] Estimating the geographic distribution of key health and development indicators on a granular subnational level is now a well-established practice. High-resolution maps of health or nutrition indicators can draw attention to vulnerable populations and help guide future interventions.[6–9]

Statistical modeling may partially account for errors or gaps in the survey data, but the validity and reliability of survey-based analyses and the policy makers that use them depend on the quality of the underlying data.[10] Data gaps and errors may be introduced at any stage of the household survey beginning with survey design[11], during data collection or data processing.[12] The potential statistical implications of low data quality are manifold and include bias in national and regional estimates as well as imprecise or misleading regression analysis due to systematic patterns of missingness. For example, illness and malnutrition are typically defined as measures lying outside a healthy range[13]. Hence, even modest inaccuracies can more than double the reported prevalence of conditions such as malnutrition in a given country, potentially resulting in less efficient allocation of resources.[14,15]

High-quality multi-topic household survey data are necessary to track the achievement of a wide range of the United Nations (UN) development goals.[9,16] Sustainable Development Goal (SDG) 17.18 explicitly aims at increasing “the availability of high-quality, timely and reliable data”.[17] In response, designers of household surveys have worked hard to identify data errors and to deliver high data quality[6,18–22] and undergo continual improvement[13,23–25]. However, we still lack knowledge about the spatial distribution and magnitudes of data errors where it matters most for human wellbeing - on subnational and district levels, where health and development interventions are implemented. Such knowledge can raise awareness among data users who are not involved in ongoing improving efforts but still want to work with household survey data. In addition, knowledge about geographic areas of low data quality can contribute to improved future survey rounds, for example by assigning more experienced teams to these localities and by providing them with more time, resources, or better equipment.[6] Finally, many data users are currently unaware of potential systematic patterns and magnitudes of the distribution of errors within countries. Examining such patterns which threaten analytical work and challenge the planning and implementation of health interventions is essential.

We aimed to investigate the magnitudes of data errors in household surveys across and within countries, and the extent to which the distribution of errors followed systematic patterns. To provide a comprehensive analysis we mapped the extent of the problem using high-resolution estimates of the distribution of data errors in 33 African countries, and then explored patterns in the prevalence of data errors by survey location using the most recent DHS household surveys for the included countries between 2009 and 2016.

Methods

Sampling errors and measurement errors

All nationally representative surveys aim for high data quality that seeks to minimize both sampling (selection) errors and non-sampling (measurement) errors. Sampling errors may occur if sample sizes are too small or if sampling frames such as census maps are incomplete or outdated.[26] Sampling errors are generally considered the relatively smaller threat to accuracy and inference, because large household surveys use probability designs and because sampling errors can often be statistically evaluated and corrected.[26–28] Measurement errors and missing information are non-sampling errors and are a more substantial concern for interpreting survey data. Non-sampling errors can occur despite rigorous efforts and enumerator training. For example, respondents may not be present or may refuse to answer a question (non-response bias), or parents may be unable to recall the birth dates of their children (information bias).[12,14,29] Procedural challenges related with the organization of household surveys include the rotation and training of fieldworkers, fieldworker fatigue, imperfect measurement equipment, language barriers, or challenges related to data cleaning.

Health professionals, who plan and execute public health programs, and researchers, who work with health and nutrition survey data, will want to know whether measurement errors are randomly distributed in the data or whether there is systematic (non-random) bias. Randomly distributed measurement errors attenuate estimates towards the null hypothesis, which is generally considered an acceptable effect on results. Non-randomly distributed errors are more problematic because the direction and the magnitude of the biases are often unknowable and because of the paucity of strategies to address the bias without additional data.[30]

If agencies, funding bodies, pre-analysis plan repositories, or journals have specific requirements or recommendations for statistical power, a systematic distribution of data errors potentially negates the possibility of using data from particularly affected regions, or from sub-samples of the population in these regions.[31] Without precise and localized information on the magnitudes and distributions of measurement errors, scholars and practitioners usually assume high survey data quality or they assume a relatively innocuous, random distribution of data errors.[15] This may distort our understanding of health and economic conditions particularly at local levels, where policies and interventions are planned and implemented.[32]

We know that data quality may be reduced for more remote populations.[33–35] However, we found no empirical studies documenting the extent or magnitude of this problem for population health data. Studies so far analyzed data quality only at the levels of whole countries or survey rounds[18,21], across large areas such as provinces, or between specific populations such as rural and urban populations.[11,23,36–39] Knowledge on data quality is therefore incomplete because interventions are often made locally or on the level of districts and should ideally be based on reliable information from the same locality. Even though information collected in household surveys are now geolocated, there is no study estimating the variation in survey data quality or magnitudes of errors at district levels or village levels for a whole region or continent.

We conducted a comprehensive high-resolution geospatial analysis of DHS data quality taken from the most recent surveys of 33 countries in central, eastern, western, and southern Africa between 2006 and 2019. Among large household surveys around the world, the DHS offers the most complete data for these regions. Result 2 directly calculated the prevalence of data errors in the most recent DHS survey round. Results 1 and 3 moved beyond the DHS survey locations and employed Bayesian model-based geostatistics, leveraging covariates from gridded high-resolution datasets to predict data quality beyond surveyed areas. Our model produced 5 × 5-km gridded estimates (roughly equivalent to villages in rural areas) of three widely used data quality indicators, which we aggregated to district (also called 'admin-2') levels and national levels using population weights.

Data quality indicators

We constructed three indicators of data quality using variables which are frequently used to assess health and well-being of women and children (appendix). The first indicator is 'incomplete age,' referring to the share of interviewed women (15-49 years) with either the year or the month of birth missing relative to all interviewed women. The second indicator is 'Age heaping,' referring to the proportion of reported ages ending in 5 or 0 of all adults between 23 and 62 (based on Whipple's index definition). A value above 20% indicates age heaping. The third indicator is 'Flagged HAZ,' referring to missing or biologically implausible values for the measured heights of children (height-for-age z-scores) according to WHO (World Health Organization) standards. These indicators were chosen as they refer to both the quality of the demographic and health (anthropometric) information in the survey data, arguably the most important dimensions for practitioners that work with DHS data. We found little correlation between the three data quality indicators themselves (figure S3 in the appendix). In addition, and in contrast to other data quality indices, the chosen indicators are expected to work well at a high spatial resolution.

Data collection

DHS data are best suited for the scope of the study, because they are publicly available, geocoded, and more comprehensive in time and space than data from other large survey programs such as MICS or LSMS.[5,40,41] The number of DHS survey rounds varied across African countries. In countries with more than one survey round since 2006, we used the most recent survey (table S9 in appendix). We extracted relevant variables from the DHS children file, from the individual women's file and the household roster file In DHS data, the individual and household survey questionnaires are linked to a survey location (termed 'cluster' or 'primary sampling unit') that represents a set of neighboring households or a village structure. The data we use comes from 18,050 DHS clusters. We assembled relevant geospatial covariates for which high-resolution spatial data or layers were available. For calculating urban settlements, we used the nighttime light emissions dataset generated by Li and colleagues[42] as well as Satellite-based high-resolution settlement data from Marconcini and colleagues[43], and the Africapolis dataset[44] that includes larger urban agglomerations of at least 10,000 inhabitants. The detailed descriptions on data collection and preparations of the variables are documented in the appendix.

Geostatistical model

For results 1 and 3, each indicator was modelled in each of the four geographic regions as is standard in literature, and the outputs of the regional models were combined to form a continuous map for the indicator. We assigned countries to one of the four regions. This speeds up computation and leverages regional relationships between the indicators and the geospatial covariates. Our INLA-SPDE approach requires the specification of a fine triangulation mesh to approximate the spatial random effect. An example spherical mesh for West Africa is shown in figure S1 in the appendix. All the covariates are standardized to have a mean of zero and variance of one prior to model-fitting.

To predict the data quality indicators at 5 x 5-km resolution and the district level, we fit a Bayesian geostatistical model with a binomial likelihood:

$$Y\,(s_i)\,|\,m(s_i) \sim \text{Binomial}(m(s_i),\, p(s_i)), \qquad (1)$$

where $y(s_i)$ denotes the number of individuals possessing the attribute being modelled at spatial location s_i (represented using the longitude and latitude coordinates), $m(s_i)$ is the number of individuals sampled from that location and $p(s_i)$ is the underlying true proportion of individuals possessing the attribute being modelled, e.g., incomplete age, at location s_i. We model $p(s_i)$ using the logistic regression model

$$\text{logit}(p(s_i)) = x(s_i)^T\beta + \omega(s_i) + \epsilon(s_i) \qquad (2)$$

where $x(s_i)$ is a vector of covariate data associated with location s_i, β are the corresponding regression coefficients, $\epsilon(s_i)$ is an independent and identically distributed (iid) Gaussian random effect with variance σ_ϵ^2 used to model non-spatial residual variation, and $\omega(s_i)$ is a Gaussian spatial random effect used to capture residual spatial correlation in the model; i.e. $\omega = (\omega(s_1),\ldots,\omega(s_n))^{\mathrm{T}} \sim N(0, \Sigma_\omega)$. $\Sigma_\omega$is assumed to follow a Matérn covariance function.[45] To assess the performance of the fitted models for out-of-sample prediction, we adopt *k*-fold cross-validation, setting *k* = 10. Additional details on data collection and analysis including modeling, estimation, model validation and limitations can be found in the appendix. Our data visualization tool provides easy access our estimates used in results 1 and 3 (https://apps.worldpop.org/SSA/data_quality/).

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Result 1: Highly variable data quality at subnational levels

DHS data quality predictions obtained from our geostatistical model varied substantially across 5 x 5-km cell levels and across district levels throughout the 33 included African countries. Within-country variation in measurement errors was apparent with all three data quality indicators for all 33 included countries (figure 1). Within-country variation in data quality was larger in magnitude in countries with lower average data quality (figure S2 in the appendix). For example, our estimates for age heaping in Nigeria (national mean 37·9%) ranged from 58·9% (standard deviation (sd) 3·8%) in Guzamala district in north-east Borno to 23·3% (sd 2·4%) in Odogbolu near Lagos, where the terminal digit preference for 0 or 5 was just 3·3% above the expected natural occurrence of 20%. Estimates of the share of incomplete information of women's age in Chad (national mean 66·1%) ranged from 91·8% (sd 2·5%) in Loug Chari in the south of Chad to only 6·8% (sd 2·2%) in Dar Tama, on the eastern border with Sudan. Flagged HAZ values in Madagascar (national mean 16·3%) ranged from 41·7% (sd 3·8%) in Bongolava to 8·7% (sd 1·7%) in Haute Matsiatra, two regions in the center of the island.

These results demonstrate that analyzing measurement errors at the country-level alone masks important local and regional patterns. In some areas, pockets of comparable data quality spread across national borders. While this was partially a result of the supranational modeling framework we applied, it also suggests that data quality was driven by common contextual factors. For example, district-level estimates of age heaping in Mayo-Tsanaga (35·1% (sd 1·3%)) in northern Cameroon (national mean 29·4%) were comparable with those in Maiha (35·6% (sd 3·9%)) across the border in north-east Nigeria (national mean 37·9%) contributing to an apparent pattern of increased age heaping in districts bordering Lake Chad. Furthermore, little correlation between the three data quality indicators themselves (figure S3 in the appendix) resulted in distinct geographic patterns for each type of measurement error within countries. For example, incomplete age records reported from Wajir West 66·2% (sd 7·1%) in northern Kenya were considerably worse than the national mean of 13·9% while in the same district flagged HAZ values (1·6% (sd 0·7%)) were close to the Kenyan national mean of 1·3%.

Result 2: Deteriorating data quality in remote areas

The distribution of measurement errors across and within 33 countries in Africa was systematic. To explore potential systematic bias in the raw data, we combine the prevalence of data errors by survey location with distance to settlements, which we approximate using luminosity emitted at night. The predictions underlying figure 2 were based on NTL emissions of a specific intensity (Digital Number, DN 15), which lies above the threshold of the luminosity of public streetlights indicating a settlement with sufficient infrastructure for survey teams to restock equipment and rest. However, the measurement bias was similarly apparent with lower and brighter light emissions threshold levels (figures S5 and S6 in the appendix). More generally, the bias was discernible with distance to settlements (with or without access to electricity) using two alternative high-resolution settlement data sets built from daylight and radar satellite imagery and electoral records. These alternative data sets include smaller and unlit settlements with shorter geographical distances between them, which made this approach less informative relative to using NTL emissions (figures S7 and S8 in the appendix).[43,44] Finally, we found little evidence that the observed bias was based on one-time effects for the majority of the included countries with multiple survey rounds (figures S9 – S11 in the appendix).

Systematically distributed measurement errors have potentially major negative implications for research, particularly if the direction and the magnitude of the bias are unknown. In the 33 included countries, increasing distance from electrified settlements emitting NTL of DN 15 was associated with decreasing data quality across all regions to varying extents. The measurement bias was relatively stronger among countries in West Africa, and relatively weaker in Central and Southern Africa (figure S4 in the appendix). The bias affected all three data quality indicators to varying extents. The bias was relatively stronger with age heaping (figure 2 a) and incomplete age records of women (figure 2 b). It is relatively less apparent for flagged HAZ values (figure 2 c). For age heaping and incomplete age of women in particular, this association grew stronger within countries of medium to high overall levels of measurement errors (figures S12 - S14 in the appendix). For example, the estimated share of interviewed women (15-49 years) with either the year or month of birth missing (’incomplete age’) in Togo (national mean 18·5%) was 20·6% at 50 km distance, increasing to 25·4% at 100 km and was 37·0% at 200 km. The proportion of reported adult ages ending in 5 or 0 (’age heaping’) increased within Kenya (national mean 26·2%) from 25·9% at 50 km from the nearest town or nighttime light emitting area to 26·8% at 100 km and 28·5% at 200 km distance. The measurement bias was weakest within countries that had overall higher quality DHS data. For example, age heaping within South Africa (national mean 20·0%) was at its natural rate of 20·0% at 50 km distance to the nearest point of nighttime light of DN 15 value. It was 20·3% at 100 km distance and 21·0% at 200 km.

Result 3: Pairing measurement errors and sampling uncertainty

Figure 3 shows the spatial distribution of two of our estimates for data quality: ’incomplete age’ and ’flagged HAZ’ (in shades of blue) and pairs them with the standard deviations of predicted estimates of related public health indicators: ‘contraceptive use’ and ’stunted children’ (in shades of green). Standard deviations are commonly used as a measure for the statistical uncertainty of key public health measures, here the prevalence of contraceptive use among sexually active women (’contraceptive use’) and stunting prevalence among children (’stunted children’). High standard deviations are mainly a result of small sample sizes at the survey cluster level or from a distribution of survey locations that does not consider geostatistical design considerations. By pairing the sampling uncertainty of ’contraceptive use’ with ’incomplete age’ and of ’stunted children’ with ’flagged HAZ’, we assessed in which districts these two types of data errors overlap. We were, for example, curious to know to which extent high sampling uncertainty of stunting overlapped with biologically implausible height-to-age measurements taken in the same population. The combination of these indicators provides health practitioners and researchers with important information on DHS data quality and future surveys with insights into required investments needed to achieve quality standards in the hardest-to-reach populations.

Across and within the 33 included countries, we found little overlap and substantial variation between these error types. Little correlation between the ’incomplete age’ and ’flagged HAZ’ (figure S3 in the appendix) further increased this heterogeneity. Some regions were challenged with either high standard deviations of both, ’contraceptive use’ and ’stunting’ (for example in

Namibia, South Africa, Zimbabwe, parts of Zambia and southern Tanzania) or with high estimates of both, ’incomplete age’ and ’flagged HAZ’ values (for example in Angola, the Democratic Republic of Congo, and Chad). On the country level, measurement errors paired with large sampling uncertainty were widely present only in Madagascar (national mean ’incomplete age’ 21·22%, national mean ’flagged HAZ’ 16·28%) and in Niger (national mean ’incomplete age’ 87·04%, national mean ’flagged HAZ’ 12·14%). On the local level, single districts were challenged by both high error estimates and large standard deviations of estimates. N’gauma district in north-western Mozambique reported estimates for ’incomplete age’ of 21·35% and a standard deviation of ’contraceptive use’ at 2·97% (mean 14·68%). In the same district, the estimate of ’flagged HAZ’ was 8·56% and the standard deviation of ’stunted children’ was 4·11% (mean 45·49%). The Mozambiquan national means of ’incomplete age’ and ‘flagged HAZ’ were 4·88% and 5·23%, respectively. No single country exhibited low rates of errors and high statistical certainty nationwide. However, districts with high data quality appeared in clusters across the subcontinent. One example were districts around Bukkuyum (’incomplete age’ 0·35%, standard deviation of ’contraceptive use’ 0·62% (mean 1·42%) in north-western Nigeria (national mean of ’incomplete age’ 4·22%). Similarly, near error-free HAZ values and high statistical certainty concentrated in districts surrounding Thiès (’flagged HAZ’ 0·16%, standard deviation of ’stunted children’ 1·88% (mean 15·59%) in western Senegal (national mean of ’flagged HAZ’ 0·31%).

Discussion

This study provides the first status-quo quantification of household survey data quality at a 5 x 5-km spatial resolution in 33 countries in Africa for household surveys conducted between 2006 and 2019. Our estimates highlighted a striking variation in data quality at district level and below, which had so far been masked in cross-country or cross-survey studies. We also found that the prevalence of data errors by survey location was systematic, resulting in worse data quality for populations that live in remote rural areas. The bias did not substantially change between surveys for most of the included countries with multiple surveys, which indicated a potentially persistent problem beyond potential one-off effects of single survey rounds.

Why is this important?

The DHS program counts as gold standard among periodic household surveys and is widely trusted. While survey experts or people involved in field enumeration may have an intuitive understanding that, for example, age-related information is hard to collect among remote populations, the exact magnitude of the problem has been unknown. High-resolution mapping studies may or may not account for (randomly distributed) measurement errors,[7] but a large community of data users, who rely on household survey data for conducting research or planning health interventions, lack the specific knowledge of experts involved in enumeration. Lacking awareness of the magnitude of this problem, researchers, practitioners or funding agencies alike have so far assumed either sufficiently high survey data quality or a relatively innocuous, random distribution of errors.

Health and development efforts are typically implemented on community and district levels. Our estimates highlight districts and regions across Africa in which survey-based anthropometric and demographic information may be too uncertain or biased to support inference, policymaking or intervention.[48] Even though researchers and funding bodies have a professional obligation to effectively communicate the uncertainty around estimates, data errors often fail to catch the attention of policymakers, government agents or fellow researchers. Mapping estimates of measurement errors can improve the awareness of these users in their work with underlying anthropometric, health, and demographic data.

Our findings corroborate and expand the details needed to understand data quality challenges of the "last mile" problem in global public health. [33,35] Our analysis is based on a twofold empirical approach. Result 2 directly documents the increase in the prevalence of errors for sampled areas in more remote locations using raw DHS data. Our predictions in results 1 and 3 go beyond merely describing the data errors in the sampled areas. Predicting data errors for non-sampled locations permit better illustrating the extent of potential data quality challenges.

What can be done about this?

The specific statistical implications of low data quality are difficult to assess in a generalized fashion. For instance, shortcomings in certain dimensions of the data have the potential to introduce bias, while other aspects of data quality may merely decrease the precision of the empirical analysis.[15] As a consequence, statistical methods that may mitigate issues arising due to low data quality need to be evaluated on a case-by-case basis and will depend on the specific inputs and desired output of the respective empirical analysis.

Notably, our intention was to illustrate the magnitude of this problem, not discuss underlying causes. Our measures, such as nighttime light emissions, do include an economic and an accessibility dimension. We hope to motivate more research to understand the causes of heterogeneity in data quality. Household survey data quality may deteriorate because of several factors and specific interventions and more research will be needed to sufficiently address them. Examples are the lack of birth registration, lower access to education[33], migration[46], or higher levels of poverty.[47] Finally and more generally, survey enumeration in remote communities that use languages other than those native to enumerators, are hard to access, or prone to of civil unrest is more demanding due to longer trips, less rest and higher risks to personal safety.

Limitations of this study

Both the data used in the analysis and our methods are subject to several limitations. First, spatial data gaps between DHS survey locations may introduce varying degrees of uncertainty in our estimates. These considerations also apply – to a lesser extent – to the aggregated analyses on the sub-national and are relevant to any geolocated household-survey based data set. Second, DHS survey locations are randomly displaced in space for data confidentiality reasons, further increasing the level of spatial uncertainty. The covariate layers we use to extract external information for each DHS cluster do not vary substantially at this high resolution. Hence, we expect that using, for example, buffering extraction surrounding GPS locations instead of point extraction using GPS locations will give extremely similar results and that our overall findings will remain essentially unchanged. Third, our modeling framework is, like any statistical model, an abstraction of reality and involves several assumptions and simplifications. Additional details and discussion about these study limitations are included in the appendix.

Implications of all the available evidence

Little apparent overlap between sampling and measurement errors and little correlation between estimates of specific measurement errors indicate that there are distinct underlying reasons for the occurrence of each type of error. Consequently, further improvements to survey activity should leverage local knowledge about prevailing causes or error. For example, in localities of a relatively large sampling uncertainty paired with relatively few measurement errors, future survey rounds may benefit from larger sampling sizes. Alternatively, in localities and areas of relatively low sampling uncertainty but high rates of measurement errors, enumerators may want to dedicate special attention and additional resources to mitigate the threats to data originating from the specific measurement error in place. In practice, the drawbacks of such changes in survey design and field work practices such as increased financial burden will need to be weighed against the benefits on a case-by-case basis. Given that many health practitioners and governments use DHS data directly to inform policy and interventions, improving understanding of data quality within local contexts is essential for public health. Estimates of the magnitudes of the errors will allow governments and agencies to plan for the next rounds of data collection with better targeted resources.

Contributors

VS, GZ, ABF, SL, SEKB and PW conceived and planned the study. ECU and GZ obtained, extracted, processed and geopositioned the data. These two authors directly accessed and verified the underlying data reported in the manuscript. SL compiled the geodata and wrote the computer code for the visualizations. ECU and GZ constructed covariate data layers. ECU and GZ wrote the computer code and designed and carried out the statistical analyses with input from ABF, AJT, SL and VS. AWS, ECU, GZ, SL and VS prepared tables and figures. AJT, SEKB and PW provided intellectual inputs into aspects of this study. MB built the online data visualization tool. VS and ABF wrote the first draft of the manuscript, and all authors contributed to subsequent revisions.

Declarations of interests

The authors declare no competing interests.

Data sharing

This study follows the Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER). The source code used to generate estimates is publicly accessible online. The study data, including full sets of estimates at the first and second administrative levels, are available online at https://data.worldpop.org/repo/prj/dhs/SSA/data_quality.zip. Our data visualization tool provides easy access our estimates used in results 1 and 3 online at https://apps.worldpop.org/SSA/data_quality/.

Acknowledgments

VS is supported by the Austrian National Bank's Anniversary Fund Grant No. 18157. PW would like to acknowledge the support of Feed the Future Food Systems for Innovation Lab, funded by the United States Agency for International Development, Cooperative Agreement No. 7200AA21LE00001. AF is supported by the University of Edinburgh.

References

1 Pelletier F. Census counts, undercounts and population estimates: The importance of data quality evaluation. United Nations, Department of Economics and Social Affairs, 2020.

2 Randall S, Coast E. The quality of demographic data on older Africans. *Demogr Res* 2016; **34**: 143–74.

3 Chan M, Kazatchkine M, Lob-Levyt J, *et al.* Meeting the Demand for Results and Accountability: A Call for Action on Health Data from Eight Global Health Agencies. *PLoS Med* 2010; **7**: e1000223.

4 Buckland AJ, Thorne-Lyman AL, Aung T, *et al.* Nutrition data use and needs: Findings from an online survey of global nutrition stakeholders. *J Glob Health* 2020; **10**. DOI:10.7189/jogh.10.020403.

5 The Demographic and Health Surveys (DHS) Program. The DHS Program. HttpsdhsprogramcompublicationsJournal-Artic.-Searchcfm. https://dhsprogram.com/publications/Journal-Articles-Search.cfm (accessed Feb 1, 2023).

6 Allen C, Croft T, Pullum TW, Namaste S. Evaluation of Indicators to Monitor Quality of Anthropometry Data During Fieldwork. DHS Program, ICF, 2019.

7 Dwyer-Lindgren L, Cork MA, Sligar A, *et al.* Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature* 2019; **570**: 189–93.

8 Kim R, Bijral AS, Xu Y, *et al.* Precision mapping child undernutrition for nearly 600,000 inhabited census villages in India. *Proc Natl Acad Sci* 2021; **118**: e2025865118.

9 Golding N, Burstein R, Longbottom J, *et al.* Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. *The Lancet* 2017; **390**: 2171–82.

10 Jerven M. Beyond precision: embracing the politics of global health numbers. *The Lancet* 2018; **392**: 468–9.

11 Akuze J, Blencowe H, Waiswa P, *et al.* Randomised comparison of two household survey modules for measuring stillbirths and neonatal deaths in five countries: the Every Newborn-INDEPTH study. *Lancet Glob Health* 2020; **8**: e555–66.

12 Perumal N, Namaste S, Qamar H, Aimone A, Bassani DG, Roth DE. Anthropometric data quality assessment in multisurvey studies of child growth. *Am J Clin Nutr* 2020; **112**: 806S-815S.

13 World Health Organization (WHO) & United Nations Children's Fund (UNICEF). Recommendations for data collection, analysis and reporting on anthropometric indicators in children under 5 years old. World Health Organization, 2019.

14 Corsi DJ, Perkins JM, Subramanian SV. Child anthropometry data quality from Demographic and Health Surveys, Multiple Indicator Cluster Surveys, and National Nutrition Surveys in the West Central Africa region: are we comparing apples and oranges? *Glob Health Action* 2017; **10**: 1328185.

15 Grellety E, Golden MH. The Effect of Random Error on Diagnostic Accuracy Illustrated with the Anthropometric Diagnosis of Malnutrition. *PLOS ONE* 2016; **11**: e0168585.

16 Espey J, Swanson E, Badiee S, *et al.* Data for Development: A Needs Assessment for SDG Monitoring and Statistical Capacity Development. Sustainable Development Solutions Network, 2015.

17 United Nations. Transforming our world: The 2030 Agenda for Sustainable Development. *N Y U N Dep Econ Soc Aff* 2015.

18 Allen CK, Fleuret J, Ahmed J. Data Quality in Demographic and Health Surveys That Used Long and Short Questionnaires. Rockville, Maryland, USA: ICF, 2020.

19 Namaste S, Benedict RK, Henry M. Enhancing Nutrition Data Quality in The DHS Program. Rockville, Maryland, USA: ICF, 2018.

20 Pullum TW, Juan C, Khan N, Staveteig S. The Effect of Interviewer Characteristics on Data Quality in DHS Surveys. Rockville, Maryland, USA: ICF, 2018.

21 Assaf S, Kothari MT, Pullum T. An Assessment of the Quality of DHS Anthropometric Data, 2005-2014. Rockville, Maryland, USA: ICF International, 2015.

22 Riese S, Assaf S, Pullum T. Measurement approaches for effective coverage estimation. Rockville, Maryland, USA: ICF, 2021 https://www.dhsprogram.com/pubs/pdf/MR31/MR31.pdf.

23 Prudhon C, de Radiguès X, Dale N, Checchi F. An algorithm to assess methodological quality of nutrition and mortality cross-sectional surveys: development and application to surveys conducted in Darfur, Sudan. *Popul Health Metr* 2011; **9**: 1–8.

24 SMART. The SMART Plausibility Check for Anthropometry. ACF Canada, 2015.

25 USAID. Anthropometric Data in Population-Based Surveys, Meeting Report, July 14-15, 2015. FHI 360/FANTA Washington, DC, 2016.

26 Utazi CE, Thorley J, Alegana VA, *et al.* Mapping vaccination coverage to explore the effects of delivery mechanisms and inform vaccination strategies. *Nat Commun* 2019; **10**. DOI:10.1038/s41467-019-09611-1.

27 Croft TN, Marshall AMJ, Allen CK, Arnold F, Assaf S, Balian S. Guide to DHS Statistics. Rockville, Maryland, USA: ICF, 2018.

28 ICF. Demographic and Health Survey: Sampling and Household Listing Manual. Calverton, Maryland, U.S.A.: ICF International, Calverton, MD, 2012.

29 Johnson K, Grant M, Khan S, Moore Z, Armstrong A, Sa Z. Fieldwork-Related Factors and Data Quality in the Demographic and Health Surveys Program. Calverton, Maryland, USA: ICF Macro, 2009.

30 Eisele TP, Rhoda DA, Cutts FT, *et al.* Measuring Coverage in MNCH: Total Survey Error and the Interpretation of Intervention Coverage Estimates from Household Surveys. *PLoS Med* 2013; **10**: e1001386.

31 Cutts FT, Claquin P, Danovaro-Holliday MC, Rhoda DA. Monitoring vaccination coverage: Defining the role of surveys. *Vaccine* 2016; **34**: 4103–9.

32 Ozodiegwu ID, Ambrose M, Battle KE, *et al.* Beyond national indicators: adapting the Demographic and Health Surveys' sampling strategies and questions to better inform subnational malaria intervention policy. *Malar J* 2021; **20**. DOI:10.1186/s12936-021-03646-w.

33 Finaret AB, Hutchinson M. Missingness of Height Data from the Demographic and Health Surveys in Africa between 1991 and 2016 Was Not Random but Is Unlikely to Have Major Implications for Biases in Estimating Stunting Prevalence or the Determinants of Child Height. *J Nutr* 2018; **148**: 781–9.

34 Balcik B, Beamon BM, Smilowitz K. Last mile distribution in humanitarian relief. *J Intell Transp Syst* 2008; **12**: 51–63.

35 Davison CM, Bartels SA, Purkey E, *et al.* Last mile research: a conceptual map. *Glob Health Action* 2021; **14**: 1893026.

36 Fayehun O, Ajayi AI, Onuegbu C, Egerson D. Age heaping among adults in Nigeria: evidence from the Nigeria Demographic and Health Surveys 2003–2013. *J Biosoc Sci* 2019; **52**: 132–9.

37 Rerimoi AJ, Jasseh M, Agbla SC, Reniers G, Roca A, Timaeus IM. Under-five mortality in The Gambia: Comparison of the results of the first demographic and health survey with those from existing inquiries. *Plos One* 2019; **14**: e0219919.

38 Harkare HV, Corsi DJ, Kim R, Vollmer S, Subramanian SV. The impact of improved data quality on the prevalence estimates of anthropometric measures using DHS datasets in India. *Sci Rep* 2021; **11**. DOI:10.1038/s41598-021-89319-9.

39 Singh M, Kashyap GC, Bango M. Age heaping among individuals in selected South Asian countries: evidence from Demographic and Health Surveys. *J Biosoc Sci* 2021; : 1–10.

40 UNICEF. Multiple Indicator Cluster Survey. MICS. https://mics.unicef.org/.

41 The World Bank Group. Living Standards Measurement Survey (LSMS). https://www.worldbank.org/en/programs/lsms.

42 Li X, Zhou Y, Zhao M, Zhao X. A harmonized global nighttime light dataset 1992–2018. *Sci Data* 2020; **7**. DOI:10.1038/s41597-020-0510-y.

43 Marconcini M, Metz-Marconcini A, Üreyen S, *et al.* Outlining where humans live, the World Settlement Footprint 2015. *Sci Data* 2020; **7**. DOI:10.1038/s41597-020-00580-5.

44 OECD/SWAC. Africapolis (database). 2020; published online Feb. www.africapolis.org.

45 Stoyan D. Matérn, B.: Spatial Variation. Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, 1986.

46 Leasure DR, Jochem WC, Weber EM, Seaman V, Tatem AJ. National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty. *Proc Natl Acad Sci* 2020; **117**: 24173–9.

47 Okwi PO, Ndeng'e G, Kristjanson P, *et al.* Spatial determinants of poverty in rural Kenya. *Proc Natl Acad Sci* 2007; **104**: 16769–74.

48 Dowell SF, Blazes D, Desmond-Hellmann S. Four steps to precision public health. Nat. News. 2016; **540**: 189–91.

49 ESA. ESA/CCI Viewer. 2020. https://maps.elie.ucl.ac.be/CCI/viewer/.

50 Center for International Earth Science Information Network – CIESIN. Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Count. 2016. DOI:10.7927/H4X63JVC.

51 GADM. Global Administrative Areas. 2021. https://gadm.org/data.html.

52 WWF. World Wildlife Fund. Global Lakes and Wetlands Database Level 1. 2004. https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database.

Figures

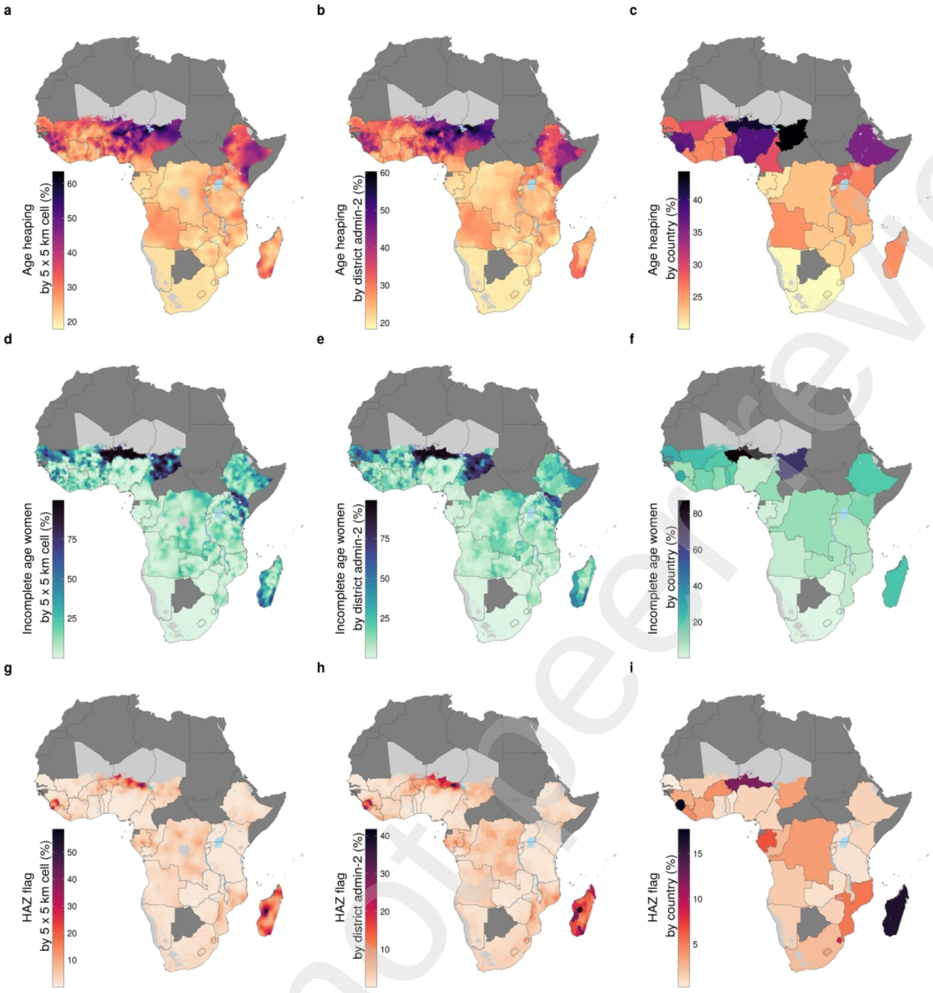
a
b
c
Age heaping by 5 x 5 km cell (%)
60
50
40
30
20
Age heaping by district admin-2 (%)
60
50
40
30
20
Age heaping by country (%)
40
35
30
25
d
e
f
Incomplete age women by 5 x 5 km cell (%)
75
50
25
Incomplete age women by district admin-2 (%)
75
50
25
Incomplete age women by country (%)
80
60
40
20
g
h
i
HAZ flag by 5 x 5 km cell (%)
50
40
30
20
10
HAZ flag by district admin-2 (%)
40
30
20
10
HAZ flag by country (%)
15
10
5

Figure 1. Distribution of measurement errors in DHS data 2009-2016. a-i, Proportion of reported ages ending in 5 or 0 of all adults between 23 and 62 ('age heaping') at **(a)** 5 × 5-km grid-cell level; **(b)** district level (admin-2); **(c)** country level. Share of interviewed women (15-49 years) with either the year or month of birth missing relative to all interviewed women ('incomplete age') at **(d)** 5 × 5-km grid-cell level; **(e)** district level (admin-2); **(f)** country level. Missing or biologically implausible values for the attained heights of children (height-for-age z-scores) according to WHO standards ('flagged HAZ') at **(g)** 5 × 5-km grid-cell level; **(h)** district level (admin-2); **(i)** country level. Countries in dark grey are not in the sample. Grid cells with fewer than 10 people per 1 × 1-km and classified as barren or sparsely vegetated or grid cells with population data not available are colored light grey [49–52].

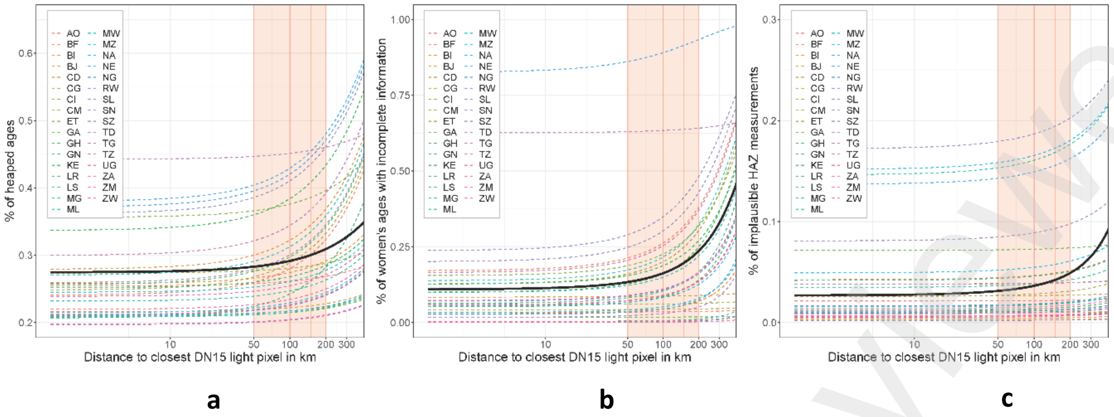

Figure 2. Predicted data quality by distance to closest DN 15 nighttime light emitting source in km (logarithmic scale). a-c, Predictions obtained from regional binomial logistic regressions on distance (in km) to closest DN 15 light pixel of **(a)** share of reported ages ending in 5 or 0 of all adults between 23 and 62 ('age heaping'), **(b)** share of interviewed women (15-49 years) with either the year or month of birth reported missing ('incomplete age'), and **(c)** implausible or missing values for the attained height-for-age z-scores of children under five according to WHO standards ('flagged HAZ'). All models include country fixed effects.

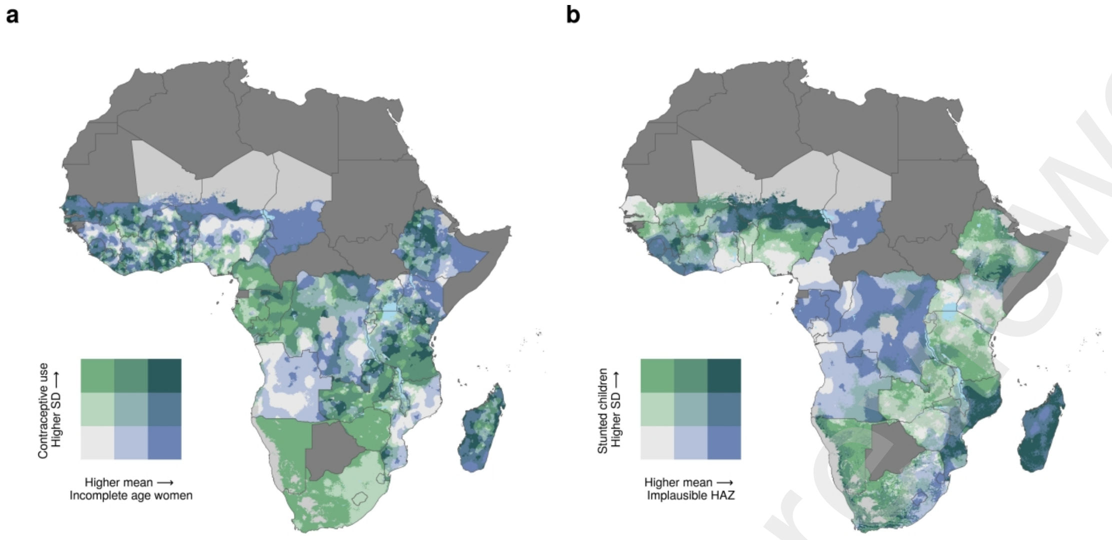

Figure 3. Distribution of measurement errors and uncertainty of predicted estimates of public health indicators in DHS data 2009-2016. a-b, standard deviations of predicted estimates of contraceptive use of sexually active women (in green) and incomplete age values (in blue) at 5 × 5-km grid-cell level **(a)**, standard deviations of predicted estimates of stunting prevalence among children (in green) and flagged HAZ values (in blue) at 5 × 5-km grid-cell level **(b)**. Countries in dark grey were not in the sample. Grid cells with fewer than 10 people per 1 × 1-km and classified as barren or sparsely vegetated or grid cells with population data not available are colored light grey [49–52].