Accepted Manuscript

Title: Integration of population census and water point mapping data–A case study of Cambodia, Liberia and Tanzania

Authors: Weiyu Yu, Nicola A. Wardrop, Robert Bain, Jim A. Wright

PII: S1438-4639(16)30542-9
DOI: http://dx.doi.org/doi:10.1016/j.ijheh.2017.04.006
Reference: IJHEH 13078

To appear in:

Received date: 21-11-2016
Revised date: 17-2-2017
Accepted date: 14-4-2017

Please cite this article as: Yu, Weiyu, Wardrop, Nicola A., Bain, Robert, Wright, Jim A., Integration of population census and water point mapping data–A case study of Cambodia, Liberia and Tanzania.International Journal of Hygiene and Environmental Health http://dx.doi.org/10.1016/j.ijheh.2017.04.006

Integration of population census and water point mapping data – A case study of Cambodia, Liberia and Tanzania

Weiyu Yu[a,*], Nicola A. Wardrop[a], Robert Bain[b], Jim A. Wright[a,*]

[a]Geography and Environment, University of Southampton, Southampton, Hampshire, SO17 1BJ, United Kingdom

[b]Division of Data, Research and Policy, United Nations Children's Fund, 3 United Nations Plaza, New York, NY 10017, USA

[*]Corresponding authors: W.Yu@soton.ac.uk (WY); J.A.Wright@soton.ac.uk (JAW)

Abstract

Sustainable Development Goal (SDG) 6 has expanded the Millennium Development Goals' focus from improved drinking-water to safely managed water services. This expanded focus to include issues such as water quality requires richer monitoring data and potentially integration of datasets from different sources. Relevant data sets include water point mapping (WPM), the survey of boreholes, wells and other water points, census and household survey data. This study examined inconsistencies between population census and WPM datasets for Cambodia, Liberia and Tanzania, and identified potential barriers to integrating the two datasets to meet monitoring needs. Literatures on numbers of people served per water point were used to convert WPM data to population served by water source type per area and compared with census reports. For Cambodia and Tanzania, discrepancies with census data suggested incomplete WPM coverage. In Liberia, where the data sets were consistent, WPM-derived data on functionality, quantity and quality of drinking water were further combined with census area statistics to generate an enhanced drinking-water access measure for protected wells and springs. The process revealed barriers to integrating census and WPM data, including exclusion of water points not used for drinking by households, matching of census and WPM source types; temporal mismatches between data sources; data quality issues such as missing or implausible data values, and underlying assumptions about population served by different water point technologies. However, integration of these two data sets could be used to

identify and rectify gaps in WPM coverage. If WPM databases become more complete and the above barriers are addressed, it could also be used to develop more realistic measures of household drinking-water access for monitoring.

Keywords: Data integration; Water point mapping; Census; WASH

1. Introduction

After the expiry of the United Nations' Millennium Development Goals (MDGs) in 2015, the international focus on water and sanitation has shifted to a broader agenda through the Open Working Group of the General Assembly's Sustainable Development Goals (SDGs) (United Nations General Assembly, 2014, WHO and UNICEF, 2015a). The ongoing development of indicators for enhanced monitoring of progress towards the SDG targets is likely to place greater demands on existing datasets. For example, percentage of population using safely managed drinking water services has been adopted as an indicator for SDG Target 6.1 (which aims to 'achieve universal and equitable access to safe and affordable drinking water for all') (WHO and UNICEF, 2015a, Division for Sustainable Development of UN-DESA, 2016). This suggests a need to integrate population-based data on household drinking water with information on water service levels related to the Human Right to Water and Sanitation (United Nations General Assembly, 2010), including quality, accessibility and availability (WHO and UNICEF, 2015a). When used in isolation, conventional data sources may not be able to meet the expanded demand for more sophisticated international monitoring.

Population census data have long been used alongside household surveys for MDGs monitoring of access to drinking water and sanitation. Censuses can be disaggregated spatially to a greater extent than household surveys (Yu et al., 2014), which facilitates their integration with other spatial datasets. However, household water sources are often classified inconsistently from country to country, census enumeration is typically decadal and population census content is restricted to core household characteristics only, for example, excluding water quality (Yu et al., 2016). Moreover, in low and middle income countries (LMICs), census small area statistics are often not publicly available. Where data are made available, they are frequently provided as aggregate data for relatively large

administrative units rather than for small areas or as micro-data with codes for provinces or districts (Ruggles et al., 2003), for data protection reasons.

The introduction of the SDGs has coincided with increased availability of alternative datasets relevant to safe drinking water access, including water point mapping (WPM). As a process, WPM involves data collection and mapping, storage, processing and analysis, relating to individual water supply points (Welle, 2007, Shantz, 2013). The water point inclusion criteria and characteristics recorded in WPM vary depending on a project's purpose. However, characteristics often include information on location and type of water point, functionality, construction information, perceived water quality, service sustainability, and other relevant characteristics. In many WPM exercises only perceived, aesthetic water characteristics are recorded (Welle, 2005, Liberia Ministry of Public Works and Liberia WASH consortium, 2011). However, in a few cases (Coast Water Services Board, 2013, Shantz, 2013), micro-biological or chemical parameters (e.g. *E.coli*, fluoride, or arsenic) are tested to assess source safety. In addition, when micro-biological or chemical tests are generally considered expensive, an enhanced water point mapping (EWPM) approach has been developed to assess water contamination at a reduced cost (de Palencia and Pérez-Foguet, 2012). In addition, although WPM data collection is sometimes centrally coordinated by government, data are often collected by a diverse range of organisations, including NGOs, so data collection protocols can be similarly diverse. The introduction of a WPM data exchange (WPDx) and related Data Exchange Standard has facilitated the standardisation of such data (Global Water Challenge, 2014) and the exchange now includes approximately 244,000 data points from 25 countries.

While population census data provide spatially disaggregated data that describe both household water sources and socio-economic characteristics, water point data could provide greater temporal resolution and supplementary information, for example, concerning water safety management, quality or quantity. However, the diversity of organisations and projects generating WPM data, along with their differing objectives, may lead to incomplete and spatially biased coverage. The future utility of WPM data will depend on our ability to integrate it with other data sources, such as population censuses. This study therefore aims to (1) quantify the apparent differences in population served by specific water source types, based on population census versus water point data (as a proxy for completeness of WPM coverage); (2) examine potential barriers to integrating population census and water point data to meet SDG monitoring needs; and (3) evaluate the potential insights

into household water use and monitoring that may be gained from integrating these two datasets.

2. Methods

2.1 Study countries and data

This study examines three case study low income countries: Cambodia, Liberia and Tanzania (Fig. 1), chosen for their relatively detailed population census content on water source categories (Yu et al., 2016) identifying a household's main drinking water source, and the availability of extensive WPM data. Geo-referenced Cambodian population census 2008 data were obtained from Open Development Cambodia (http://www.opendevelopmentcambodia.net/). These data consist of four administrative levels: 1 (province); 2 (district); and 3 (commune) as vector polygons; and level 4 (village) as vector points. The water point data for Cambodia were obtained from Cambodia WellMap (http://cambodiawellmap.com/) which includes 59,759 records of drilled, dug and combination well water points, originally sourced from several non-governmental organisations (NGOs) and government entities. Tabular data for households by main source of drinking water in Liberia were obtained from the 2008 Liberian population census (Liberia Institute of Statistics and Geo-Information Services, 2011). The Liberia 2011 WPM dataset was acquired from WASH Liberia (http://wash-liberia.org/), and consists of 10,001 improved water points (data submitted date from 2010-2011), with associated demographic data (population and household numbers) from the 2008 population census at administrative level 1 (county). For Tanzania, 2012 population census data were obtained from the Tanzania National Bureau of Statistics portal (http://www.nbs.go.tz/); water information at administrative level 2 (district) was derived from their regional basic demographic and socio-economic profiles. Tanzania water point data were acquired from the WPDx portal (http://waterpointdata.org/; data acquired: 25 January 2016), covering 23,352 records contributed by different data collection organisations in 1978, 1982, 2002-2009, and 2013-2014, but having excluded water points with missing GPS coordinates.

2.2 Water point data pre-processing

In order to match water point data with population census data, the initial database was filtered to remove disused or non-domestic water points and those constructed following census enumeration. Water points that met any of the following criteria were therefore removed: (1) water points recorded as disused (e.g. abandoned, closed due to lack of payment, etc.); (2) water points installed after the census enumeration date; (3) water points recorded as serving a facility or workplace (e.g.

school, health centre, place of worship, etc.) rather than households; (4) water points used for purposes other than household drinking water (e.g. cattle troughs). Water points with ambiguous characteristics were retained. (Detailed characteristics used to identify water points for exclusion were listed in Supplementary Information A.)

Water point coordinates were used to link WPM data with census data at the commune level (administrative level 3) in Cambodia, county level (administrative level 1) in Liberia, and district level in Tanzania (administrative level 2). Water points lacking coordinates and those with implausible coordinates (e.g. outside national boundaries or in the sea) were excluded. Where there was a mismatch between GPS coordinates and recorded administrative area, administrative area information (name, ID) was corrected according to the GPS location.

Note to Publisher: Insert **Fig. 1** about here

Besides location, WPM and census data linkage required definitional matching of their respective water source classifications. WPM classifications were generally more detailed. For example, Tanzanian water point data contained information such as the original source of water (e.g. groundwater, surface water, rainwater), type of water point (e.g. standpipe, hand pump, etc.), and water extraction or lifting system (e.g. Afridev hand pump, electrically driven mechanised pump, gravity scheme, etc.). In comparison, water source categories in population census data were user-based and generally less detailed.

The Cambodian WPM data contained three different types of well water point categories, namely drilled well, dug well, and combination well (e.g. an open-well constructed above an underlying tube-well), whilst the Cambodian population census differentiated protected dug wells, unprotected dug wells and tube wells. Since water point data do not distinguish protected dug wells from tube wells and since combination wells could not be unambiguously placed into the tube well or dug well category, these three categories were collapsed together in both datasets to facilitate comparison.

In Liberia, because of the difficulties of matching other source types across the two datasets, the census-based ‘protected dug well and protected spring’ class was matched to the combined ‘manual pump on dug well’ and ‘protected spring’ WPM classes. Other source categories were discarded

from the analysis.

For Tanzania, WPM data contained detailed information on original water source, water extraction system, and water point type, which enabled most source categories to be matched with their equivalents in population census data (except for 'cart with small tank/drum'). The matched WPM classes were therefore each of 'standpipe', 'tube well or borehole', 'protected dug well' and 'protected spring', with other WPM classes with very few associated records being discarded (e.g. there were only two 'piped into dwelling' water points).

2.3 Measuring water point data coverage for matching census data

To identify gaps in water point coverage in each administrative area, we calculated the ratio of census-based population using each water source type to the maximum population potentially served by the recorded water points. This gave an index of areal water point shortage:

$$I_x(A)=\frac{\sum[N_{xi}(A)S_{xi}]}{P_x(A)} \qquad \text{(Eq. 1)}$$

where $P_x(A)$ represents the total census population using water source type x (e.g. tube well, dug well, spring, etc.) as their main drinking source within administrative area A; when $P_x(A)=0$ (no record of certain water source user), $I_x(A)=1$; $N_{xi}(A)$ is the number of water points of type x with water extraction/lifting method i (e.g. hand pump, powered pump, etc.) within administrative area A; and S_{xi} refers to the maximum population served by water point type x with water extraction/lifting method i, as we assume all water points are serving the maximum design capacity population.

Table 1 shows the maximum population served (S_{xi}) by different types of water point and extraction/lifting techniques in the Cambodian, Liberian and Tanzanian WPM datasets. To estimate the maximum number of people theoretically being served for each water point, we assumed plausible numbers based on the type of water point and the type of water extraction or lifting method, according to technical information and empirical evidence from previous studies (Jordan, 1984, Baumann, 2000, 2011, Smet and Wijk-Sijbesma, 2002, Mwakali, 2006, Baumann et al., 2010) (see Supplementary Information B). Where technical data could not be obtained (e.g. 'other' pump model), we made different assumptions based on other available information (e.g. according to well depths if available, see Table 2) and logical inferences. We examined the sensitivity of the index to the largest capacity-based assumption (ASI) from literature, smallest capacity-based assumption (ASII), and the commonly-made assumption that all water points serve 300 people (ASIII).

The index of areal water point shortage $I_x(A)$ measures the completeness of WPM coverage for an administrative areal unit A. When $I_x(A)$ is less than 1, it suggests incomplete WPM coverage in area A, since the maximum capacity of recorded water points could not serve the headcount recorded in the census for that area. $I_x(A)$ greater than 1 is harder to interpret, but suggests more complete WPM coverage, potentially that WPM features represent seasonal, secondary drinking water sources or water for other purposes, or potentially that individual water points served fewer people than suggested by the sources in Supplementary Information B. When the census contains no recorded households using water source type x within area A, water points of type x recorded via WPM are likely to be secondary household drinking water sources or used for other purposes.

2.4 Analysis of $I_x(A)$ values

In contrast to Liberia, WPM exercises in Tanzania and Cambodia were undertaken by multiple organisations and merged from multiple data sources in Cambodia. The data collection and mapping approaches therefore might potentially vary depending on the organisation involved or original data source. $I_x(A)$ values for Cambodian communes were therefore classified according to the predominant water point inventory source in the commune (as the institution collecting data was not reported). Boxplots were used to examine the distribution of $I_x(A)$ values for each of these original data sources. One-way ANOVA was used to test for significant differences in mean data coverage (as measured by $I_x(A)$) between different original water point inventory sources, with Scheffé's method (Scheffé, 1959) as post hoc test. In Tanzania, the organisations collecting WPM were recorded; however, a subsequent, equivalent analysis to Cambodia was not feasible in Tanzania, since WaterAid and SNV were the predominant data collection organisations in most of the Tanzanian districts, precluding assessment of index values by data collection organisation.

2.5 Data integration

To illustrate the potential benefits of integrating census and WPM data, we focussed on Liberia where WPM was centrally coordinated and our index values were all greater than 1. In each county, we adjusted the census-based proportion of population using protected wells downwards to exclude source with three types of problem identified through WPM. These problems included user-reported water quality issues (e.g. salty taste; rust-coloured water); water points whose functionality was described as 'working but with problems'; and point sources that were unable to provide sufficient water all year round according to users. In doing so, we assumed each water point served the same

number of people within each county. We therefore sought to adjust coverage figures to take account of perceived water quality, impaired source functionality, and user-reported sufficiency of supply.

3. Results

Fig. 2 shows the number of water points excluded during pre-processing for the three study countries based the exclusion criteria. More than half (55.5%) of Liberian water points were excluded either because they were abandoned (8.7%), installed after the 2008 population census (24.6%), serving workplaces or facilities (2.4%), not used for drinking (11.5%), had implausible coordinates (<0.1%), or were without an equivalent source type in the census (8.3%). In comparison, only 3.3% and 9.8% of the water points were excluded from Cambodian and Tanzanian datasets respectively. Table 3 provides summary information for the water points included for each study country.

Note to Publisher: Insert **Fig. 2** about here

Note to Publisher: Insert **Fig. 3** about here

Note to Publisher: Insert **Fig. 4** about here

Note to Publisher: Insert **Fig. 5** about here

Table 4 summarises the measured $I_x(A)$ index by census administrative level and country, based on the three different assumptions about water point capacity. Based on ASI (highest capacity estimates), 43.3% of the communes in Cambodia had incomplete water point coverage when compared with population census 2008 data (Table 4), and this number increased to 61.8% when based on ASII (lowest capacity estimates). There were 42.0% and 37.5% respectively of Cambodian districts and provinces with incomplete coverage based on ASI, increasing to 63.2% and 54.2% based on ASII. For Tanzania, most regions (administrative level 1) and districts had incomplete water point coverage of public taps/standpipes, tube wells/boreholes, protected dug wells, and protected springs

in comparison to population census 2012 data. At administrative level 2, water point coverage appeared greater for standpipes (77.5%-85.2% of districts with incomplete coverage, depending on assumptions) than for springs (91.7% of districts with incomplete coverage). At administrative level 1, over 90% of regions had incomplete coverage for all source types. None of the 15 counties in Liberia had apparently incomplete water point coverage. Fig. 3 shows the WPM-derived coverage measure versus household drinking water source use in the census by country and source type, based on a moderate assumption about population served by water points (ASIII). The two measures were strongly correlated for Liberian counties ($r=0.87$, $p<0.01$, $n=15$) and Tanzania regions for protected springs ($r=0.74$, $p<0.01$, $n=30$), but weakly correlated for Cambodian districts and communes ($r=0.21$, $p<0.01$, $n=193$, and $r=0.09$, $p<0.01$, $n=1,621$, respectively), Tanzanian regions for tube wells ($r=0.06$, $p=0.76$, $n=30$) and districts for standpipes ($r=-0.04$, $p=0.62$, $n=169$), tube wells ($r=0.04$, $p=0.61$, $n=169$), and protected dug wells ($r=0.21$, $p<0.01$, $n=169$). Fig. 4 shows the geographic patterns in the index measuring the ratio of census-based to WPM-derived source coverage. WPM coverage appeared incomplete in most parts (56.5% of the districts) of Cambodia, primarily distributed in Cardamom and Elephant Mountains and North-western regions, with the exception of Pailin province; and also incomplete in most parts of Tanzania (78.1%, 87.6%, 83.4% and 91.7% for standpipes, tube wells, protected dug wells and protected springs, respectively); however, WPM-derived coverage of all the four water source types are close to census-based figures in Bukoba, Chemba, Iramba, Kisarawe, Manyoni, and Missenyi districts. WPM-census agreement changes with scales, as shown in Fig. 5 for example.

Note to Publisher: Insert **Fig. 6** about here

When $I_x(A)$ index values were disaggregated by the predominant data source used to compile the Cambodian WPM database in each commune (Fig. 6), index values for two data sources, namely the RDI Rope Pump Table and World Vision Hard Copies, suggest complete water point coverage for all such communes. The one-way ANOVA found statistically significant differences in mean $I_x(A)$ values between predominant data sources at $\alpha=0.05$ ($F=234.519$; P-value<<0.001; F-critical=2.105); Scheffé's test and boxplots suggested that communes where data from the RDI Rope Pump Table predominant had statistically higher mean index values at $\alpha=0.05$ than other communes.

Note to Publisher: Insert **Fig. 7** about here

Note to Publisher: Insert **Fig. 8** about here

Following integration of census and WPM data, most of protected dug well users were found served by water points without perceived quality issues that the percentages by county ranged from 76.2% to 95.4% (Fig. 7 B). Grand Kru had the highest percentage (93.6%) of protected dug well/spring users using fully functioning sources, whilst Grand Bassa had the lowest (55.2%) (Fig. 7 A). Maryland and River Gee had the highest percentages (82.4% and 83.5%, respectively) of protected dug well/spring users using water sources reported sufficient for year-round needs, whilst the percentages dropped to lower than 50% in Nimba, Grand Cape Mount, and Gbarpolu (Fig. 7 C). Overall, only 36.9% of protected dug well and spring users in Liberia were served by water points without functionality, sufficiency or quality issues; more than half of users were affected by one or more of these issues in most counties (Fig. 8), with the exceptions of River Gee (71.6%), Grand Kru (64.2%), and Maryland (61.9%).

4. Discussion

This study suggests that for some countries, in this instance Cambodia and Tanzania, available WPM data have incomplete coverage. Where WPM requires coordination of field activities by multiple agencies with scarce data collection resources, the approach presented here could help inform future WPM planning by identifying gaps in dataset coverage and prioritising areas for future surveys. Spatial representation of the ratio of census-based population using different water sources to the maximum population potentially served by recorded water points (Fig. 4) highlighted areas in Cambodia and Tanzania where many households reported using groundwater point sources, yet there were insufficient water points in the WPM datasets to account for such household use. Working towards comprehensive WPM coverage, these areas could be prioritised for follow-up water point mapping.

In Cambodia, we examined the values of $I_x(A)$ relative to the predominant source of water point mapping data in each commune. This suggested that there was coverage of water points more

consistent with census data in communes where data were predominantly drawn from the RDI Rope Table. The reasons for the apparently greater coverage require further investigation, but could reflect RDI's involvement as the project partner in the development of the Cambodian rope pump version and perhaps detailed knowledge of their implementation in these communes. Such an approach could be expanded to expanded to assess other potential influences on WPM-census data comparison, such as the time lag between WPM fieldwork and census enumeration dates. It would also be possible to examine whether census-derived measures of household use of a particular source types are correlated with measures derived via WPM databases.

This study highlights several potential barriers to the spatial integration of WPM and population census data. A fundamental issue is that censuses in general only record a household's main drinking water source, whereas WPM typically records sources used for any purpose. In Liberia, for example, according to the 2014-15 Household Income and Expenditure Survey, closed wells were reported to be used by 6.0% of households for drinking, but 25.0% of households for cooking and 26.4% for washing (The World Bank, 2016). Relative to these differences, seasonal variation in source use is less pronounced with, for example, reported closed well use for drinking only increasingly marginally to 6.6% in the Liberian dry season. In contrast, in Cambodia, source use varies substantially by season, with 40.8% of rural households using rainwater in the wet season but only 10.1% in the dry season (National Institute of Statistics, Directorate General for Health and ICF International, 2015). In some parts of Cambodia (blue areas in Fig. 4 A), and to a lesser extent in parts of Liberia and Tanzania, we observed apparently large numbers of water points relative to the population reporting groundwater source use for drinking via the population census (see also Fig. 3). This may indicate that many recorded water points are used for non-domestic purposes such as watering animals or irrigation. Recent studies have emphasised multiple use water services (van Koppen, Moriarty and Boelee, 2006, van Koppen et al., 2014), recognising that households often use different types of water source for different purposes. WPM data implicitly capture multiple use water services, whereas population census data generally only capture the main domestic or drinking water source. WPM data may thus better reflect a wider range of water uses.

Methodological work on spatial database integration suggests these problems exemplify more generic problems in combining separate databases (Devogele, Parent and Spaccapietra, 1998). Exact equivalence of entities in two spatial databases is rare and often, entities in one database are a

subset, super-set, or only partially overlap with those in a second database. In this instance, provided WPM has complete coverage, among the set of all water sources appearing in WPM data (regardless of their usage or functionality), population census data only relate to the subset of these that are functional household domestic drinking water sources constructed before the census date. Conversely, census data capture piped connections, delivered water (e.g. tanker-truck, small cart with tank/drum) and packaged water (e.g. bottled water, sachet water), but these seldom feature in WPM data. One-to-one relationships between entities in different databases are similarly rare in spatial data aggregation. It is more common to find one-to-many ('aggregation-fragmentation') relationships because databases capture entities at different scales. This also occurs when integrating population census and WPM datasets, since a single census area and water source type typically relates to multiple WPM water points and source types. Water source categories in several cases are collapsed to solve the conflict when undertaking integration, which can result in merging of improved and unimproved water source types. This undermines their utility for national monitoring of drinking water access.

In addition, temporal mismatches between databases and the accuracy of both databases can affect their integration (Flowerdew, 1991). Alongside seasonal source use as noted above, it is often difficult to establish from WPM data when water points were abandoned or became operational, and thus unclear which sources were used on a given census enumeration date. Moreover, water points lacking plausible locations or source type information could not be matched to census data.

This study built a bridge between the two types of information by converting WPM data to plausible numbers of population served per water point, based on technical information and experiences from previous studies; however, where technical data are missing or unclear, assumptions must be made in accordance with logical inferences. This may also impact the accuracy of the integration between population census and WPM datasets, as the difference can be significant between different assumptions. Furthermore, our integration of the two data sets will be affected by census data quality issues, such as under- or over-enumeration and misreporting by households of the souce types that they use.

Our results confirm the value of collecting GPS coordinates for water points, as required under the WPDx standard (Global Water Challenge, 2014). They also highlight several ways that WPM protocols could be modified to facilitate their subsequent integration with census data. Firstly, collecting data

on water point usage (e.g. where they serve a health facility or school, or are used for agriculture) would enable removal of non-domestic water points prior to integration with census data. If such data are not available, non-domestic water points could potentially be identified through map overlap with for example school and health facility locations recorded in OpenStreetMap (OSM). Secondly, the government-coordinated WPM exercise in Liberia produced data that were more consistent with census outputs than in Tanzania and Cambodia (Figs. 3 and 4), where WPM was undertaken by multiple organisations. This highlights the data quality benefits of coordinated data collection. Finally, there is a need to better quantify the populations served by different water points. The number of users and different uses of water points could be quantified by introducing direct observation of users at a sample of water points, or else by interviewing a sample of community leaders or supply managers.

Cross-scale comparison indicates that our measure of WPM-census agreement is greater for larger areal units (e.g. Fig. 5), which may be because proportionately more of the imprecisely georeferenced water points will be displaced across district or commune boundaries. Compared to provinces, more people may cross district or commune boundaries to reach water sources, with for example 2.1% of rural Cambodian households travelling long distances (more than 30 minutes round trip) to obtain drinking water in the wet season and 7.1% in the dry season (National Institute of Statistics, Directorate General for Health and ICF International, 2015).

In Tanzania, the only country where multiple water source types were investigated in this study, the degree of inconsistency between census and WPM-derived coverage varied by water source type (Fig. 4 C – F). Almost all (99.89%) of the Tanzanian water points in the WPDx database were originally sourced from a previous WPM exercise which targeted Improved Community Water Points (ICWPs) but excluded private water points (Welle, 2005) that are more numerous, sometimes inaccessible in household compounds, and thus harder to enumerate. In contrast, the population census covers both communal and private sources. Non-enumeration of privately owned wells and to lesser extent boreholes in Tanzanian WPM may explain the apparently better agreement with census data for standpipes relative to these other source types (Fig. 4 C – F). More generally, our comparison of WPM and census-derived populations served depends on capacity estimates for each water point type (Arlosoroff et al., 1984, DHV, 1984, International Development Research Centre, 1984, Jordan, 1984, The World Bank, 1985, UNICEF, 1997, Baumann, 2000, 2011, Mwakali, 2006, Baumann et al.,

2010, Jiménez and Pérez-Foguet, 2011, USAID, 2016). These may vary by country and be much lower than we have assumed for privately owned water points.

WPM received considerable attention during the MDGs period (Welle, 2005) and has been discussed as a potential candidate to support SDG monitoring. SDG monitoring focuses on availability ('available when needed'), accessibility ('located on premises'), and quality ('free from micro-biological and priority chemical contamination') as key criteria in national systems. Since some WPM datasets cover water quality and sufficiency of sources (WHO and UNICEF, 2015a), their combination with census data on accessibility could support monitoring of progress towards SDG target 6.1. We illustrate a potential method for combining census and WPM data (Figure 6) in this way for Liberia, where suitable data are available.

Currently, WPM data mainly facilitates studies on improved rural water supplies and their functionality, operation and maintenance; limited data exist on actual numbers of population or households served by individual water sources as opposed to their maximum technical capacity. An exception is the Ethiopian National WASH Inventory (NWI) which combined information from water supply scheme inventories (e.g. water quality, functionality, population served) with a household survey on service use and water quantity, and has been used to estimate the population served by improved drinking-water sources (WHO and UNICEF, 2015b). Another previous study (Giné Garriga, de Palencia and Pérez Foguet, 2013) also combined WPM with household survey at local level in Tanzania, Kenya and Mozambique and produced reliable estimates of water coverage and service level. Even where WPM data are not recorded within household surveys, such survey data could still be integrated with WPM. Household surveys are more frequently conducted, have a more internationally standardised water classification system and can capture characteristics such as collection time, supply interruption, seasonality, etc. which are generally beyond the reach of population censuses. As evidenced by greater improved water source coverage in many countries, a large number of water points have been installed in recent years; however, population censuses are generally conducted every ten years. In contrast, household surveys are often conducted more frequently and sometimes even annually, providing data more likely to be contemporaneous with WPM. Some household surveys also capture water use for purposes other than drinking (e.g. cooking and hygiene) and may therefore more closely match the multiple-use sources have not been mapped as part of household survey fieldwork, such surveys' lack of full population coverage remains a

barrier to integration with WPM databases. This is because it makes spatially disaggregated estimation of source coverage challenging.

To incorporate issues such as water quality, affordability and availability into monitoring of household water access, some more recent household surveys have included water quality modules (Wright et al., 2016). When genuinely integrated household surveys data are likely becoming increasingly important source of data for SDG monitoring, integration of WPM with population censuses or household surveys may be an alternative means of addressing these issues.

5. Conclusion

By converting water source-based information to population served based on the technical capacity of each water point, this paper integrated population census data with WPM data in three countries. Several challenges in integrating population census and WPM datasets were identified, including: difficulties in identifying and excluding water sources not used for drinking by households; matching of census and WPM source types; temporal mismatches between data sources, reflected in seasonality of source use and water point functionality; data quality issues such as missing or implausible data values, and assumptions about population served by different water point technologies. Some of these issues may be addressed as government, international bodies, and NGOs coordinate WPM data collection and adopt data standards to address incomplete or inconsistent data capture. In addition, this analysis highlights variation over space and by source type in existing WPM data coverage, with coverage gaps apparent in Cambodia and Tanzania. Such gaps in WPM coverage could be investigated further, for example through follow-up field survey; or could be brought together to an open platform if data exist but remain inaccessible. For Liberia, where WPM was centrally coordinated, our analysis suggested consistency between WPM census data for protected dug wells and springs by county. Here, WPM-based information on functionality, sufficiency and quality of drinking water were combined with census to refine county-level measures of these drinking-water services. The Liberian results demonstrate the potential for census-WPM integrated data to support monitoring progress towards drinking water-related SDGs.

Acknowledgements

The authors gratefully acknowledge Andrew J Tatem and Alessandro Sorichetta from the University of Southampton for their comments on the manuscript.

References

Arlosoroff, S., Grey, D., Journey, W., Karp, A., Langenegger, O., Rosenhall, L. and Tschannerl, G. (1984) Rural Water Supply Handpumps Project Handpumps Testing and Development: Progress Report on Field and Laboratory Testing. Washington, D.C., U.S.A.

Baumann, E. (2000) Water Lifting. 1st ed. St Gallen, Switzerland: SKAT.

Baumann, E. (2011) Low Cost Hand Pumps [Online]. RWSN Field Note 2011-3. St Gallen, Switzerland. Available: http://www.sswm.info/sites/default/files/reference_attachments/BAUMANN 2011 Low Cost Hand Pumps.pdf.

Baumann, E., Montangero, A., Sutton, S. and Erpf, K. (2010) WASH Technology Information Packages - for UNICEF WASH Programme and Supply Personnel [Online]. 1st ed. UNICEF; Skat. Available: http://www.ircwash.org/sites/default/files/unicef_wash_technology_web_0.pdf.

Coast Water Services Board (2013) Water Point Mapping Report: Turkana County (November).

Devogele, T., Parent, C. and Spaccapietra, S. (1998) On spatial database integration. International Journal of Geographical Information Science. [Online] Vol.12 (4), pp.335–352. Available: http://www.tandfonline.com/doi/abs/10.1080/136588198241824.

DHV (1984) Low cost water supply for human consumption, cattle watering, small scale irrigation. Part 2 : Pumping equipment. Amersfoort, The Netherlands.

Division for Sustainable Development of UN-DESA (2016) Sustainable Development Knowledge Platform [Online]. Available: https://sustainabledevelopment.un.org/sdg6 [Accessed 29 May 2016].

Flowerdew, R. (1991) Spatial Data Integration. Geographic Information Systems and Science. Vol.1, pp.375–387.

Giné Garriga, R., de Palencia, A. J.-F. and Pérez Foguet, A. (2013) Water–sanitation–hygiene mapping: An improved approach for data collection at local level. Science of The Total Environment.

[Online] Vol.463-464, pp.700–711. Available: http://www.sciencedirect.com/science/article/pii/S0048969713006578.

Global Water Challenge (2014) WASH Data Sharing Update – September 2014 [Online]. Arlington, VA. Available: http://sustainablewash.org/sites/sustainablewash.org/files/wash_datapoint_update_september_2014_compiled_with_appendices.pdf.

International Development Research Centre (1984) Proceedings of a workshop on Hydraulic Ram Pump (HYDRAM) technology. Arusha, Tanzania.

Jiménez, A. and Pérez-Foguet, A. (2011) Water Point Mapping for the Analysis of Rural Water Supply Plans: Case Study from Tanzania. Journal of Water Resources Planning and Management. [Online] Vol.137 (5), pp.439–447. Available: http://ascelibrary.org/doi/abs/10.1061/(ASCE)WR.1943-5452.0000135.

Jordan, T. D. J. (1984) A handbook of gravity-flow water systems for small communities. London, UK: IT Publications.

Van Koppen, B., Moriarty, P. and Boelee, E. (2006) Multiple-use water services to advance the millennium development goals [Online]. Research Report 98. Colombo, Sri Lanka. Available: http://search.ebscohost.com/login.aspx?direct=true&db=lhh&AN=20093261021&site=ehost-live\nhttp://www.cabi.org/cabdirect/showpdf.aspx?PAN=20093261021\nhttp://www.iwmi.cgiar.org/Publications/IWMI_Research_Reports/PDF/pub098/RR98.pdf.

Van Koppen, B., Smits, S., Rumbaitis del Rio, C. and Thomas, J. B. (2014) Scaling up Multiple Use Water Services. Practical Action.

Liberia Institute of Statistics and Geo-Information Services (2011) Vol.5 Analytical report on housing conditions and housing facilities. 2008 population and housing census. Liberia Institute of Statistics and Geo-Information Services.

Liberia Ministry of Public Works and Liberia WASH consortium (2011) Liberia Waterpoint Atlas. [Online] p.43. Available: http://wash-liberia.org/wp-content/blogs.dir/6/files/sites/6/2013/01/Final_Review_Version_-_Waterpoint_Atlas___Investment_Plan_x1.pdf.

Mwakali, J. (2006) Proceedings from the International Conference on Advances in Engineering and Technology. Elsevier.

National Institute of Statistics, Directorate General for Health and ICF International (2015) Cambodia Demographic and Health Survey 2014. Phnom Penh, Cambodia; and Rockville, Maryland, USA.

De Palencia, A. J. F. and Pérez-Foguet, A. (2012) Quality and year-round availability of water delivered by improved water points in rural Tanzania: effects on coverage. Water Policy. Vol.14 (3), pp.509–523.

Ruggles, S., King, M. L., Levison, D., McCaa, R. and Sobek, M. (2003) IPUMS-International. Historical Methods. [Online] Vol.36 (2), pp.60–65. Available: http://www.tandfonline.com/doi/abs/10.1080/01615440309601215.

Scheffé, H. (1959) The Analysis of Variance. New York: John Wiley & Sons.

Shantz, A. (2013) A Review of Functionality of an Existing National Well Database in Cambodia – Final Report.

Smet, J. and Wijk-Sijbesma, C. Van (2002) Small community water supplies : technology, people and partnership [Online]. IRC - International Water and Sanitation center. Delft, The Netherlands: IRC International Water and Sanitation Centre. Available: http://www.ircwash.org/sites/default/files/Smet-2002-Small_TP40.pdf.

The World Bank (1985) World Bank Technical Paper Number 48. Washington, D.C., U.S.A.

The World Bank (2016) Liberia - Household Income and Expenditure Survey 2014-2015 [Online]. Available: http://microdata.worldbank.org/index.php/catalog/2563 [Accessed 20 Jun 2007].

UNICEF (1997) Report on Evaluation of UNICEF Assisted Rural Water Supply and Sanitation Programme Activities 1992 -1997.

United Nations General Assembly (2010) The human right to water and sanitation (A/RES/64/292).

United Nations General Assembly (2014) Report of the Open Working Group of the General Assembly on Sustainable Development Goals (A/68/970) [Online]. Available: http://www.un.org/ga/search/view_doc.asp?symbol=A/68/970.

USAID (2016) Elephant Pump Innovations [Online]. Available: http://www.divportfolio.org/innovations/elephant-pump [Accessed 21 Mar 2016].

Welle, K. (2005) Learning for Advocacy and Good Practice - WaterAid Water Point Mapping.

Welle, K. (2007) WaterAid learning for advocacy and good practice - Water and sanitation mapping: a synthesis of findings. London.

WHO and UNICEF (2015a) JMP Green Paper : Global monitoring of water , sanitation and hygiene post-2015 (Zero Draft) [Online]. Available: http://www.wssinfo.org/fileadmin/user_upload/resources/JMP-Green-Paper-15-Oct-2015.pdf.

WHO and UNICEF (2015b) Progress on water and sanitation: 2015 Update and MDG Assessment.

Wright, J., Dzodzomenyo, M., Wardrop, N. a., Johnston, R., Hill, A., Aryeetey, G. and Adanu, R. (2016) Effects of sachet water consumption on exposure to microbe-contaminated drinking water: Household survey evidence from Ghana. International Journal of Environmental Research and Public Health. Vol.13 (3).

Yu, W., Bain, R., Mansour, S. and Wright, J. A. (2014) A cross-sectional ecological study of spatial scale and geographic inequality in access to drinking-water and sanitation. International Journal for Equity in Health. [Online] Vol.13 (1), p.113. Available: http://www.equityhealthj.com/content/13/1/113.

Yu, W., Wardrop, N. a., Bain, R. E. S., Lin, Y., Zhang, C. and Wright, J. a. (2016) A Global Perspective on Drinking-Water and Sanitation Classification: An Evaluation of Census Content. Plos One. [Online] Vol.11 (3), p.e0151645. Available: http://dx.plos.org/10.1371/journal.pone.0151645.

Fig. 1. Cambodian (KHM), Liberian (LBR) and Tanzanian (TZA) water point datasets with population census data.

Fig. 2. Numbers of water points excluded when linking Cambodian (KHM), Liberian (LBR) and Tanzanian (TZA) water point datasets with population census data.

Fig. 3. Scatter plot using base-10 log scale showing census-based population using different types of drinking water source (X) versus maximum population potentially served by the recorded water points (Y) for the study countries and water categories based on ASIII. Each dot represents an administrative areal unit; dot shapes represent different administrative levels; KHM, LBR, and TZA represent Cambodia, Liberia and Tanzania respectively. Areal units without record were excluded in the figure.

Fig. 4. Ratio of census-based population using different water sources to the maximum population potentially served by recorded water points based on ASIII; $I_x(A)$ for (A) Cambodian districts, (B) Liberian counties, and (C – F) Tanzanian districts (standpipe, tube well, protected dug well and protected spring respectively); each bar on the right side of map shows percentages of total areal units in different $I_x(A)$ classes.

Fig. 5. Example of cross-scale changes in WPM-census agreement in Prey Veng Province, Cambodia.

Fig. 6. Boxplots of Cambodian commune-level $I_x(A)$s by original data sources. The bottom (light blue) and top (dark blue) of the box represent the 25th and 75th percentiles respectively. Y-axis represents the $I_x(A)$ value; x-axis is the name of original water point inventory source with corresponding number of communes with $I_x(A)<1$ and the total number of communes in brackets.

Fig. 7. Liberian WPM-census data showing respectively the percentage of users by county with protected wells/springs that were (A) fully functioning, (B) without perceived water quality issues and (C) sufficient for year-round needs according to users.

Fig. 8. Liberian WPM-census integrated data showing the percentage of protected well/spring users by county using water sources that were fully functioning, without perceived water quality issues and where users reported sufficient water for all year round.

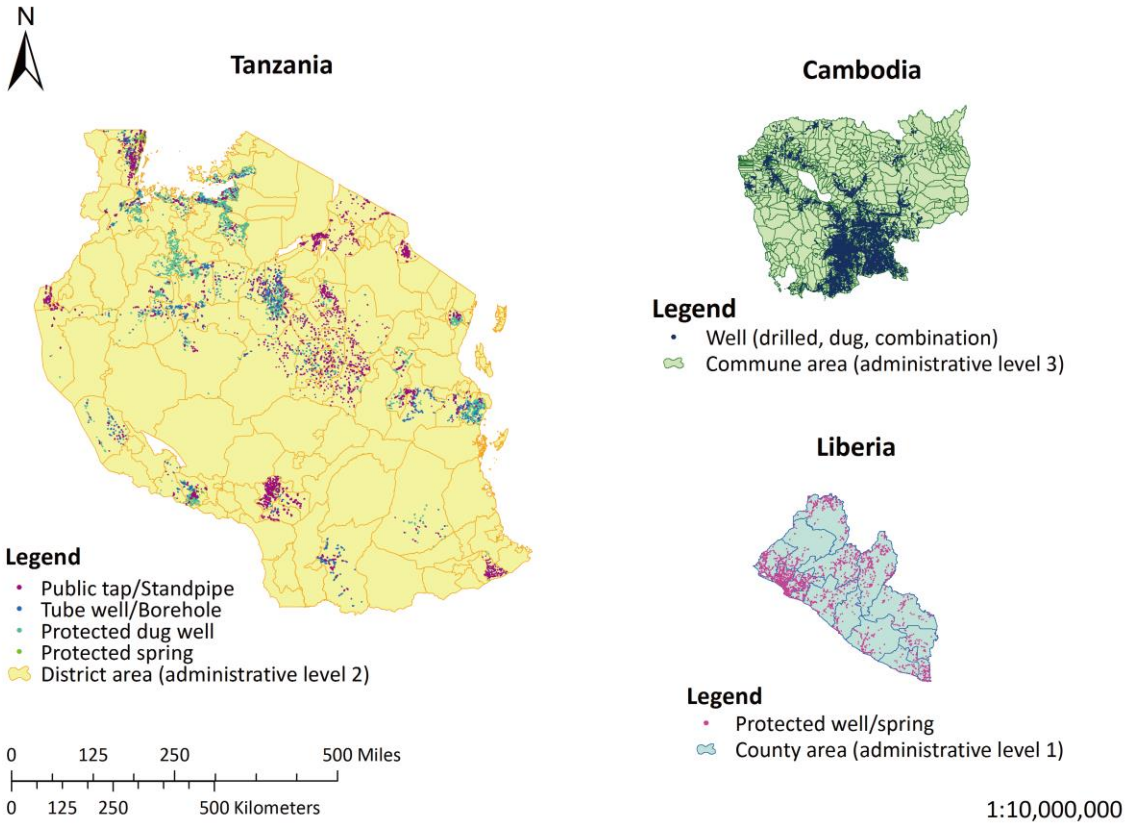

Fig: 1

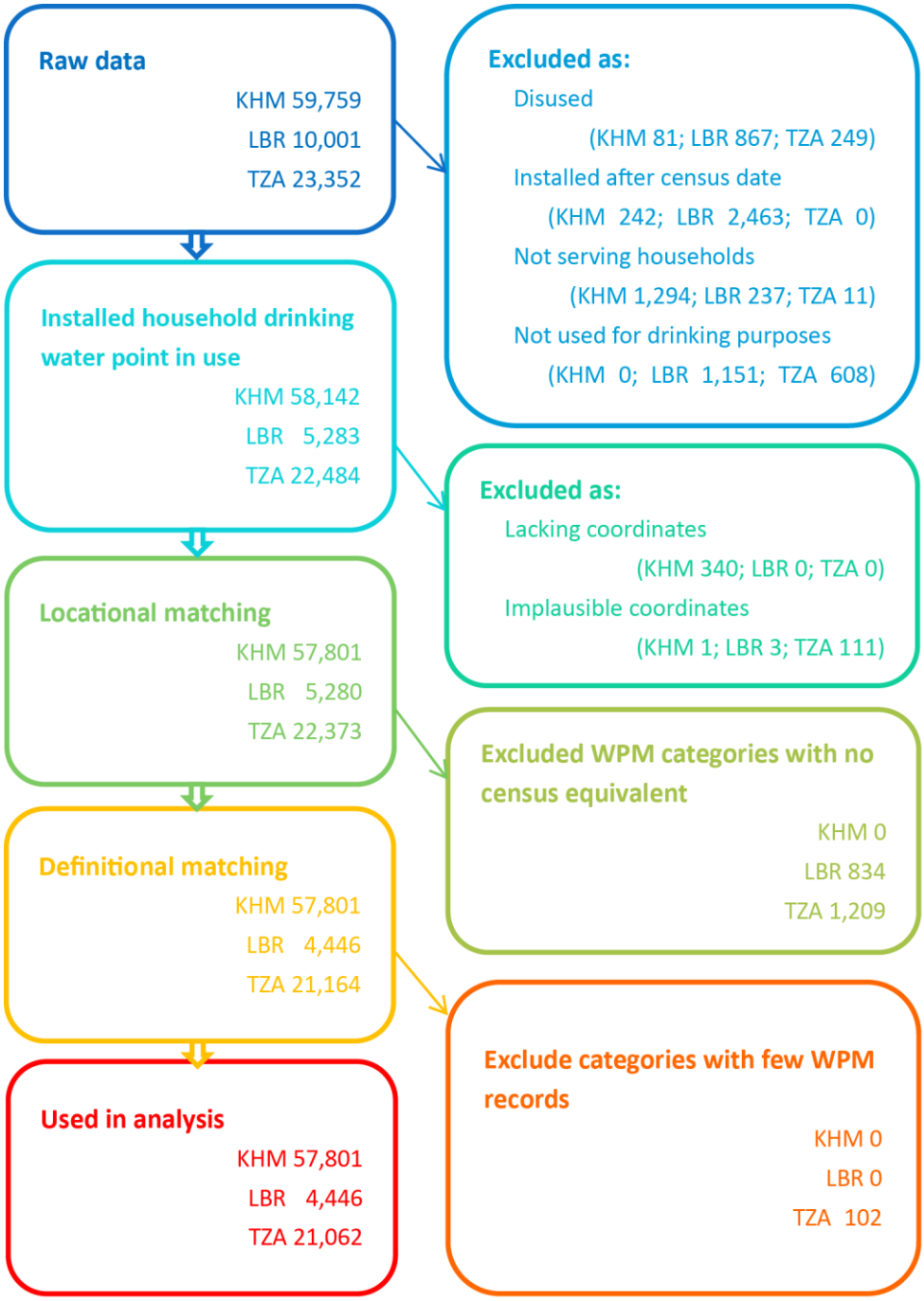

Fig: 2

Fig: 3

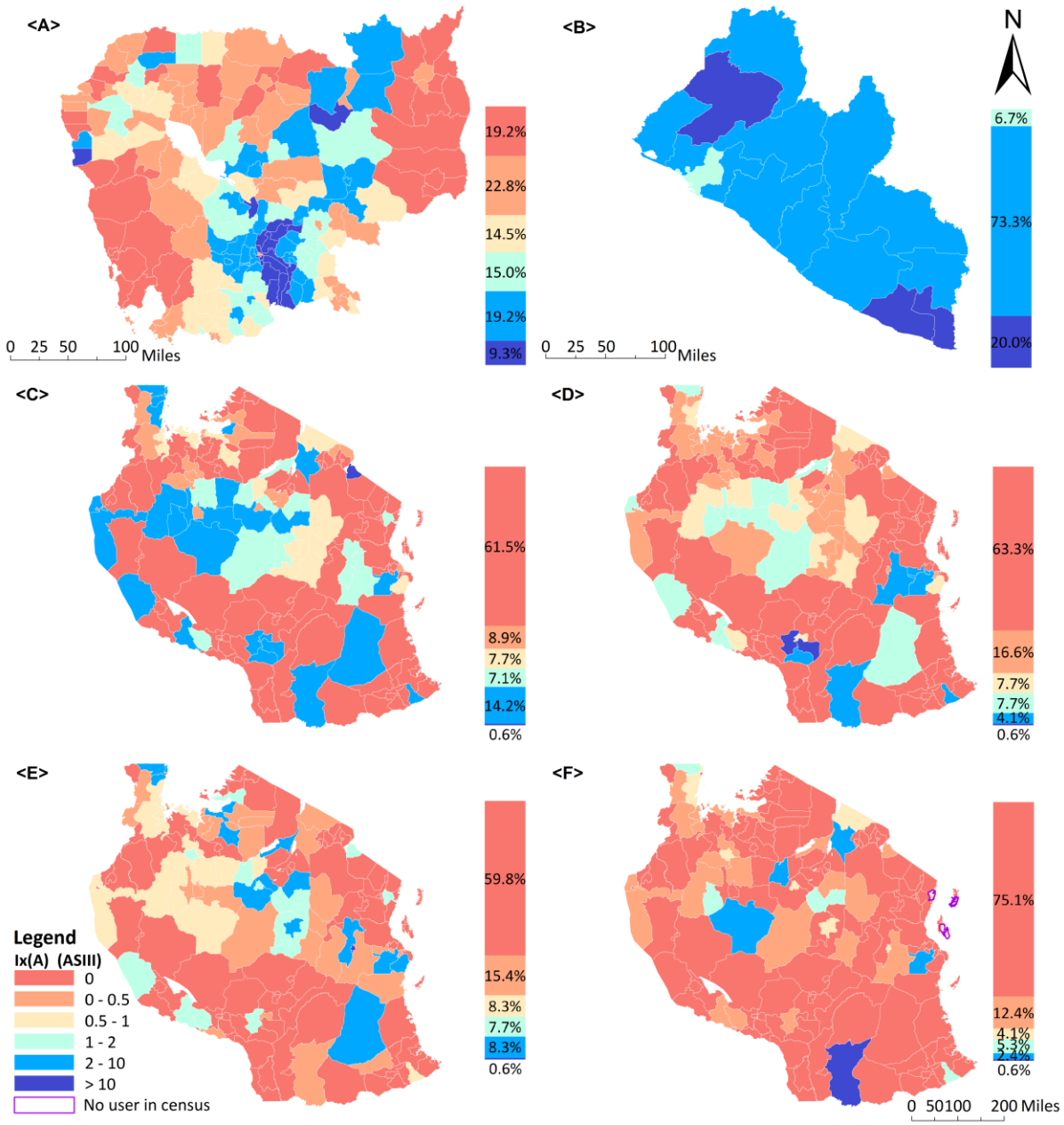

Fig: 4

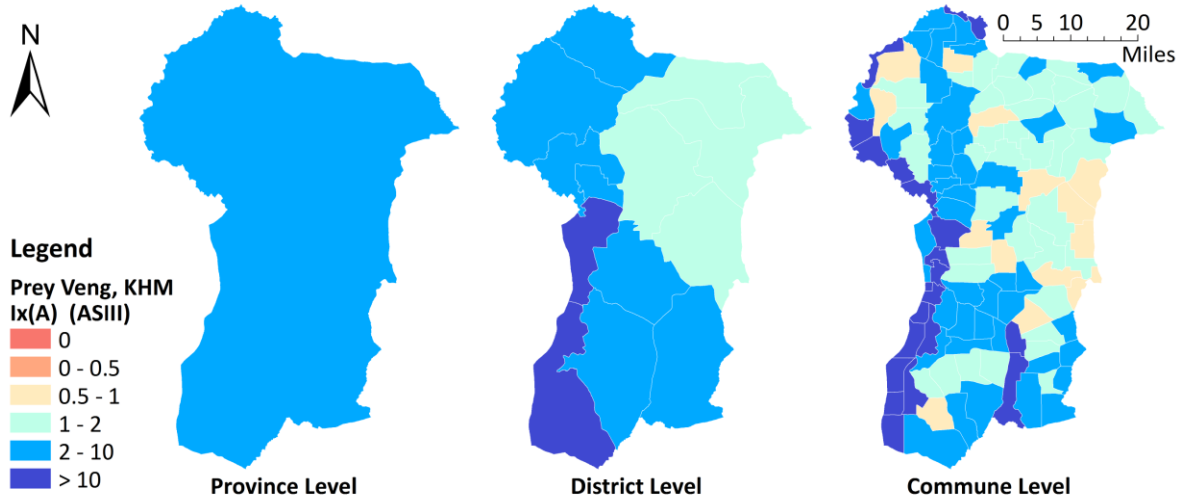

Fig: 5

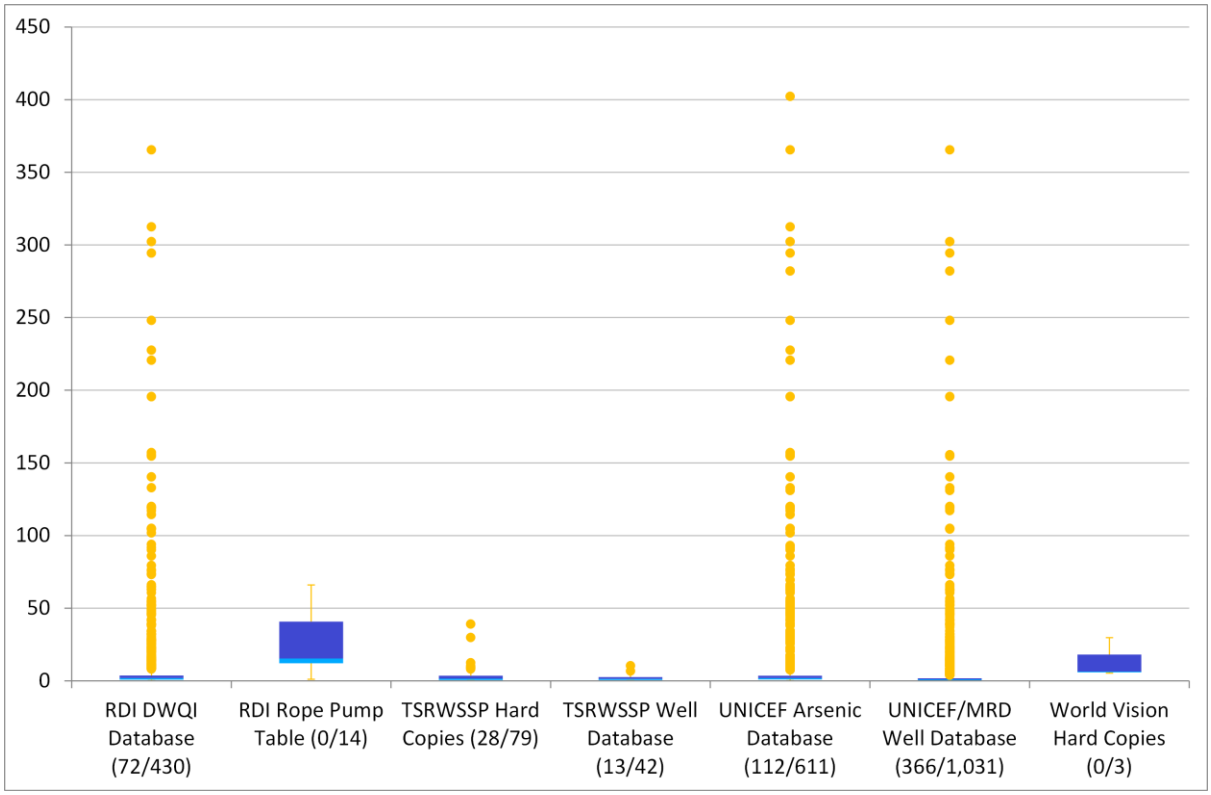

Fig: 6

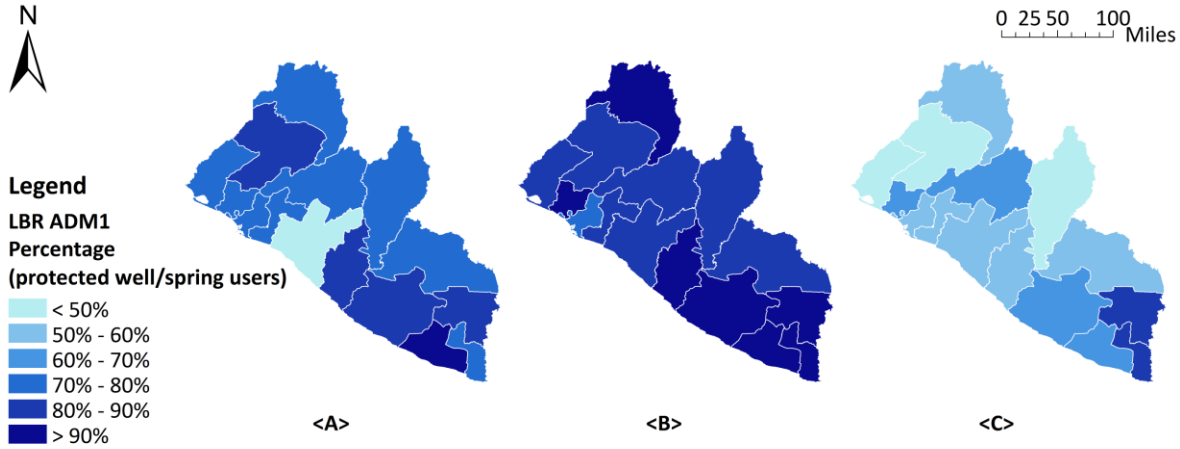

Fig: 7

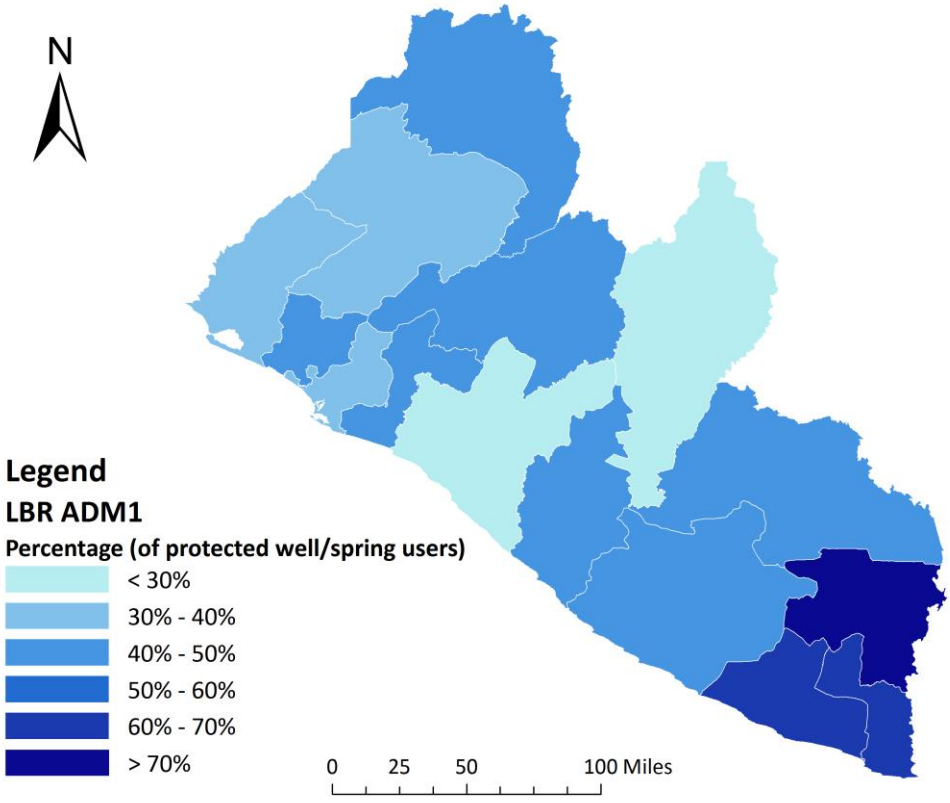

Fig: 8

Table 1

Assumed maximum population served (S_{xi}) for different types of water point and extraction/lifting methods in Cambodia, Liberia and Tanzania water point datasets

Water point type	Water lifting / extraction method	Maximum population served (S_{xi})
Standpipe	Single tap	70-150
Standpipe	Multiple taps	300-500
Well/borehole	Mechanised (powered) pump	500-5000
Well/borehole	Manual pump for heavy duty	250-300
Well/borehole	Manual pump for median duty	150-200
Well/borehole	Manual pump for light duty	70-100
Well/borehole	Water lifting with rope and bucket or similar	50-70
Well/borehole	Lifting method unclear	**See Table 2**
Spring	Protected spring	300

S_{xi} shows the range of assumed maximum population served; the specific S_{xi} number depends on the detailed lifting method, e.g. as a light duty manual pump, a rope pump serves 70 people, whilst a Tara pump serves 100 people.

Table 2

Example of assumptions about maximum population served based on depth for groundwater points lacking technical data or with unclear extraction methods

Assumption	Description	Deep well (≥25m)	Median well (7-25m)	Shallow well (<7m)
Assumption I (ASI)	Water lifting method with the largest capacity assumed to be employed	5000 (autopump) 300 (handpump)	300	150
Assumption II (ASII)	Water lifting method with the smallest capacity assumed to be employed	500 (autopump) 300 (handpump)	300	50
Assumption III (ASIII)	Regardless of water extraction technique, all groundwater points serve 300 people	300	300	300

Numbers for each well depth represent corresponding assumed maximum population served.

Table 3

Characteristics of water points included in Cambodia, Liberia and Tanzania

Characteristics	Cambodia	Liberia	Tanzania
Year of installation			
Pre-1990	3,180 (5.50%)	95 (2.14%)	7,426 (35.26%)
1990-1999	18,801 (32.53%)	270 (6.07%)	5,860 (27.82%)
2000-2009	32,426 (56.10%)	3,601 (80.99%)	7,507 (35.64%)
After 2009	Not applicable	Not applicable	0 (0%)
Unknown	3,394 (5.87%)	480 (10.80%)	269 (1.28%)
Water point type			
Drilled well	21,622 (37.41%)	0 (0%)	2,540 (12.06%)
Auto-pump (500-5000)	1 (<0.01%)	0 (0%)	102 (0.48%)
Heavy duty hand-pump (250-300)	7,151 (12.37%)	0 (0%)	2,029 (9.63%)
Median duty hand-pump (150-200)	0 (0%)	0 (0%)	4 (0.02%)
Light duty hand-pump (70-100)	4,516 (7.81%)	0 (0%)	70 (0.33%)
Hand-pump (other/unknown/unclear)	4 (<0.01%)	0 (0%)	242 (1.15%)
Other/unknown	9,950 (17.21%)	0 (0%)	93 (0.44%)
Dug well	955 (1.65%)	4,438 (99.82%)	5,833 (27.69%)
Heavy duty hand-pump (250-300)	0 (0%)	4,401 (98.99%)	4,273 (20.29%)
Median duty hand-pump (150-200)	0 (0%)	1 (0.02%)	0 (0%)
Light duty hand-pump (70-100)	287 (0.50%)	0 (0%)	3 (0.01%)
Rope & bucket/windlass (50-70)	0 (0%)	0 (0%)	566 (2.69%)
Hand-pump (other/unknown/unclear)	0 (0%)	2 (0.04%)	423 (2.01%)
Other/unknown	668 (1.16%)	42 (0.94%)	568 (2.70%)
Protected dug well	287 (0.50%)	4,396 (98.88%)	5,833 (27.69%)
Other/unknown well	35,224 (60.94%)	0 (0%)	0 (0%)
Improved spring	Not applicable	8 (0.18%)	614 (2.92%)
Standpipe	Not applicable	0 (0%)	12,075 (57.33%)

Standpipe – single tap	Not applicable	0 (0%)	8,129 (38.60%)
Standpipe – multiple taps	Not applicable	0 (0%)	3,940 (18.71%)
Standpipe – unknown	Not applicable	0 (0%)	6 (0.03%)
Total included water points	57,801 (100%)	4,446 (100%)	21,062 (100%)

Table 4

Descriptive statistics for the ratio of census-based population using different water sources to the maximum population potentially served by the recorded water points ($I_x(A)$) for Cambodia, Liberia and Tanzania by census administrative level

	KHM adm1	KHM adm2	KHM adm3	LBR amd1	TZA-p adm1	TZA-t adm1	TZA-d adm1	TZA-s adm1	TZA-p adm2	TZA-t adm2	TZA-d adm2	TZA-s adm2
N	24	193	1621	15	30	30	30	30	169	169	169	169
ASI (largest capacity-based assumption)												
N_i	9	81	702	0	27	24	27	28	131	135	141	155
P_i	37.5%	42.0%	43.3%	0.0%	90.0%	80.0%	90.0%	93.3%	77.5%	79.9%	83.4%	91.7%
Max	35.30	178.89	2093.7	27.98	1.76	3.54	1.56	1.00	8.19	35.29	13.27	10.04
Min	0.00	0.00	0.00	1.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg	5.13	7.60	23.46	8.06	0.41	0.57	0.35	0.20	0.67	0.80	0.50	0.30
SD	7.73	18.38	110.47	6.22	0.49	0.82	0.41	0.29	1.35	3.01	1.35	1.16
ASII (smallest capacity-based assumption)												
N_i	13	122	1002	0	30	28	28	28	144	152	145	159
P_i	54.2%	63.2%	61.8%	0.0%	100%	93.3%	93.3%	93.3%	85.2%	89.9%	85.8%	94.1%
Max	9.42	22.59	342.21	27.77	0.97	2.26	1.51	1.00	4.83	8.14	13.14	5.02
Min	0.00	0.00	0.00	1.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg	1.57	2.00	4.87	8.03	0.23	0.27	0.32	0.13	0.36	0.36	0.46	0.16
SD	2.36	3.96	19.12	6.19	0.27	0.47	0.39	0.25	0.74	1.11	1.32	0.59
ASIII (empirical number-based assumption)												
N_i	11	109	848	0	26	28	26	28	132	148	141	155
P_i	45.8%	56.5%	52.3%	0.0%	86.7%	93.3%	86.7%	93.3%	78.1%	87.6%	83.4%	91.7%
Max	16.63	53.22	402.29	27.98	2.87	5.20	1.97	1.00	14.83	13.29	14.72	10.04
Min	0.00	0.00	0.00	1.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg	2.40	3.37	7.53	8.06	0.46	0.41	0.42	0.20	0.78	0.51	0.60	0.30
SD	3.78	7.46	27.72	6.22	0.62	0.95	0.51	0.29	1.76	1.68	1.59	1.16

N represents the number of administrative units; N_i and P_i represent the number and percentage of administrative units that have incomplete water point coverage (in relation to census data); Max, Min, Avg, and SD respectively are maximum, minimum, mean values, and standard deviations of measured I_x(A)s; p, t, d, s following TZA represent data on public tap/standpipe, tube well/borehole, protected dug well, and protected spring respectively.