Comparing water point based and household survey based water access estimates with publicly available data

Nicolas Dickinson, WASHNote

13/06/2022

Contents

1 Abstract

The purpose of the study is to determine how water point level estimates for rural basic water service coverage from the Water Point Data Exchange (WPdx) compare to the household level estimates from the Joint Monitoring Programme of WHO/UNICEF (JMP) in multiple geographies. The study describes how these different estimates are produced and proposes the comparison of JMP basic minus services on premises to WPdx basic access. WPdx basic access estimates the population with 1km of an improved water point.

Comparing between metrics and triangulating different measured results can be useful to validate conclusions and inform decision making. This study finds a relatively strong correlation and linear trend between these two estimates in four countries that suggests that using household surveys and water point inventories together can be useful to decision makers who may only have one or the other data sources or may want to validate the conclusions from one against another.

The WPdx basic estimates allow a more granular geographical level of access estimates that can be useful to districts and enable national vulnerability assessments. This could strengthen the type of analysis provided in JMP inequality charts showing the differences between country regions. At the same time, further research is needed to validate these trends at these lower geographical levels.

Rural water leaders, including national and local governments, development partners, service providers and civil society should continue to advocate for the publication of water point data and the validation of access estimates on the basis of publicly available information. This plays an important role in improving the quality of both public and private data sets and analyses used by researchers and decision makers.

2 Abbreviations

Abbreviation	Full name
AIS	AIDS Indicator Survey
DHS	Demographic Health Survey
GHS	Nigeria General Household Survey
GRID3	Geo-Referenced Infrastructure and Demographic Data for Development
HDX	Humanitarian Data Exchange
JMP	Joint Monitoring Programme of WHO/UNICEF
MICS	Multiple Indicator Cluster Survey
MIS	Malaria Indicators Survey
NHS	Uganda National Household Survey
NORM	Nigeria Water, Sanitation, Hygiene National Outcome Routine Mapping
NPS	Uganda National Panel Survey
PMA	Performance Monitoring for Action Survey

Abbreviation	Full name
R^2	R squared, coefficient of determination: this represents the proportion of variation in the dependent variable that is predictable from the independent variables.
RW_1	The proportion of population using improved sources not exceeding 30 minutes collection time
RW_2	The proportion of population using improved sources which are accessible on premises
SDG	Sustainable Development Goal
SDG6	Sustainable Development Goal 6 aiming for "clean water and sanitation for all"
UNICEF	United Nations Children's Fund
W_1	The proportion of population that uses improved drinking water sources (all sources including piped) of the total population
W_7	The proportion of the population using basic drinking water services
$W_{7,!premises}$	The proportion of the population using basic drinking water services that are not accessible on premises
$W_{7,premises}$	The proportion of the population using basic drinking water services that are accessible on premises
WHO	World Health Organization
WPdx	Water Point Data Exchange
$W_{premises}$	The proportion of population using improved sources which are accessible on premises

3 Acknowledgements

This study could not have been possible without the contribution of open data on water points by data providers to WPdx. Members of the Water Point Data Exchange (WPdx) working group reviewed both the proposal and findings of this work. Katy Sill of WPdx first recognized the potential of the work, provided invaluable feedback, and responded quickly with explanations about how the WPdx algorithms work while investigating and delivering improvements to the tools when required to make this comparison possible.

Similarly, the National Statistics Offices (NSOs) and the Demographic and Health Surveys (DHS) Program of the United States Agency for International Development (USAID) made it possible to use household survey data from different countries. I would like to thank the Joint Monitoring Programme of WHO/UNICEF (JMP) team for sharing country, regional and global estimates of progress on drinking water, sanitation and hygiene (WASH) in households as well as the estimates for the sub-indicators required to generate those estimates, for providing clarifications about the JMP methodology, and for taking time to reflect on study findings

This material is based upon work supported by USAID under award number 7200AA18CA00033.

4 Introduction

Multiple sources of data are available to decision makers on the state of water access and services. There is relatively strong agreement that reliable data for decision making is needed. At the same time, it is not always clear which data sources are both available and appropriate to answer the questions about where and how to invest resources in water services and how to appropriately target the poorest. This study seeks to determine how water point coverage estimates based on publicly available data from the Water Point Data Exchange (WPdx) compare and contrast with the official Joint Monitoring Programme of WHO/UNICEF (JMP) figures. The goal is to provide recommendations about how these different estimates could be used in tandem and what their respective strength and limitations are.

In order to monitor progress towards national drinking water goals, national decision makers and administrators in water ministries have developed rural water coverage estimates, based on data on infrastructure, specifically the number of nationally acceptable facilities installed and the number of people who by design could have access to them. As the Millennium Development Goals (MDGs) aimed to half the number of people without improved facilities by 2015, national actors were encouraged to harmonize their coverage estimates and facility types with the JMP standard definition of an 'improved' facility. With the advent of the Sustainable Development Goals (SDGs), service level requirements have been added to international goals, such as spending less than 30 minutes fetching water to meet a 'basic' level of service. This is typically estimated on the basis of recall in household surveys. (JMP 2018) (JMP 2021) Over time, decision makers have had to work towards both incorporating new goals and definitions and tracking progress in a consistent way over time. New definitions, even improvements, create inconsistencies in monitoring indicators that can initially be difficult to explain to decision makers. As a result, national debates sometimes even put to question legitimacy of globally developed estimates versus national coverage indicators that had been developed and tracked in prior years by national ministries and implementation agencies. In general, decision makers have often sought to understand the difference between the two and seek for their alignment.

The Water Point Data Exchange (WPdx) is a data standard (WPdx 2021) for sharing water point data and a platform ("WPdx – The Water Point Data Exchange Is the Global Platform for Sharing Water Point Data," n.d.) which harmonizes data from disparate sources (government, NGOs) into a consolidated data set for analysis to help identify access and data gaps to prioritize investments. WPdx has taken the infrastructure-based approach similar to the approach for estimating national coverage figures. However, WPdx compiles data on infrastructure functionality and estimates the likely population served by functional water points based on their geographic location and high resolution population maps. ("WPdx Decision Support Tools," n.d.a) These additions promise a fine grained approach to estimating access to public water points in rural areas.

According to the JMP published global figures, as of 2020, it can be estimated that around 13% of rural households depend on improved water sources that are not on premises but within 30 minutes from their household. This represents 1 billion people in countries for which these figures are available. These figures underline the importance of public water points for a large percentage of rural population, even if ultimately, the SDG 6.1 target aims at safely managed services on premises. (*Jmpwashdata* 2022)

The JMP provides a globally agreed standard set of indicators and estimates for water supply in countries all around the world. Through an intensive partnership between the UN and the National Statistic Office (NSO) in each country, the JMP produces comparable estimates at national scale, disaggregated by rural and urban areas. Additionally, the JMP provides additional analysis of inequalities using quintiles and regional estimates from the data sources found in each country. These sources are representative surveys, census, and administrative data sets. These estimates, mostly derived from statistical studies, are often regarded as a gold standard for national level estimates in cross-country comparisons, but the JMP estimates are linear projections and do not disaggregate into sub-national regions. While useful for tracking and comparing national progress in the world, this limits the use of the JMP estimates in national decision making, especially within a country's sub-national regions.

National governments will typically use more detailed findings from the same statistical surveys and administrative data sources used by JMP, together with other available data sources, to measure their progress against national policy objectives, including SDG 6.1. These officially recognized data sources are updated, in the best cases, on an annual or rolling basis. At times, administrative sources are only updated when new projects install new infrastructure. In other cases, these figures are collected locally at decentralized levels and reported to the national level on a quarterly or annual basis, providing more complete picture.

The UN Water 2021 Summary Progress Update on SDG6 states that "credible and timely water and sanitation data provide numerous social, economic, and environmental benefits in both public and private sectors, such as stronger political accountability and commitment, as well as public and private investments. [High quality data] also enables evidence-based policy-making, regulations, planning and investments at all levels, to ensure the most effective deployment of resources." However, UN Water identifies specific data challenges including gaps in data collection and sharing and suggests that improved data generation, validation, standardization, and information exchange can build trust so leaders can make informed decisions and increase

accountability. The U.S. Global Water Strategy (U. S. 2017) identifies the lack of data for decision-making as a limiting factor in service provision sustainability and includes a strategic approach to promote common data exchange formats and access to data for decision-making to help support reaching Strategy objectives.

4.1 Purpose

The purpose of the study is to determine how WPdx (water point level) estimates for rural basic water service coverage compare to the JMP data (household level) in multiple geographies. The study will seek to understand the strengths and limitations of each set of estimates and parameters and the underlying data sources used to produce these estimates.

Comparing between metrics and triangulating different measured results can be useful to validate conclusions and inform decision making. This study will reflect on the applicability of using these estimates together for more insights and better decision making on rural water supplies.

4.2 Areas for comparison

In order to make the best comparison between access estimates from WPdx and JMP, countries have been selected where a national water point inventory or other form of asset mapping has been conducted and included in the WPdx tools database. In each country, there has also been a significant number of recent data sources for JMP estimates, primarily household surveys, which have reported on water services and specifically information used to determine whether or not at least a basic level of service has been achieved.

Tables 2 and 3 provide a summary of the water point data available in WPdx for the countries considered in this study. The number of water points updated in WPdx varies substantially per year.

Similarly, Table 4 provides a list of representative surveys that have provided information about rural areas used in the JMP estimates as of February 2022. There appears to be good overlap in the time periods for which surveys are available for JMP and water points data has been collected and reported into WPdx. However, while the JMP provides linear estimates projected across all years on the basis of these available surveys, WPdx at the moment only provides a single snapshot based on the latest water point data available at each site where data was collected and reported.

The WPdx data standard supports any water points regardless of its urban or rural status, however, the focus of this analysis is on rural areas for which WPdx tools provide service level estimates. Even so, this initial comparison may be of interest to evaluate whether including urban areas in future evaluations would be fruitful.

The WPdx geographical unit of analysis is more granular than that of DHS surveys in general. Due to the fact that access is calculated based on high resolution population grids, access estimates are calculated for multiple levels of administrative sub-divisions based on available shape files. See Figure 1

4.2.1 At least basic water supply as defined by JMP

While SDG 6.1 aims for universal access to safely managed drinking water, SDG 1.4 includes a target for universal access to basic services including at least a basic water supply. This study will compare at least basic water supplies, which are drinking water from an improved water source, provided that collection time is no more than 30 minutes for a round trip including queuing. Some of these supplies may also be safely managed but this data has not been compared during this study as on premise water supplies are not included in WPdx. Improved water supplies are those that have the potential to provide safe water in terms of their nature and construction and while the definition is well-established, interpretations vary per country based on local context and technologies. The JMP ladder categories are presented in Table 5. (JMP 2021)

The main source of data for JMP estimates are representative household surveys for measuring the use of water services by households in a country. These surveys typically include questions on the main or primary

Table 2: Data sources and number of times an update to WPdx+ water points has been made per country

Country	Data sources (number of water point reports)	Total updates
Eswatini	eSwatini (Swaziland) Department of Affairs (12842), WaterAid Eswatini (94), WaterAid UK (28), WaterAid Swaziland (8)	12,972
Liberia	WASH Liberia (27299), Living Water International (734), WaterAid UK (20)	28,053
Nigeria	Federal Ministry of Water Resources, Nigeria (88543), GRID3 (10738), iMMAP (425), Living Water International (37), WaterAid UK (19), WaterAid (18), Water Mission (4)	99,784
Sierra Leone	Ministry of Water Resources, Sierra Leone (49974), Ministry of Basic and Senior Secondary Education - Sierra Leone (3882), GOAL (1858), Inter Aide (1443), UNICEF (327), Living Water International (84), CRS RAIN Project (71), Kenema Water Directorate (59), WaterAid UK (22)	57,720
Uganda	Ministry of Water and Environment, Uganda (82671), Evidence Action (5913), Ugandan Water Project (1702), C & D (1646), Lifeline (1219), Water For People (1087), Water for People (878), IRC (782), Living Water International (576), World Vision (575), TTC Mobile (385), UNHCR (321), Water Mission (71), The Water Trust Uganda (60), CARE-Uganda-Otuke (27), Drop in the Bucket (13), WaterAid UK (6)	97,932

Table 3: WPdx+ water points reported by year after de-duplication

Year	Eswatini	Liberia	Nigeria	Sierra Leone	Uganda
2010	0	504	31	1,799	74,321
2011	0	9,418	0	0	2,223
2012	0	0	0	28,182	4,272
2013	859	35	69	5	3,977
2014	7,241	115	81	41	4,985
2015	4,740	2,639	88,388	26	3,221
2016	1	109	11	21,804	681
2017	0	14,807	0	0	1,010
2018	7	66	10,236	213	230
2019	1	262	936	4,201	1,399
2020	121	73	28	1,443	492
2021	0	10	0	0	1,045
2022	0	0	0	0	57

Table 4: JMP data sources by year with information on household water use within a 30 minute round trip to the water supply in rural areas (see table of abbreviations names of surveys)

Year	Eswatini	Liberia	Nigeria	Sierra Leone	Uganda
2000	MICS			MICS	
2001					DHS
2003	WHS		DHS		
2005				MICS	
2006					NHS, DHS
2007	DHS	DHS	MICS		
2009					NHS
2010				MICS	
2011		MIS			AIS, DHS
2013		DHS	GHS, DHS	MIS	
2014	MICS				NPS, PMA
2015		HIES	MIS		MIS, PMA
2016		MIS	GHS, PMA	MIS	NPS, DHS, PMA
2017			PMA, MICS	MICS	PMA
2018			PMA, NORM, DHS	IHS	PMA
2019			NORM	DHS	MIS, NPS

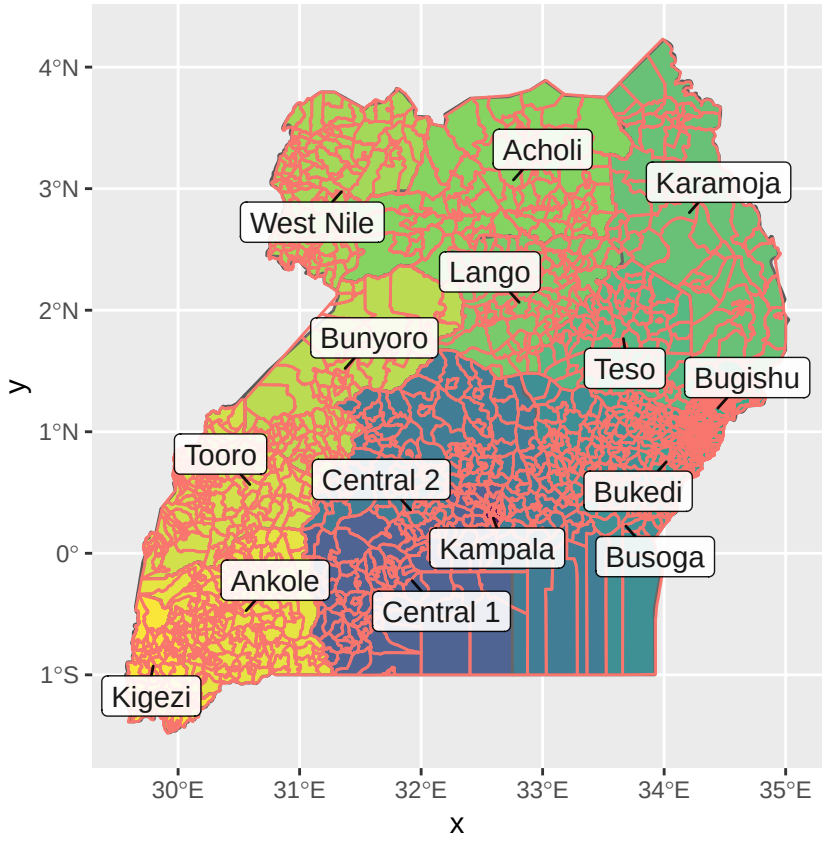

Figure 1: The DHS regions in Uganda (labeled) are larger than the WPdx regions in red.

Table 5: JMP drinking water ladder definitions for each level of service

Category	Definition
Safely managed	Drinking water from an improved water source that is accessible on premises, available when needed and free from faecal and priority chemical contamination
Basic	Drinking water from an improved source, provided collection time is not more than 30 minutes for a roundtrip including queuing
Limited	Drinking water from an improved source for which collection time exceeds 30 minutes for a roundtrip including queuing
Unimproved	Drinking water from an unprotected dug well or unprotected spring
Surface water	Drinking water directly from a river, dam, lake, pond, stream, canal or irrigation canal

water source used by the household. Over the years, the type of questions and information collected in household surveys has largely been standardized and yet again improved with the advent of the JMP and SDG 6.1 indicator definitions. Perhaps the best known household surveys that address household water services may be the DHS and the MICS, implemented by National Statistics Offices. These definitions are implemented in MICS and DHS surveys in many countries around the world and are a critical data source for the JMP estimates.

In order to avoid the challenge associated with using national studies using different definitions, this study will use the household surveys results as shared by JMP in the country inequalities files as these provide a level of geographic disaggregation down to survey regions. As DHS survey region geographical shape files are available publicly online, the DHS surveys were used as the primary source of JMP data points at sub-national level.

4.2.2 National coverage estimates

Coverage estimates at national level can be calculated in different ways. Typically ministries and agencies that track the implementation of rural water supply programs keep lists of facilities that have been installed. They often also have a standard list of design populations expected to be served by each facility. These design populations can be summed up and divided by the total population in an area to come up with a 'coverage' estimate. Other variations will include refinements like taking into account the functionality of the facilities and capping the population served if the population around the water point is lower than the design population. While it is often possible to track these parameters on the basis of administrative data alone from project completion reports, site visit reports and quarterly or annual reporting, it can be difficult to track functionality and population movements with regular frequency.

However, since the MDGs, the SDGs have introduced important new targets about the level of service beyond only access to a water point. For the basic service targets of SDG 1.4, it is now important to consider travel and queuing time. For the SDG 6.2 targets, it is additionally critical to take into account water quality, the availability of water when needed and whether access is on premises for households.

There are also clear limitations to any type of infrastructure inventory based estimates if they are not systematically updated. At times water points are covered or abandoned and this may not be updated in the inventory. The age of the last report during a site visit or spot check can vary enormously and can affect also the quality or validity of these estimates when the data is too old. Additionally, new infrastructure may replace existing systems but this may not be reflected in the inventory. Aside from methodological issues, it is expected that there will be a margin of error in any of these estimates based on how well managed the inventory is. This error may be harder to estimate in comparison with the the statistical power of the household survey estimates that vary based on some known factors such as the sample size and effect size.

4.2.3 WPdx access estimates

For public water points and community water supplies, the WPdx estimates take steps to address some of the limitations found in some national coverage estimates. WPdx basic access addresses the lack of round trip time estimates and by using high resolution population grids and a 1km radius as a proxy for the population served within a 30 min round trip. The analysis does not currently take into account how overcrowding may affect queuing. The high resolution population grid enables more granular and efficient comparisons of the amount of people served per water point rather than per administrative area with available population data. WPdx is a source of data that can be updated by governments, NGOs and service providers, and thus has the potential to have more frequent functionality updates. As such, it may provide a bridge between the basic water supply definitions and coverage estimates specifically for communities that depend on public water points.

In brief, the WPdx access algorithm can be interpreted as an estimated minimum population with access to water from an improved public water source within a 1km radius. Data from different sources are combined into a single data set with internal data validation to ensure consistency with the WPdx standard and a

statement that the publisher has the right to share the information. Overlapping data points, whether from the same or different data sources, are put through a de-duplication and WPdx validation algorithm. There is no formal external validation of the combined data sources although some come from official sources such as a national ministry of water. Water points are classified by source and technology type and a table of specific design populations have been used as the maximum possible population served within 1km of each water point. These maximum capacity populations are kept constant across each system and country for the moment.

The following maximum capacity values extended from Sphere Guidelines and Yu et al. (2017) are currently in use:

- 100 people per rainwater catchment system
- 250 people per tap [tapstand, kiosk]
- 300 people per protected spring
- 400 people per hand pump [most hand pumps, lower for some such as Walimi, Tara, rope pump, EMAS, etc.]
- 1000 people per mechanized (powered) well

If the local water supplies either outperforms or underperforms the maximum capacity population used by the WPdx algorithm, this may either under or over estimate access to at least basic water services. There can be local and regional and even national variations as a result but it is expected to be an acceptable approximation. Water points that are clearly unimproved, such as unprotected springs or surface water, have been excluded.

Several different variables are generated from the access estimate and summed over the administrative area:

- **Rural population with basic access**: The number of people served based on the technology type and the population within 1km up to the maximum capacity as shown above.
- **Rural population without basic access**: The number of people within 1km of a non-functional water supply technology from a public point source. Non-functional in this context means that the water point was reported as not having any water flowing when visited.
- **Rural population uncharted**: Population which is not within 1km of a reported water point technology and/or is within an urban area.
- **Overcapacity population**: The number of people within 1km of the water point technology over the capacity of that source type.

WPdx faces some of the same limitations as national coverage estimates. As stated on waterpointdata.org, "functionality datasets represent a snap-shot at one point in time. They do not indicate whether sources identified as non-functional will be fixed the next day, the next week or never – sources identified as non-functional are not necessarily permanently out of service. . . ".

While accepting data from multiple sources is a strength, at the same time, "data is provided 'as-is'. No additional validation or verification is done by WPdx. Lastly, data on WPDX has been uploaded by multiple sources and may not be statistically representative of national water point functionality." (https://www.waterpointdata.org/resources/#FAQ)

4.3 Aligning national water service access estimates

Different estimates of water access from different sources of information on water access have sometimes acted as a source of confusion for decision makers and stakeholders who struggle to understand the difference between each metric. As a result, national governments and the JMP have sought to align the national frameworks and the SDG target indicators in order to reduce these differences and reconcile national standards and international standards in a coherent manner.

Definitions and methodologies matter when aligning data sources. National household surveys results can differ from JMP estimates when different definitions are used for key concepts such as "improved sources" or different parameters are used to assess access. In some countries, a distance metric may be used instead of a time metric or surveys may collect responses about time using different time ranges other than 30 minutes and these may or may not include queuing. As a result, to develop internationally comparable SDG estimates, the JMP must align the data from national sources with the SDG target indicator definitions. The National Statistics Office and JMP review each data source before including it into these estimates of access. Some data sources used at national level may remain unused in JMP estimates if the data somehow is incompatible with the international standards. This study has only used data sources included by JMP in order to be able to consistently assess WPdx estimates against household survey results across different countries.

The lessons for this study, while specific to WPdx are also relevant for alignment at national level with regard to coverage estimates. For example, when national inventories of water facilities are used to track the outputs of their water supply infrastructure programs and estimate access based on the number of systems installed, these coverage figures and overall access trends may be expected to differ, sometimes substantially, with those reported by the JMP and also those reported in national household surveys by the National Statistics Office. This study can help answer the question of how closely inventory based coverage figures may be aligned with household based survey results.

4.4 Rural public water points

The use of public water points reported in WPdx is expected to be lower in areas with household connections to piped supplies and other water sources on premise such as a hand pump in a compound. However, these rural households would still be counted if they are within 1km of water point. The difference between household surveys and WPdx access estimates is expected to be sensitive to the behaviors of household water users (their choice of primary water access points), access to piped water systems and household connections. Findings of this study may also be sensitive to the type of water point data that has been collected in a particular area or by a particular entity reporting to WPdx.

In order to reduce these differences and to focus on the greatest added value of the WPdx estimates, this study focuses specifically on rural public water points.

5 Methodology

5.1 Data sources and tools

The study methodology consists of gathering the publicly available data primarily from WPdx, JMP and DHS in order to run the comparison of WPdx and JMP estimates.

Data sets used include :

- WPdx+: a de-duplicated, cleaned and enriched list of water points reported and the data source for access estimates (https://data.waterpointdata.org/dataset/Water-Point-Data-Exchange-Plus-WPdx-/eqje-vguj/data)
- WPdx access estimates: the "Adm Region Analysis" tool data set of access estimates per administrative region in each country as well as other indicators such as total population and rural population derived from third party data such as the Meta high resolution population grid and EU Global Human Settlement database. (https://tools.waterpointdata.org/)
- JMP country files and inequality files: detailed survey source data used to calculate the JMP water service level estimates in the study countries and the DHS codes of the data sources used in this study. The inequality files were used a source of sub-regional data. These were access using the jmpwashdata R package. (https://github.com/WASHNote/jmpwashdata and https://washdata.org)

Table 6: WPdx rural access figures (percent)

Country	Basic access (1km)	Non-functional WP (1km)	Uncharted access
Eswatini	41.6	10.7	47.6
Liberia	74.3	7.2	18.5
Nigeria	29.1	20.8	50.1
Sierra Leone	61.4	16.3	22.3
Uganda	61.9	12.1	26.0

Table 7: JMP rural service figures

Country	Safely managed	Basic	Limited	Unimproved	Surface water
Eswatini	NA	62	12	12	13
Liberia	NA	64	7	3	26
Nigeria	17	60	7	21	11
Sierra Leone	9	51	5	23	21
Uganda	7	46	32	15	6

- Urbanization and population figures from the World Bank databank to validate data sources used by JMP using the WDI R package.
- DHS survey metadata - DHS survey GIS boundary files - GADM administrative region boundary files: shape files prepared by GADM hosted at UC Davis. (https://gadm.org)
- OCHA administrative region boundary files: files prepared by OCHA and available via the Humanitarian Data Exchange (HDX). The rhdx R package was used to download these files per country.

R and RStudio were used to write and run the analysis. The source code is available and all packages and package versions have been recorded using the renv R package.

5.2 Deriving a comparable estimate

As WPdx is focused on rural public water points, a number of rural basic water services usually included within JMP definitions of basic services are not necessarily reported in WPdx data set. For example, packaged or delivered water is considered improved if households use an improved source for cooking and handwashing. Household connections and most improved sources available on premises are most often not included in WPdx. Additionally, private water sources that are off premises may also not be available in the WPdx repository if not captured in a national inventory of public water points. This study assumes that if these water "points" are included in WPdx+, it is only in a handful of records.

Indeed, when comparing national WPdx basic access indicator estimates with basic water as reported in the JMP world files, there is quite a large variation and no clear correlation between the two.

Table 8: Comparison of rural access figures

Country	WPdx basic access	JMP Basic	Percent difference
Eswatini	42	62	20
Liberia	74	64	-10
Nigeria	29	60	31
Sierra Leone	61	51	-10
Uganda	62	46	-16

In order to better compare these two estimates, this study will recalculate the JMP indicator in each DHS region in each country to exclude the population which has access on premises. This will not address all differences in methodology but is expected to address many of the excluded sources with some variation across countries based on service provision models in use and household practices.

The JMP calculates the access to basic services on a number of parameters derived from household surveys. These include:

- W_1: the proportion of population that uses improved drinking water sources (all sources including piped) of the total population
- RW_1: the proportion of population using improved sources not exceeding 30 minutes collection time

W_7 or basic drinking water services are calculated by JMP as follows:

$$W_7 = W_1 * RW_1$$

Additionally, the JMP also estimate the following ratio in order to estimate the proportion of safely managed services (on premises, available when needed, and free from faecal and priority chemical contamination):

- RW_2: the proportion of population using improved sources which are accessible on premises

This study additional defines a new parameter:

- W_{premises}: the proportion of population using improved sources which are accesible on premises

$$W_{7,premises} = W_1 * RW_2$$

Using these parameters, available in JMP country inequality files, it is possible to estimate the proportion of the population with access to improved water sources that are not on premises ($W_{7,!\text{premises}}$) and could possibly be captured by WPdx. This is a new estimate created for the purpose of this study:

$$W_{7,!premises} = W_1 * RW_1 - W_1 * RW_2 = W_1 * (RW_1 - RW_2)$$

In this study, the JMP inequality file country region parameters derived by JMP from DHS surveys are used to calculate this proportion. Each region parameter represents both rural and urban population of that region. It is expected that urban areas will have more on premises services.

For this study, it is not possible to disaggregate between rural and urban population in a region with the provided JMP data. Rather, the proportion $W_{7,!\text{premises}}$ was multiplied by the total rural population as given by WPdx in order to get an estimated population of rural people with services within 30 minutes but not on premises. The assumption is that as a region has a larger urban population $W_{7,!\text{premises}}$ will shrink as the number of people with on premises services increases. This should reduce the bias from using a data source that includes urban populations. However, it could also lead to an underestimation of rural $W_{7,!\text{premises}}$ in areas that are partially urban.

In order to examine whether this may be an issue, this study is able to divide regions up into a group of mostly rural regions and mostly urban regions. The expectation is that the fit between the JMP derived off-premises basic services and the WPdx access within 1km would be closer in mostly rural areas.

This new derived proportion of water services should provide a relatively comparable estimate to the use of public drinking water sources as mapped by WPdx. This new estimate, as well as, other proportions are converted using population figures in order to run linear regressions to evaluate if they are indeed linearly correlated with one another.

The model used is the following, where $\text{WPdx}_{\text{population}}$ is the population with WPdx access within 1km:

$$W_{7,!premises,population} = a + b * WPdx_{population}$$

This is not a perfect estimate. There will be households close to public water points who will access services on premises and so should feature in the WPdx access population. However, if we can assume that the majority of population on premise are clustered together (such as in a city), then it may still be a good estimate.

It is possible to turn the model and predict the $\text{WPdx}_{\text{population}}$ variable on the basis of the JMP parameters, basic and on premises, and the urban and rural population according to WPdx. We can also see if there is an interaction between the urban population and the on premises population in the region. This model uses the total population of the region according to WPdx to convert the W_7 and $\text{W}_{\text{premises}}$ into a population figures:

$$WPdx_{population} = a + b*W_{7,population} + c*W_{premises,population} + d*ruralpop + e*urbanpop + f*W_{premises,population}*urbanpop$$

This model provides an opportunity to test the way that each of these parameters (with the model estimates for a, b, c, d, e, and f) are related to $\text{WPdx}_{\text{population}}$ variable and validate some of the assumptions used in the simpler model.

This study does not use a model, such as a beta regression (Cribari-Neto and Zeileis 2010), which can work with proportions and rather estimated the populations associated with each proportion. Future studies may choose to further examine the impact of urbanization and how a 10% urban vs. a 90% urban area may impact findings.

5.3 Considerations regarding population, urbanisation, sampling and country definitions

The JMP use the following sources for population and urbanization figures in all countries and available from the World Bank DataBank:

- For population, the UN Population Division's World Population Prospects: 2019 revision,census reports and other sources. For more details see: https://databank.worldbank.org/reports.aspx?source=2&type=metadata&series=SP.POP.TOTL
- For urbanization, United Nations Population Division. World Urbanization Prospects: 2018 Revision. See: https://databank.worldbank.org/reports.aspx?source=2&type=metadata&series=SP.URB.TOTL.IN.ZS).

DHS surveys are developed by National Statistics Offices on the basis of a representative sample of the total population per sub-region. Sample sizes are typically chosen on the basis of census data and other data sources or factors deemed appropriate by national statisticians within the local context. Other factors can be related to seasonality of data collection, cost, and risks to enumerators and can lead to the exclusion or less representative samples of some study areas. While samples and weights are based on these factors, the parameters estimated by JMP are expected to be representative of the proportion of the total population. Due to differences in methodology, JMP population estimates are not likely to match exactly estimates used by the National Statistics Office in country.

WPdx uses the high resolution population grid provided by Meta's (previously Facebook) Data for Good initiative, which has a 30 meter x 30 meter grid resolution and is available from the Humanitarian Data Exchange for each country. WPdx tools use an urban geographic grid to calculate the level of urbanization and exclude urban areas from their access estimates (see Figure 2). The urbanization GIS raster comes from

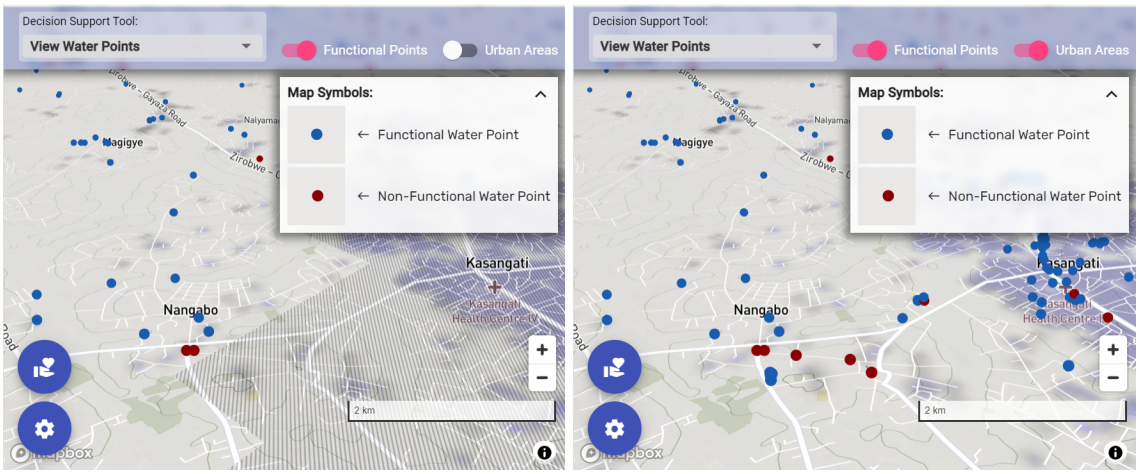

Figure 2: WPdx uses the 1km by 1km urban grid to mask out the 30m x 30m population grid shown here together with water points in the decision support tool.

the EU Global Human Settlement Database albeit with a much less granular resolution of 1 km x 1 km (see https://ghsl.jrc.ec.europa.eu/ghs_stat_ucdb2015mt_r2019a.php).

In order to derive comparable populations with access to rural water supplies and services, all proportions from JMP have been multiplied by the population of that area from WPdx. These population figures are necessarily derived from WPdx as JMP and DHS surveys only offer a proportion of people for each service level parameter. It is not expected that this will greatly affect findings but nonetheless the total country population differences between JMP and WPdx are compared in this analysis.

Similarly, rural and urban have varying definitions in each country so that both figures used by JMP and WPdx estimates may not align with the figures in country. The detailed comparison of these nationally defined methods for data collection and analysis in each country is beyond the scope of the current study. Rather it is proposed that if one can correlate strongly the JMP and WPdx estimates, WPdx estimates may be used as a proxy to further disaggregate access and service level trends at sub-regional level in a way that is comparable across countries. Ultimately this should lead to a better understanding of off-site basic service provision at sub-national level, which may be critical to the most vulnerable populations. Even so, these potential insights should not be construed as the 'actual' rural population served due to differences in methodology in each country. A direct comparison with national figures would require a detailed review of the differences in definitions and methodologies.

5.4 Administrative divisions and the selection of surveys for comparison

Administrative divisions, both in terms of official boundary shapes and names or codes are notoriously difficult to harmonize across different data sources, especially over time. For this reason, shape files for each individual country and data source have been used in order to ensure a geographically sound comparison rather than basing these on name alone or using single set of boundary files. These files include the specific shapes for each DHS survey and the administrative boundaries used to aggregate populations served by region. While figures are provided by JMP for each DHS region, WPdx figures are aggregated by identifying each administrative area that fits into each DHS region.

At sub-national level, the R jmpwashdata library, which consolidates the JMP country inequality files, made it possible to consolidate the JMP service level parameters and indicators used in each region of each country. In the case of DHS surveys, it is possible to associate these values to the geographical boundaries available in the publicly available DHS shape files. In those cases, total WPdx estimates for these larger areas were calculated.

Table 9: WPdx geographical data sources for boundaries

Country	Source
Eswatini	OCHA
Liberia	GADM
Nigeria	GADM
Sierra Leone	GADM
Uganda	OCHA

Table 10: DHS surveys found in JMP inequality files and the number of households and regions included

Country	Survey	Year	DHS households	Regions
Eswatini	DHS	2007	4,843	4
Liberia	DHS	2020	9,068	5
Liberia	DHS	2013	9,333	5
Liberia	MIS	2011	4,162	6
Liberia	MIS	2009	4,162	6
Liberia	DHS	2007	6,824	6
Nigeria	DHS	2018	40,427	6
Nigeria	MIS	2015	7,745	6
Nigeria	DHS	2013	38,522	6
Nigeria	MIS	2010	5,895	6
Nigeria	DHS	2008	34,070	6
Sierra Leone	DHS	2019	13,399	5
Sierra Leone	DHS	2013	12,629	4
Sierra Leone	DHS	2008	7,284	4
Uganda	MIS	2019	8,351	15
Uganda	DHS	2016	19,588	15
Uganda	AIS	2011	11,340	10
Uganda	DHS	2006	8,870	9
Uganda	DHS	2001	7,885	4

Identifying the correct DHS survey and boundary shape file was achieved by using the names of the data sources together with the year of the data source given in in the jmpwashdata R package to match to the corresponding DHS survey. Matching was possible for Demographic Health Surveys, Malaria Indicator Surveys, and AIDS Indicator Surveys registered in the DHS database. Any year during the survey period (start or finish) was used to match the survey year identified by JMP as long as it identified only one unique match.

Table 9 shows the sources of administrative boundary shape files selected by WPdx. The sources differ per country based on the experience of partners in country and feedback about which match more closely current boundaries. The OCHA boundaries were downloaded from HDX and the GADM boundaries were downloaded from the GADM website.

Table 10 shows surveys that were comparable with the WPdx access estimates on the basis of having shape files available from DHS and JMP parameters calculated in the JMP inequality files. Table 11 shows the years with the most water points reported. Eswatini has one matching survey from 2007, while the majority of WPdx water points have been reported between 2013 and 2015. Liberia has five surveys in the period from 2007 to 2020, and a large number of water points reported since 2010 and a peak in reporting in 2017. Nigeria has five surveys in the period between 2008 and 2018. Sierra Leone has three surveys in the period from 2008 to 2019 and extensive water point reporting in the years 2012, 2016 and 2019. Finally, Uganda has 5 surveys from 2001 to 2019 and 90,000 water points reported in 2010 and in the order of 5,000 each subsequent year. As a result, these countries provide some interesting points of comparisons with recent

Table 11: Top three reporting years in WPdx per country

Country	Year	Number of water points reported
Liberia	2017	14,807
Liberia	2011	9,418
Liberia	2015	2,639
Nigeria	2015	88,388
Nigeria	2018	10,236
Nigeria	2019	936
Sierra Leone	2012	28,182
Sierra Leone	2016	21,804
Sierra Leone	2019	4,201
Eswatini	2014	7,241
Eswatini	2015	4,740
Eswatini	2013	859
Uganda	2010	74,321
Uganda	2014	4,985
Uganda	2012	4,272

Table 12: All WPdx regions in Sierra Leone that overlap less than 95 percent with a single DHS region

WPdx region	DHS region	Most overlap with a single DHS region
Eastern - Kailahun - Dia	Eastern	92.8
Eastern - Kailahun - Kissi Kama	Eastern	92.3
Eastern - Kailahun - Kissi Teng	Eastern	91.7
Eastern - Kailahun - Kissi Tongi	Eastern	90.1
Eastern - Kailahun - Luawa	Eastern	94.9
Eastern - Kenema - Nomo	Eastern	92.2
Southern - Bonthe - Imperi	Southern	91.4

surveys with the exception of Eswatini where the survey precedes the water point mapping by at least 5 years. Even so Eswatini has been kept in the analysis.

Based on peak years, in Nigeria figures are compared with the 2015 MIS survey and separately with 2018 DHS survey that has a higher sample size. Liberia is compared with with DHS 2020, Sierra Leone with DHS 2013, and Uganda with AIS 2011 and DHS 2016 which has a higher sample size.

Generally, there has is a good overlap between each WPdx region and DHS regions. As shown in Figure 3 and 12 when there is less overlap, this tends to be on country borders or regions with water bodies but is not expected to effect results.

6 Findings

6.1 Population and urbanization

The analysis used WPdx population to compare different access estimates at sub-national level. This provides a consistent population figure to compare access ratios. This section notes some differences in overall population figures between WPdx figures, derived from the Meta's data for good geospatial population grid, and JMP figures, derived from UN population and urbanization sources. It is worth noting these UN sources do not provide sub-national population figures and are not geospatial in nature. The detailed analysis is summarized in the tables in the appendix.

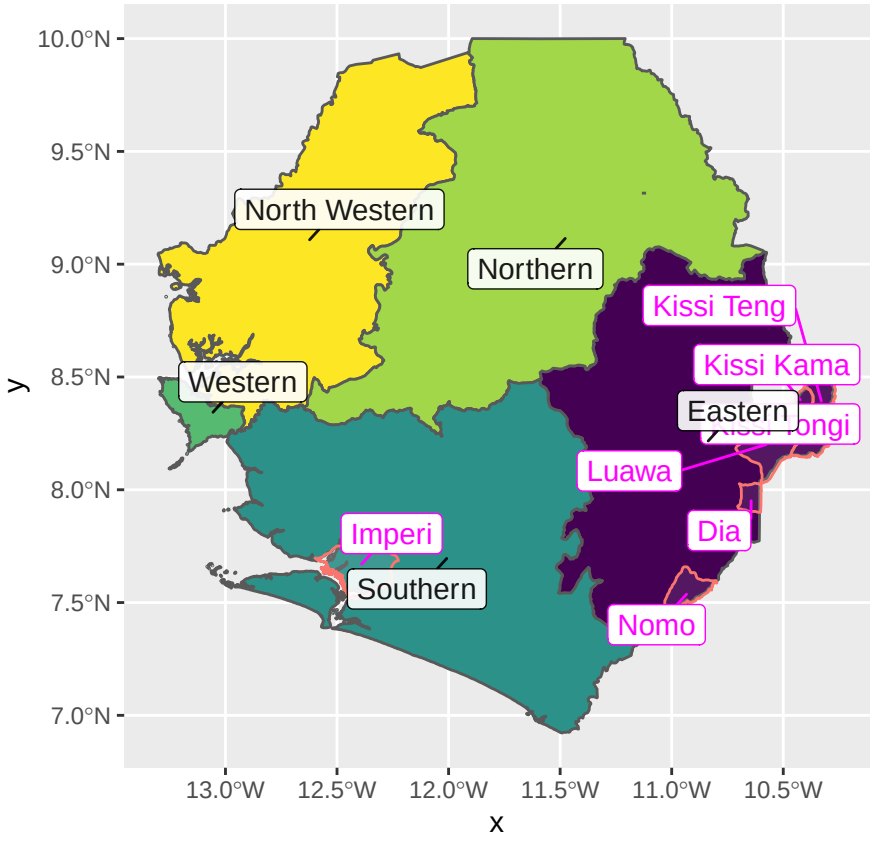

Figure 3: Map of areas with less than 95% overlap between WPdx and DHS.

The WPdx total national population is close to that used by JMP in Nigeria and Sierra Leone but differs in other countries as shown in Table 19 in Appendix A. This table also shows that WPdx predicts lower levels of urbanization than the UN prospectus.

With the exception of Uganda, the WPdx estimated rural population is 26.4% to 18.2% lower than the rural population estimated from the UN data sources. In Uganda, the WPdx national rural population estimate is rather 4% higher than estimates from UN data.

6.2 Comparison of WPdx basic access and JMP basic off premises

There appears to be a linear trend when plotting the different estimates per country and correcting for the proportion of basic water services on premises (see Figure 4).

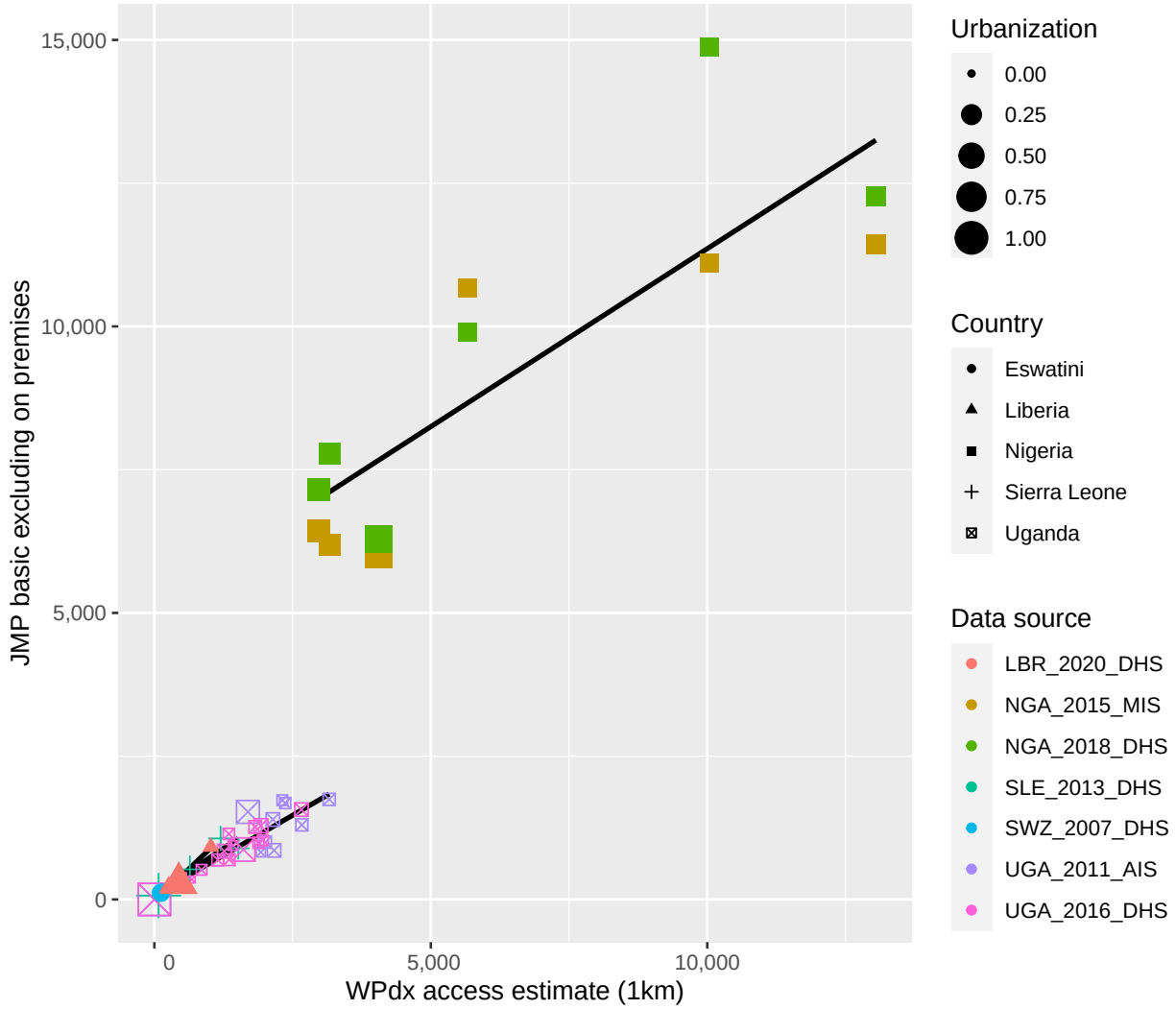

Figure 4: WPdx access and JMP access appear to have linear trends.

A correlation is confirmed using the Pearson correlation for a survey in each country. For Nigeria and Uganda, the DHS survey from 2018 and 2016 are used respectively to calculate correlations. Overall, correlations between both estimates are strong. Additionally, there is also a strong correlation between the total rural

Table 13: Pearson correlation between the WPdx estimated population served within 1km against other estimates per country

Country	JMP basic not on premises (est. rural population)	Rural population (WPdx)	Number of regions
Eswatini	0.98	0.98	4
Liberia	0.99	1.00	5
Nigeria	0.85	0.90	6
Sierra Leone	0.93	1.00	4
Uganda	0.95	0.86	15
Combined countries	0.92	0.94	34
Combined countries without Nigeria	0.96	0.94	28

population and the WPdx access population. This is not unexpected since as the population of the region grows there would be more people with access (see Table 13).

The correlations beg the question of whether the rural population variable may be just as predictive of JMP basic not on premises as the WPdx basic access. However, when one adds rural population in the same linear regression as WPDx basic access, the rural population coefficient shows little explanatory power ($p = 0.67$). For this reason, the rural population can be safely ignored and is dropped from the regressions presented in the following section.

6.3 Combining countries

Due to few observations per country, it was necessary to test the relationship between WPdx and JMP basic (not on premises) using a linear model with all countries together. Single country regressions could not provide any strong conclusions due to the low number of regions (i.e. observations) except in the case of Uganda whose results are similar to the combined model. A combined analysis provides the opportunity to see if there are overall trends in access that may be derived when there are more observations and which could apply to more than one country. However, Nigeria is excluded from the combined regression as it shows completely different trends than the other countries (see Figure 4).

As this combined group of countries can mask some differences between countries, a mixed model regression was tested to account for variation in the slope and intercept across countries. However, this model yields the same relationship found in a simple linear regression. For this reason, the simpler model has been retained. The mixed model coefficients can be found in the appendix in Table 25.

Table 14: Variable names in tables of coefficients for the models presented

Model variable name	Description	JMP style notation
r_w_bas_estimate_pop	Rural population estimated to have JMP basic drinking water services not on premises	$W_{7,!premises,population}$
wpdx_r_pop_with_basic_access	Rural population with WPDx basic access within 1km	$WPdx_{population}$
w_bas_pop	Population with JMP basic drinking water services	$W_{7,population}$
w_premises_pop	Population with improved drinking water on premises	$W_{premises,population}$
rural_pop	Rural population	
urban_pop	Urban population	

Table 15: Regression coefficients and statistics for the model predicting JMP basic excluding on premises from WPDx basic

Variable	Lower bound	Coefficient estimate	Upper bound	P-value
(Intercept)	-18,119.6060	63,978.1215	146,075.8491	0.1213
wpdx_r_pop_with_basic_access	0.5184	0.5873	0.6562	0.0000

6.3.1 Modeling regional JMP basic not on premises using WPdx basic access

The new combined regression is shown in Figure 5.

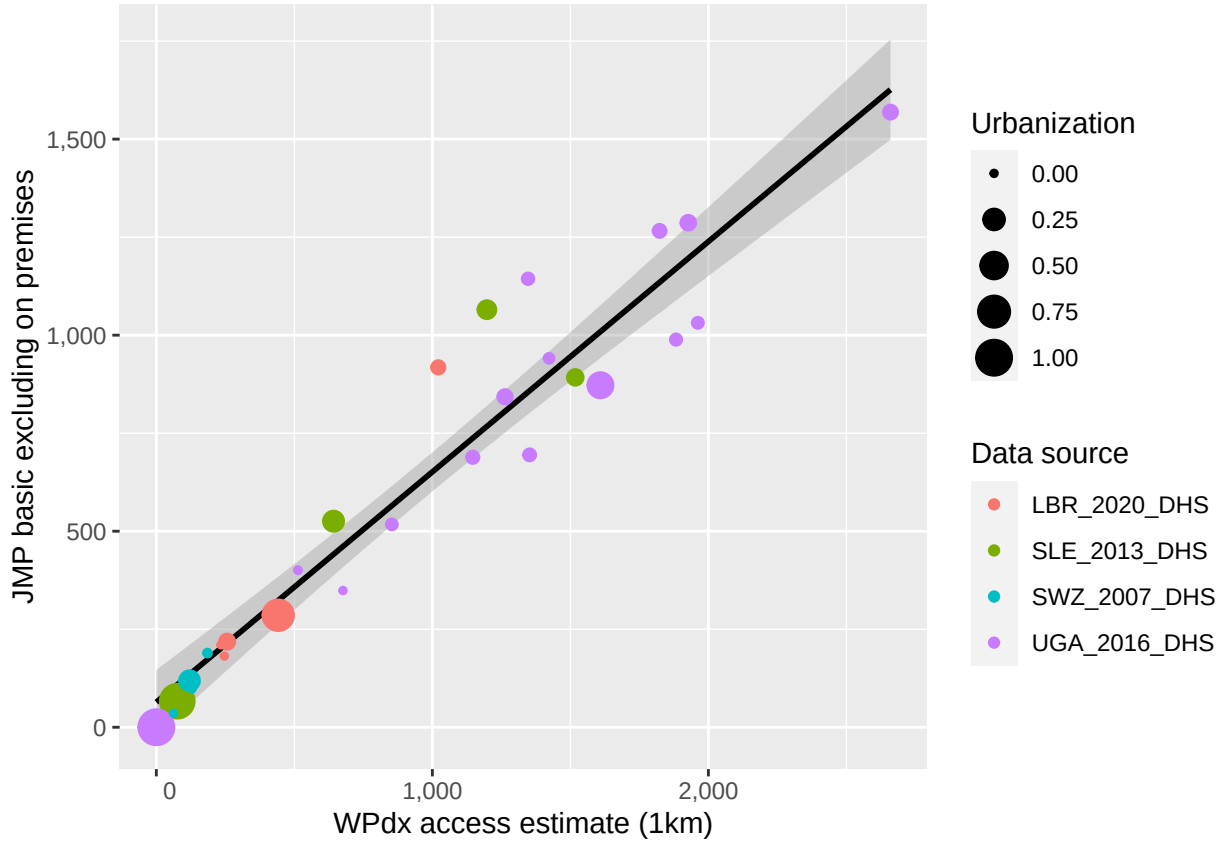

Figure 5: Population (per 1000 persons) with access without Nigeria. The 95% confidence bands show how on average JMP basic off premises is expected to change with WPdx access estimates.

The combined country regression shows strong trends with a very high adjusted R^2 of 0.9189439 ($p = 6.4 \times 10^{-16}$). The average of JMP basic not on premises is expected to be between 51.8% and 65.6% within a 95% confidence interval (mean of 58.7%).

We can do the same regression with half of the data set, which includes the 14 most rural regions. The plot is shown in Figure 6. There is no region that is more than 3% urban included in this group and it excludes Sierra Leone.

Even with half the number of regions, the regression of the most rural areas continues to show a strong trends with a very high adjusted R^2 of 0.8970438 ($p = 1.7 \times 10^{-7}$). The average of JMP basic not on premises is expected to be between 43.6% and 65.9% within a 95% confidence interval (mean 54.7). While the WPdx

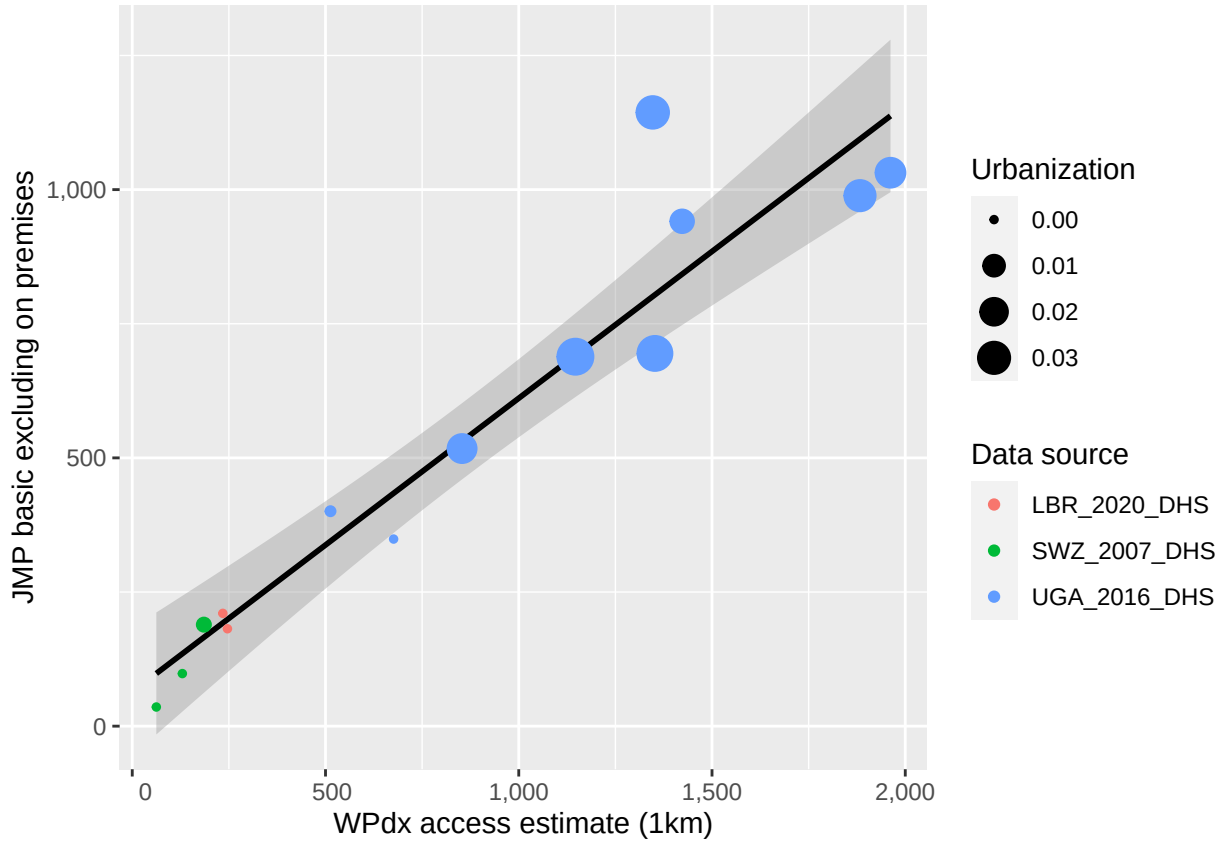

Figure 6: Plotting the 50% most rural regions; none are more than 3% urban. The 95% confidence bands show how on average JMP basic off premises is expected to change with WPdx access estimates.

Table 16: Regression coefficients and statistics for the model predicting JMP basic excluding on premises from WPDx basic with the 50 percent most rural areas

Variable	Lower bound	Coefficient estimate	Upper bound	P-value
(Intercept)	-55,126.8911	64,064.0973	183,255.0857	0.2643
wpdx_r_pop_with_basic_access	0.4356	0.5471	0.6586	0.0000

Table 17: Regression coefficients and statistics for model with separated parameters (all regions)

Variable	Lower bound	Coefficient estimate	Upper bound	P-value
(Intercept)	-169,138.7657	-44,956.6528	79,225.4601	0.4607
w_bas_pop	0.4578	0.8292	1.2006	0.0001
rural_pop	0.1685	0.3456	0.5227	0.0005
urban_pop	-0.7589	-0.3803	-0.0017	0.0491
w_premises_pop	-1.8535	-1.1181	-0.3827	0.0046
urban_pop:w_premises_pop	0.0000	0.0000	0.0000	0.2326

access estimate in these mostly rural regions over predicts access a bit more than when including more urban areas, the positive linear relationship still holds with similar coefficients. This confirms that there may be very minor under prediction of access in more urban areas because of using the on premise proportion of the whole region including urban areas to calculate the rural population with JMP basic off premises services ($W_{7,!premises}$).

6.3.2 Modeling WPdx basic access from DHS survey data and regional population

It is possible to turn the relationship around and estimate WPdx basic access within 1km from the JMP parameters from DHS surveys used to derive our estimate of $W_{7,!premises}$ and the regional rural and urban population figures. This model can validate the direction of each parameter in the linear regression.

This separated parameter model has a strong model fit with an adjusted R^2 of 0.9489964 ($p = 1.9422418 \times 10^{-14}$). Table 17 shows that in this separated model the JMP basic population served and overall rural population are positively correlated with WPdx access and that the population on premises is negatively correlated with WPdx access. These findings do not contradict the previous model or the assumptions we used to develop our estimate of rural access to JMP basic off premises.

If we decide to only focus on the three quarters of regions that are most rural, we would expect a the urban population parameter to be less important and possibly a stronger relationship with the other parameters albeit with less observations. It is necessary to take more than half of the regions as the model has more independent variables. However, these regions remain very rural and represents 21 regions from 0% to 10% urban (4.5% total). Indeed, as shown in Table 18, the urban population becomes less important in this regression while the on premises parameter becomes more important reinforcing the hypothesis that these estimates are related, at least in some countries.

7 Discussion

This study provides evidence that WPdx access coverage and JMP basic drinking water service estimates trend in the same direction in country regions when accounting for the population using an improved source on premises. This has been done in two ways. First, by deriving a comparable estimation of JMP basic off premises using data from the JMP inequalities files and testing it's relationship with WPdx figures in a linear regression. The second way was by estimating average WPdx figures on the basis of regional JMP

Table 18: Regression coefficients and statistics for model with separated parameters for the 75 percent most rural regions

Variable	Lower bound	Coefficient estimate	Upper bound	P-value
(Intercept)	-162,845.3504	-1,660.7342	159,523.8819	0.9828
w_bas_pop	0.4431	0.8542	1.2654	0.0005
rural_pop	0.1679	0.3703	0.5726	0.0014
urban_pop	-2.7389	-0.8332	1.0725	0.3661
w_premises_pop	-3.2168	-1.9772	-0.7376	0.0040
urban_pop:w_premises_pop	0.0000	0.0000	0.0000	0.1710

basic and JMP on premises figures from DHS surveys as well as the rural and urban populations of an area as calculated by WPdx.

Overall, WPdx access figures are higher than JMP figures. This can be due to differences in the methods and definitions of each and it can also be due to the behaviors of households and the water sources they choose to use as their primary source when a choice is present.

In terms of definitions and methodologies, it is possible that the radius used by WPdx and the lack of an estimate for queuing (based on overcrowding or time restrictions for example) may also be a cause for over-estimating access when compared to the household survey sources. It could also be that the maximum population served of each technology should be adjusted downward. The appendix B.1.1 shows how the number of people per water point varies between the household based and WPdx based estimates. These would require addition research to see how changing the radius and maximum populations and potentially using population density to estimate crowding would effect the comparison with JMP data. Additionally, the population maps used can also be a source of differences in estimates if for some reason, population densities in areas around water points are estimated to be higher than the number of people on the ground. It should be noted that each time definitions or methodologies are changed in the WPdx access estimate, it would be good to update this analysis to update the relationship between the household-based estimates and the WPdx estimates of access.

Behavior is also a potential source of differences. Public water points are part of the infrastructure of multiple-use services where people may use the water obtained from public water points for their animals or industry in addition to household uses. Hypothetically, there may be a desire to have access to more water points than are used as a primary source of drinking water for households that could lead, in some cases, to overestimation with population maps. Some water points may be provided by institutions that are not close to homes and people may choose to use other primary sources that are potentially unimproved or even surface water. Further research is needed to better quantify these other uses and choices based on surveys that contain information on other uses and alternative sources.

It is recommended to also explore the reason for differences in total national population between the UN sources and the Meta for Good 30m x 30m population grid. These may be due to mistakes in the population grids and it may be required to contact the producers of these datasets to validate population sources. Alignment between national, JMP and WPdx population sources would support the use of these different sources of information especially with regard to the total number of people with a level of service.

The overall higher rural population figures may be explained in part by higher estimates of urbanization in the UN prospectus and the way that WPdx estimates and removes urban populations (see methodology). It may be possible to use a buffer to extend estimated urban areas based on distance and/or population density around the lower resolution urban grids used (1km x 1km). In order to compare WPdx access with household surveys that disaggregate between rural and urban areas, it will be important to further align urban and rural areas more closely to the definitions and delimitations used in countries and used by the UN urbanization prospectus.

Nonetheless the fact that a linear relationship could be found provides experts and decision makers with potential new tools to be able to combine inventory based and household based surveys to quantify and

better understand access:

- For local governments and NGOs to have better estimates of populations served per water point (see B.1.1) at district level.
- For governments and potential service providers evaluating service delivery areas and service models to estimate the total number of public water points in an area on the basis of household surveys. This can be useful in cases where only piped systems and household connections have good data and these planned to be extended. It can also be useful when planning a data collection exercise for a national water point inventory to estimate the time and effort required to conduct the exercise.
- For policy makers to understand access to services by vulnerable populations at a lower spatial scale than possible with only the DHS surveys used in the JMP inequality files. Some of the most vulnerable populations do not have access to services on premises. Household surveys together with a water point inventory could provide estimates of the spread of services at district level.
- For government and service providers to have an improved understanding of access on the basis of administrative data at district level. For areas without recent household surveys, WPdx basic access together with administrative information on the number of household piped connections and self-supply may be used to estimate JMP basic drinking water services at municipal and district level. This study found the average of JMP basic not on premises in DHS regions was around 50% and 70% of the WPdx access within a 95% confidence interval. Adminstrative data on piped systems and household connections could be added to provide an estimate for JMP basic drinking water, including access on premises.

Because of the limited number of DHS regions per country, there was not enough data to validate how these trends may vary between countries, which would be useful to further contextualize and adjust parameters such as the maximum number of people served and/or the appropriate radius around water points. More detailed household surveys could also help to better understand the number of people served by each technology in a country. Dissaggregation between rural and urban survey results at regional level would enable a more accurate trend line to be drawn. Future research could use DHS microdata and census or other national survey microdata to further disaggregate household data used in these estimates and draw stronger conclusions.

Interestingly, if the relation found in eSwatini, Liberia, Sierra Leone, and Uganda together holds in each country and in Nigeria, then it may suggest that there may be a significant number of water points that are not included in the WPdx database of Nigeria.

Overall, these findings reinforce the view that combining data from different data sources provide a stronger picture of services, infrastructure and vulnerability over time.

For WPdx, there are a number of potential avenues to explore to further validate the use of WPdx basic access estimates in districts and regions and improve WPdx:

- Explore adding urban WPdx access estimates to further align and compare with household surveys.
- Adding a layer to record available information on premise services, such as source on self-supply and the number of household or compound piped system connections. These DHS surveys could also be added as a layer but do not provide a high resolution for use a district level.
- Aligning urban boundaries and calculated urbanization more closely to other data sources. An alternative would be to create a new urban grid based on the population grid. One method identified by Katy Sill of WPdx is described in (Dijkstra et al. 2020).
- Explore with Meta whether the population grids can be more closely aligned with national and UN sources or why not. Some initial exploration has been conducted in Sierra Leone.

- Future studies may choose to further examine the impact of urbanization and how a 10% urban vs. a 90% urban area may differ but would require a different type of regression.
- Try to estimate from the JMP estimates the number of water points in WPdx uncharted areas to estimate the number of water points that still could be added and guide attention to those areas.
- Focus studies on specific districts with household surveys and a good water point inventory reported in the WPdx database such as in Kabarole, Uganda, to further validate and cross check whether the trends in this report can be replicated at lower scales.
- Explore how changing the WPdx access parameters such as the radius used for distance, the maximum capacity per technology and other parameters could be used to improve the alignment of WPdx estimates and those from household survey sources.
- Explore the addition of a field or a label in the WPdx data standard to better identify self-supply and private water points that may be restricted in access. Alternatively, this could be a field to provide information about the types of users allowed to use the drinking water supply whether it is the general public, a household and/or neighbors.

This study has been able to develop a number of new analyses cutting across several African countries on the basis of national inventories made public through WPdx and the data provided from NGOs and other data providers in those countries ("WPdx Decision Support Tools," n.d.b). It also builds on the publicly available and nationally validated JMP datasets that have been compiled together in an the jmpwashdata package (Dickinson 2021).

Without open access to these data sets and publication of data by each data provider, it would not have been possible to make these results available. Open access can increase the validation, trust and quality in access estimates and identify issues that may need to be resolved. During this study, the tool producing the WPdx access estimates and the data sources were provided with feedback during data cleaning and analysis and as a result the estimates were updated to address different issues, such as missing water points or incorrect calculations. Countries and districts that do share their inventories may benefit from more research and analysis by third parties to validate its use. Public access can potentially enable community based organisations to check if the estimates and water point data match services on the ground and improve the usefulness of the data and tools.

It is recommended to WPdx to continue advocacy to add more countries across different continents and at different income levels to extend the coverage of these estimates. Together with household surveys, WPdx basic access estimates may be used to better estimate the distribution and location of potentially vulnerable populations and target resources more equitably.

It is almost always possible to use the WPdx data standard to format water point data that organizations are collecting as part of their normal operations, but it is not always feasible to publish this water point data. There is also a need to publish the methods used to create estimates so that they can be replicated in a validated manner in areas where, for example, water point data cannot be shared, whether due to political, security or other conditions. It could also be possible to offer ways to share data with WPdx for the purpose of the WPdx access estimates but without making those specific water points publicly available. This could be made possible either within the WPdx platform or by implementing the validated methods have been made public in other information systems. Ultimately, WPdx will be an important resource for anyone developing these tools.

It is important that household data, both the sources used by JMP and other household survey results at lower geographical levels share their findings on the key parameters used to calculate basic and safely managed service levels. This may be an area for improved sector coordination and sharing with a gap that could be filled with an initiative such as WPdx.

This study has demonstrated that even dated water point data can be useful in estimating access. As more parties contribute to complete inventories in each country, there will more possibilities to make better estimates. While service levels are often conceptualized from the perspective of the household, the cost, effort

and complexity of household surveys and data management make it difficult to collect this data more than every few years and with only a few country regions. Estimations at the level of the facility or water point, such as the water point reliability or downtime, can be useful to keep track of services and identify problems within days instead of months or years (Dickinson et al. 2017) especially when it is updated as part of regular operations. It is important that sector leaders from national and local government, international partners, service providers, and civil society require the collection and reporting of water point data, especially within national frameworks and insist that, when possible, this is made publicly available.

8 References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." *Journal of Statistical Software* 67 (1): 148. https://doi.org/10.18637/jss.v067.i01.

Cribari-Neto, Francisco, and Achim Zeileis. 2010. "Beta Regression in *R*." *Journal of Statistical Software* 34 (2). https://doi.org/10.18637/jss.v034.i02.

Dickinson, Nicolas. 2021. *Jmpwashdata: WHO/UNICEF Joint Monitoring Programme Water and Sanitation Data*. WASHNote. https://github.com/WASHNote/jmpwashdata.

Dickinson, Nicolas, Felix Knipschild, Peter Magara, and Gerald Kwizera. 2017. "Harnessing Water Point Data to Improve Drinking Water Services." Rotterdam, The Netherlands. https://washnote.com/2017/08/14/white-paper-available-harnessing-water-point-data-for-improved-water-services/.

Dijkstra, Lewis, Ellen Hamilton, Somik Lall, and Sameh Wahba. 2020. "How Do We Define Cities, Towns, and Rural Areas?" https://blogs.worldbank.org/sustainablecities/how-do-we-define-cities-towns-and-rural-areas.

JMP. 2018. "JMP Methodology: 2017 Update and SDG Baselines," April. https://washdata.org/sites/default/files/JMP%20methodology-Apr-2018-5.pdf.

———. 2021. *Progress on Household Drinking Water, Sanitation and Hygiene 2000-2020: Five Years into the SDGs.* Geneva: World Health Organization (WHO); the United Nations Children's Fund (UNICEF).

Jmpwashdata. 2022. WASHNote. https://github.com/WASHNote/jmpwashdata.

Thulin, Måns. 2021. *Boot.pval: Bootstrap p-Values*.

U. S. 2017. "U.S. Government Global Water Strategy 2017," 70.

WPdx. 2021. "WPDx Data Standard," January. https://www.waterpointdata.org/wp-content/uploads/2021/04/WPDx_Data_Standard.pdf.

"WPdx Decision Support Tools." n.d.a. http://tools.waterpointdata.org/?all_waterpoints=true&any_waterpoints=true&show_population_density=false&show_landcover=false&show_adm_borders=false&show_point_counts=false&adman_view=%22unserved%22&show_adman_pies=true&show_adman_labels=true&mode=%22basic%22&adman_level=%22best%22&bounds=%5B%5B-168.75000000000088%2C-65.18303007291351%5D%2C%5B168.74999999999807%2C65.18303007291394%5D%5D.

———. n.d.b. http://tools.waterpointdata.org/?all_waterpoints=true&any_waterpoints=true&show_population_density=false&show_landcover=false&show_adm_borders=false&show_point_counts=false&adman_view=%22unserved%22&show_adman_pies=true&show_adman_labels=true&mode=%22basic%22&adman_level=%22best%22&bounds=%5B%5B-168.75000000000088%2C-65.18303007291351%5D%2C%5B168.74999999999807%2C65.18303007291394%5D%5D.

"WPdx – The Water Point Data Exchange Is the Global Platform for Sharing Water Point Data." n.d. https://www.waterpointdata.org/.

Table 19: Summary of population differences

Country	Facebook population grid	JMP/WB Population	WPdx urbanization	JMP/WB urbanization	Percent difference in population	Percent difference in urbanization
Liberia	4,500,762	5,057,677	34	52	12	-18
Nigeria	205,979,343	206,139,587	35	52	0	-17
Sierra Leone	7,972,347	7,976,985	30	43	0	-13
Eswatini	1,281,479	1,160,164	7	24	-10	-17
Uganda	37,000,694	45,741,000	11	25	24	-14

Table 20: Summary of rural population differences

Country	WPdx rural population	JMP/WB Rural Population	Percent difference
Liberia	2,961,975	2,423,184	-18
Nigeria	133,841,163	99,033,580	-26
Sierra Leone	5,594,091	4,553,024	-19
Eswatini	1,194,756	879,741	-26
Uganda	33,000,695	34,326,791	4

A Population figures and sampling

A.1 National figures

A.2 Sampling and population per region

The most recent DHS surveys added to the JMP files contain figures for the sampled population for water access indicators. The number of households sampled and the number of regions are taken from the DHS database and shape files. Together the average household size can be derived to validate these data sources (see Table 21). With the exception of Liberia, household samples per region do not vary proportionally with regional population in WPdx. It is likely that other methodological factors and local considerations have influenced the sampling and weighting across regions. It can also be that the national population figures used by the National Statistics Office are different in each region in comparison to those in WPdx but this out of the current study's scope.

B National estimates

It is possible to sum up the estimated population for each of the JMP/DHS and WPdx parameters per country. While the purpose of this study was not to compare national level figures but rather to compare sub-national trends, Table 23 is provided for reference.

Table 21: Survey sample sizes

Country	Year	Survey	DHS households	JMP population	Regions
Liberia	2020	DHS	9,068	42,093	5
Nigeria	2018	DHS	40,427	189,656	6
Sierra Leone	2019	DHS	13,399	71,408	5
Uganda	2019	MIS	8,351	44,602	15

Table 22: Regional population summary

Country	JMP ID	DHS region name	Percent sampled	Meta population grid	JMP population sampled	Total WPdx regions	WPdx regions without population
Liberia	LBR_2020_DHS	South Eastern B	0.85	285,122	2,432	29	0
Liberia	LBR_2020_DHS	South Eastern A	0.79	332,064	2,624	56	1
Liberia	LBR_2020_DHS	North Western	0.96	367,292	3,522	32	0
Liberia	LBR_2020_DHS	North Central	0.95	1,508,734	14,277	116	0
Liberia	LBR_2020_DHS	South Central	0.93	2,007,550	18,711	72	0
Nigeria	NGA_2018_DHS	South East	0.08	24,552,208	20,653	103	0
Nigeria	NGA_2018_DHS	South South	0.08	26,272,152	20,725	104	0
Nigeria	NGA_2018_DHS	North East	0.11	29,292,650	32,694	119	1
Nigeria	NGA_2018_DHS	South West	0.08	37,384,247	30,724	121	0
Nigeria	NGA_2018_DHS	North West	0.14	42,782,296	58,911	154	0
Nigeria	NGA_2018_DHS	North Central	0.06	45,695,790	25,772	174	0
Sierra Leone	SLE_2019_DHS	Northern	1.89	775,836	14,675	21	0
Sierra Leone	SLE_2019_DHS	Southern	1.01	1,396,185	14,082	38	0
Sierra Leone	SLE_2019_DHS	Western	0.94	1,557,117	14,665	3	0
Sierra Leone	SLE_2019_DHS	North Western	0.63	1,978,273	12,546	35	0
Sierra Leone	SLE_2019_DHS	Eastern	0.67	2,264,936	15,244	56	0
Uganda	UGA_2019_MIS	Kampala	1.80	76,929	1,387	1	0
Uganda	UGA_2019_MIS	Karamoja	0.10	976,577	942	60	0
Uganda	UGA_2019_MIS	Teso	0.17	1,250,087	2,173	57	0
Uganda	UGA_2019_MIS	Kigezi	0.13	1,305,608	1,660	70	0
Uganda	UGA_2019_MIS	Lango	0.12	1,835,450	2,197	74	0
Uganda	UGA_2019_MIS	Bugisu	0.11	2,247,254	2,519	173	1
Uganda	UGA_2019_MIS	Tooro	0.14	2,280,475	3,218	118	0
Uganda	UGA_2019_MIS	Bukedi	0.08	2,382,364	1,997	107	0
Uganda	UGA_2019_MIS	Acholi	0.07	2,485,739	1,734	114	0
Uganda	UGA_2019_MIS	South Buganda	0.24	2,614,920	6,273	105	0
Uganda	UGA_2019_MIS	West Nile	0.17	2,629,913	4,442	96	0
Uganda	UGA_2019_MIS	Bunyoro	0.06	2,671,567	1,726	105	0
Uganda	UGA_2019_MIS	Ankole	0.09	3,959,536	3,597	183	0
Uganda	UGA_2019_MIS	Busoga	0.09	3,961,016	3,613	123	0
Uganda	UGA_2019_MIS	North Buganda	0.11	6,323,260	6,845	134	0

Table 23: National estimates derived from the sum of regional estimates from JMP/DHS and WPdx

Country	JMP basic (national)	On premises (national)	Basic off premises (rural estimate)	WPdx basic access (rural)	WPdx waterpoint functionality
Eswatini	63.5	26.5	36.9	41.6	76.9
Liberia	75.1	17.7	61.2	74.3	80.3
Nigeria	70.7	26.5	43.5	29.1	57.2
Sierra Leone	53.7	7.0	45.6	61.4	67.7
Uganda	51.6	14.3	38.2	61.9	80.1

Table 24: Sub-national variation in population with access per water point

Country	Mean basic off premises per WP	Std. dev. basic off premises per WP	Mean WPdx basic access per WP	Std. dev. WPdx basic access per WP
Eswatini	51	9	67	31
Liberia	112	49	135	50
Nigeria	1649	1128	867	330
Sierra Leone	68	36	90	50
Uganda	185	59	299	95

B.1 Access

B.1.1 People served per water point

The WPdx basic access is calculated on the basis of the water points in the WPdx inventory. Overall, it is higher than the JMP basic off premises in each country and tends to estimate a higher number of people served per water point as a result.

C Mixed model

A mixed model of the linear regression using the lme4 R package (Bates et al. 2015) was attempted to see if allowing for random effects in the slope and intercept between each country in the regression would yield different results. The mixed model coefficients including 95% confidence intervals are included here using the bootstrapping method provided by the boot.pval R package (Thulin 2021). The result is provided in Table 25. The findings confirms there is little difference with the simple linear regression with all countries in the same group.

Table 25: Coefficients of the mixed model without Nigeria with 95 percent confidence intervals

	Coefficient	Lower bound	Upper bound	P-value
(Intercept)	63,406.636	-21,089.511	149,769.001	0.147
wpdx_r_pop_with_basic_access	0.591	0.518	0.665	0.000