OLAS

WATER AND SANITATION OBSERVATORY FOR LATIN AMERICA AND THE CARIBBEAN

Water and Sanitation Household Survey Dataset

for Latin America and the Caribbean

Methodology Documentation — 2026 Update

Description of OLAS Household Survey Dataset Methodology

Inter-American Development Bank · Water and Sanitation Division

Contact: olasdata@iadb.org

Table of Contents

1. Dataset

1.1 Purpose

The Household Survey Dataset is generated from data from National Household Surveys throughout Latin America and the Caribbean (LAC) with the objective of providing detailed information on water and sanitation access for the region. The dataset is produced by the Water and Sanitation Observatory for Latin America and the Caribbean (OLAS) via a collaboration between the Inter-American Development Bank's Social Division. The OLAS's primary purpose is to serve as a publicly accessible platform to:

- Provide sufficient, consistent, and updated data and information to contribute to the integration of the SDGs in the sector and to understand the real gaps in the water and sanitation sector at the national and regional levels.
- Standardize the methodology implemented by each country when compiling information and calculating sector indicators.
- Promote the coordination between statistical and information agencies, sectoral entities, and national and local government actors, so that the information corresponds to the reality of the country.
- Encourage research in the water and sanitation sector at both the national and regional levels through the creation of the OLAS Research and Development Network (RID), so that information is available to better guide public policy in the countries.

The dataset serves to address and expose data gaps, providing detailed information about water and sanitation access for countries in LAC that can be disaggregated by several socioeconomic and demographic dimensions. This document describes the OLAS dataset, the methodology used to generate it and provides guidance on data use.

Addressing Data Gaps

Various agencies of the United Nations are responsible for integrating information for monitoring the Sustainable Development Goals (SDGs) at the global level, with information submitted by countries. For SDG 6, related to access to clean water and sanitation, the UN created the Joint Monitoring Program (JMP) involving UNICEF, WHO, UNEP, OECD, UNESCO, FAO, and UNECE. Despite this support, there are major information gaps throughout LAC. Five major challenges face the collection of SDG-relevant information:

1. There is a significant gap in the information available, which varies from country to country.
2. There is no regional agency to perform the integration and validation of information, which would allow homogenization and harmonization for subsequent analysis, nor are there incentives for the implementation of a common methodology.
3. Sector information is often outdated, uses inconsistent metrics, and is generated under different methodologies.

4. At the national level, there is a disarticulation between statistical agencies, information systems, and sector entities, and a disconnect with local government actors and municipalities.
5. The data needs of the region are different from those at the global scale, due to its own specific water and sanitation challenges.

1.2 Description of the OLAS Household Survey Dataset

The OLAS Household Survey Dataset contains water and sanitation related indicators generated from microdata from national household surveys throughout the region. The dataset contains information from 2003 to the present for countries throughout Latin America and the Caribbean. Indicators are provided as both a percentage of households and a total household count for each category, and can be broken down by socioeconomic dimensions, including area (urban or rural), income quintile, sex, and ethnicity of the head of household.

Using microdata from National Household Surveys is consistent with the OLAS priority of operationalizing country-generated data. The key benefits include:

- High country coverage: most LAC countries have these surveys and most include modules on household living conditions, including questions on the source of water and type of sanitation.
- Representativity and sample size: surveys are usually representative at the national, urban, and rural levels, and generally have much larger sample sizes than other regional surveys.
- Country-generated: the dataset is generated from microdata published by the countries using a transparent process, making it a regional effort subject to review by the countries

Using country-level data also presents specific challenges. While the methodology for categorizing data in household surveys is standardized, the diversity of questions and response options across different surveys creates harmonization challenges and can impair cross-country comparison. The methodology for generating indicators minimizes this risk, but in some cases uncontextualized comparison is not recommended. These cases are discussed transparently in the Data Use section.

Dataset Products

The OLAS Household Survey Dataset is delivered one dataset but contains three related products:

- Base Dataset: country-year survey microdata harmonized and processed into indicators, with observed values only.
- Interpolated Dataset: the base dataset with missing years filled using linear interpolation between observed survey years, enabling consistent time series analysis. No extrapolation beyond observed data ranges is performed.
- Regional Dataset: population-weighted regional estimates for LAC overall and four IDB subregions, derived from the interpolated dataset, with population coverage metadata for each estimate.

This document covers the methodology for all three products, as well as recommendations on how to use the fields in the final dataset.

Periodicity

Household surveys are carried out on different schedules throughout LAC, so this dataset includes data from different years for different countries. The dataset contains data from 2003 to the present. A list of countries and their corresponding survey years is available in the Data Sources section. Countries generally carry out surveys continuously or every other year. Due to the near constant availability of new source datasets, this dataset is updated annually.

Indicators

The dataset contains water and sanitation-related indicators, with each provided as both a percentage of households (field: value) and a total number of households (field: level). These indicators are designed to provide granular information on water and sanitation access in the region and are organized by topic. A complete list of indicator definitions is available in Table 1. Many indicators are designed to be examined in compositional groups — groups that should sum to 100% for any given country, year, and disaggregation slice (Table 2).

Indicator	Description
aguafred_ch	Percent of households with a piped water connection inside the dwelling
aguafpublico_ch	Percent of households that use a public tap or standpipe
aguafembotellada_ch	Percent of households that use bottled water as their primary source
aguafpozoprot_ch	Percent of households that use protected wells
aguaflluvia_ch	Percent of households that use rainwater harvesting
aguafcamion_ch	Percent of households that receive trucked water
aguafotramej_ch	Percent of households using other improved sources (e.g. protected springs)
aguafsuperficial_ch	Percent of households sourcing water directly from surface water bodies
aguafotranm_ch	Percent of households using other non-improved sources
aguafdesconocido_ch	Percent of households with an unclassifiable water source
aguafnomejorada_ch	Percent of households using an unimproved water source
aguafmejorada_ch	Percent of households using an improved water source
aguafredcon_ch	Percent of households that use the piped water distribution network as their primary source for water used for human consumption (not including public standpipes)
aguafpublicocon_ch	Percent of households that use public standpipes as their primary water source used for human consumption
aguafembotcon_ch	Percent of households that use bottled water as their primary source for water used for human consumption
aguafpozoprotcon_ch	Percent of households that use protected wells as their primary source for water used for human consumption
aguaflluviacon_ch	Percent of households that use rainwater harvesting for human consumption
aguafcamioncon_ch	Percent of households that receive trucked water for human consumption
aguafotramejcon_ch	Percent of households that source water from other improved sources (protected springs) for human consumption
aguafsupercon_ch	Percent of households that source water for human consumption directly from surface water bodies.
aguafotranmcon_ch	Percent of households that source water for human consumption directly from other non-improved sources

aguafdesconcon_ch	Percent of households using an unclassifiable water source as their primary source for water used for human consumption
aguafmejoradacon_ch	Percent of households that use an improved water source as their primary source for water used for human consumption
aguafnmcon_ch	Percent of households using an unimproved water source for consumption
aguared_ch	Percent of households with any piped water network connection
aguadisp_ch	Percent of households with water available as needed (JMP SDG 6 definition)
aguadispcontinuo_ch	Percent of households reporting no water shutoffs
aguadispnulo_ch	Percent of households with no information on water continuity
aguamide_ch	Percent of households with a water meter
aguatrat_ch	Percent of households that treat water at the point of use
sanred_ch	Percent of households with a toilet connected to the sewer network
sanseptic_ch	Percent of households with a toilet connected to a septic system
sanotramejorado_ch	Percent of households with another improved sanitation facility
sanotranm_ch	Percent of households with another unimproved sanitation facility
sanambiente_ch	Percent of households whose sanitation drains to the surface
sandesconocido_ch	Percent of households with an unclassifiable sanitation facility
sanmejorado_ch	Percent of households with an improved sanitation facility
sannomejorado_ch	Percent of households with unimproved sanitation facilities
sinsan_ch	Percent of households without any sanitation facility
nosanvecino_ch	Percent of households relying on public or neighbors' facilities
nosanairelibre_ch	Percent of households practicing open defecation
nosandesc_ch	Percent of households without sanitation with unknown alternatives
sanexclusivo_ch	Percent of households with exclusive access to sanitation facilities
sanexclusivonulo_ch	Percent of households with no information on sanitation exclusivity

Table 1: Indicator data dictionary

Indicator Groups and Hierarchy

Several sets of indicators form compositional groups, meaning that within a given country, year, and disaggregation stratum, all members of the group should sum to approximately 100%. These groups are critical for validating data quality and for understanding the complete distribution of a household population across categories. The 3 compositional groups are:

- Group 1 — Primary Water Source: all water source indicators (aguafred_ch through aguafdesconocido_ch)
- Group 2 — Sanitation Facility Type: all sanitation type indicators (sanred_ch through sandesconocido_ch)
- Group 3 — Water sources for human consumption (differentiated in some surveys from water sources for other uses): aguafredcon_ch through aguafdesconcon_ch

The groups follow a two-level hierarchy. The three primary groups are detailed water source and sanitation installations breakdowns whose members sum directly to 1. Each has a corresponding summary group, which is calculated independently from the surveys (value_original), but recomputed from their primary group for the interpolation and regional aggregation process (See sections 2.2 and 2.3). Additionally, there is the subgroup for households without sanitation facilities (No sanitation in Table 2), which sum to the value of the indicator “sinsan_ch”.

Table 2 shows the mapping of detailed indicators to their high-level summaries by topic. The high-level indicators represent aggregations of the detailed indicators within the same subtopic. Compositional groups — groups that must sum to 100% — are indicated within the same subtopic row.

Topic	High-Level Indicator	Detailed Indicators
Water source for general use (each column in this group sums to 100)	aguafmejorada_ch	aguafred_ch, aguafpublico_ch, aguafembotellada_ch, aguafpozoprot_ch, aguaflluvia_ch, aguafcamion_ch, aguafotramej_ch
	aguafnomejorada_ch	aguafsuperficial_ch, aguafotranm_ch
	aguafdesconocido_ch	aguafdesconocido_ch
Water source for consumption (each column in this group sums to 100)	aguafmejoradacon_ch	aguafredcon_ch, aguafpublicocon_ch, aguafembotcon_ch, aguafpozoprotcon_ch, aguaflluviacon_ch, aguafcamioncon_ch, aguafotramejcon_ch
	aguafnmcon_ch	aguafsupercon_ch, aguafotranmcon_ch
	aguafdesconcon_ch	aguafdesconcon_ch
Sanitation facility type (each column in this group sums to 100)	sanmejorado_ch	sanred_ch, sanseptic_ch, sanotramejorado_ch
	sannomejorado_ch	sanotranm_ch, sanambiente_ch
	sandesconocido_ch	Unclassifiable sanitation facility
No sanitation	sinsan_ch (detailed indicators sum to sinsan_ch value)	nosanvecino_ch, nosanairelibre_ch, nosandesc_ch

Table 2: Indicator groups and hierarchy

Data Dictionary

The final dataset contains fields representing the 3 data products (original harmonized household survey values, interpolated values closing temporal gaps, and regional and subregional estimates). These products have been combined into one dataset for transparency purposes, as having all three makes interpretation of the datasets easier.

Fields in the final dataset are listed below, color coded by their origin product.

Field	Description
agg_region	Region identifier: ISO3 code for country rows; region name (e.g. 'LAC', 'Andean') for regional rows
year	Year of estimate
indicator	Indicator name (see Table 1)
area	Disaggregation: Total, Urban, Rural

Field	Description
quintile	Income quintile: Total, Q1–Q5
sex	Sex of head of household: Total, Male, Female
ethnicity	Ethnicity: Total, Indigenous, Afrodescendant, Other
value_final	Final indicator value (proportion, 0–1). Indicator value field across original data, interpolated data and regional data, representimg interpolated or smoothed data if the field underwent renormalizations and the original data point if not. For regional rows it is the population weighted mean of the indicator for the geographic area covered (subregion or region).
level_final	Final total households in category. Households in indicator slice across original data, interpolated data and regional data, representing interpolated or smoothed data if the field underwent renormalization and the original data point if not. For regional rows it is household sum across the geographic area covered (subregion or region).
value_original	Original observed value (country rows only; NA for regional rows)
level_original	Original observed household count (country rows only; NA for regional rows)
se_original	Original standard error (country rows only; NA for regional rows)
value_smooth	Audit column: value after renormalization (country rows only)
level_smooth	Audit column: level after renormalization (country rows only)
status_value	Data status: 'observed', 'interpolated', or 'no_data'
status_level	Data status for level field
n_countries	Number of countries contributing to estimate (1 for country rows, multiple for subregional and regional rows)
n_observed	Number of contributing countries with observed (not interpolated) values
n_interpolated	Number of contributing countries with interpolated values
countries_included	Comma-separated list of ISO3 codes of countries contributing to estimate
status_region	Regional status: 'all_observed', 'mixed', 'all_interpolated', or 'all_imputed'
hh_contributing	Total households represented by contributing countries for this indicator-year
hh_potential	Total households across all subregion members with any data in this year
n_countries_potential	Number of subregion members with any household data in this year
pct_population_covered	Percentage of subregion households represented by this estimate (hh_contributing / hh_potential × 100)
group_en	Indicator group name in English (e.g. “Water sources”, “Drinking water classification”). “Independent variable” for indicators not belonging to any compositional group.
group_es	Indicator group name in Spanish.
country_name_en	Full country name in English (country rows only; subregion/region name for aggregate rows).
country_name_es	Full country name in Spanish.
subregion_en	IDB subregion name in English (e.g. “Andean Region”, “Caribbean”). For regional rows, same as country_name_en.
subregion_es	IDB subregion name in Spanish.

Field	Description
region_en	Always “Latin America and the Caribbean”.
region_es	Always “América Latina y el Caribe”.
area_en	Area disaggregation label in English: “Total”, “Urban”, “Rural”.
area_es	Area disaggregation label in Spanish: “Total”, “Urbano”, “Rural”.
quintile_en	Income quintile label in English: “Total”, “Quintile 1” through “Quintile 5”.
quintile_es	Income quintile label in Spanish: “Total”, “Quintil 1” through “Quintil 5”.
sex_en	Sex label in English: “Total”, “Male”, “Female”.
sex_es	Sex label in Spanish: “Total”, “Hombre”, “Mujer”.
ethnicity_en	Ethnicity label in English: “Total”, “Indigenous”, “Afro-descendant”, “Other”.
ethnicity_es	Ethnicity label in Spanish: “Total”, “Indígena”, “Afrodescendiente”, “Otro”.
quality_score	Data quality score (1–5, higher = better). For observed country rows: based on coefficient of variation (SE / value). For interpolated country rows: based on years to nearest observed survey round. For regional rows: based on pct_population_covered. NA when SE is unavailable for observed rows.

Table 3: Data dictionary

Data Sources

The dataset is generated from microdata from National Household Surveys throughout Latin America and the Caribbean, harmonized by the Social Division of the Inter-American Development Bank. Table 3 lists the surveys included in the dataset, along with the national statistical agency responsible and the years for which data are available.

Note: *This table is updated annually. The survey year ranges shown reflect data incorporated as of the 2026 update.*

ISO	Country	Survey Name	Agency	Years Available	IDB Subregion
ARG	Argentina	Encuesta Permanente de Hogares Continua (EPH) and its continual version	INDEC	2003–2024	Southern Cone
BHS	Bahamas	Labour Force Survey (LFS)	DoS	2009–2022	Caribbean
BLZ	Belize	Labour Force Survey (LFS)	SIB	2005–2009	Caribbean
BOL	Bolivia	Encuesta de Hogares (EH), and its continual version	INE	2005–2023	Andean
BRA	Brazil	Pesquisa Nacional por Amostra de Domicílios (PNAD) and its continual version	IBGE	2003–2023	Southern Cone
BRB	Barbados	Labour Force Survey (LFS)	BiSS	2024	Caribbean
CHL	Chile	Encuesta de Caracterización Socioeconómica (CASEN)	MDS	2003–2022	Southern Cone
COL	Colombia	Gran Encuesta Integrada de Hogares (GEIH)	DANE	2003–2024	Andean
CRI	Costa Rica	Encuesta de Hogares de Propósitos Múltiples (EHPM) and Encuesta Nacional de Hogares (ENAHO)	INEC	2003–2024	CA & Mexico
DOM	Dominican Republic	Encuesta Nacional de Fuerza de Trabajo (ENFT) and its continual version	ONE	2005–2024	Caribbean
ECU	Ecuador	Encuesta Nacional de Empleo y Desempleo (ENEMDU)	INEC	2003–2024	Andean
GTM	Guatemala	Encuesta Nacional de Empleo e Ingresos (ENEI) and Encuesta Nacional de Condiciones de Vida (ENCOVI)	INE	2004–2024	CA & Mexico
GUY	Guyana	Labour Force Survey (LFS)	Bureau of Statistics	2019–2021	Caribbean

ISO	Country	Survey Name	Agency	Years Available	IDB Subregion
HND	Honduras	Encuesta Permanente de Hogares de Propósitos Múltiples (EPHPM)	INE	2003–2024	CA & Mexico
HTI	Haiti	Enquête sur les Conditions de Vie des Ménages (ECVMAS)	IHSI	2012–2014	Caribbean
JAM	Jamaica	Labour Force Survey (LFS)	STATIN	2012	Caribbean
MEX	Mexico	Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH)	INEGI	2004–2024	CA & Mexico
NIC	Nicaragua	Encuesta de Medición de Nivel de Vida (EMNV) and Encuesta Continúa de Hogares ECH	INIDE	2005–2016	CA & Mexico
PAN	Panama	Encuesta de Hogares (EH), Encuesta de Hogares de Propósitos Múltiples (EHPM), Encuesta de Mercado Laboral (EML)	INEC	2014–2024	CA & Mexico
PER	Peru	Encuesta Nacional de Hogares (ENAHO)	INEI	2003–2024	Andean
PRY	Paraguay	Encuesta Permanente de Hogares (EPH) and its continuous versión	DGEEC	2003–2024	Southern Cone
SLV	El Salvador	Encuesta de Hogares de Propósitos Múltiples (EHPM)	DIGESTYC	2003–2024	CA & Mexico
SUR	Suriname	Survey of Living Conditions (SLC)	ABS	2017–2024	Caribbean
TTO	Trinidad & Tobago	Continuous Sample Survey of Population (CSSP)	CSO	2004–2016	Caribbean
URY	Uruguay	Encuesta Continua de Hogares (ECH)	INE	2003–2024	Southern Cone
VEN	Venezuela	Encuesta Nacional de Condiciones de Vida (ENCOVI)	INE	2003–2014	Andean

Table 4: Data sources by country — update annually when new survey years are incorporated.

COUNTRY	SURVEY	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
ARG	ARG-EPHC	2457	2433	3635	3677	3677	3647	3639	3621	3463	3507	3667	3429	3407	3379	3363	3413	3349	3255	3267	3349	3331	3169
BHS	BHS-LFS	1345	1339	1381	1381	1415	1363	1381		1115	469	1265	1323	1027	983	1109	1157	903				107	
BLZ	BLZ-LFS	1159	6330	7879		5600											133						
BOL	BOL-ECH	7491		7879	6631	5089	6701	6806		9279	12492	11740	11807	11790	16991	16903	16408	16377	12228	16577	12743	16221	
BRA	BRA-PNAD	7464	7945	7509	7496	7603	7595	7514		7727	7557	7620	7668	7534									
BRA	BRA-PNADC														7657	8037	8144	8130	1543	1552	8196	8179	849
BRB	BRB-CLFS		361	349	55	116	91		52	55		49		46	349								
BRB	BRB-BSLC																						2096
CHL	CHL-CASEN	6409			6476			6571		8527		8486		8810		11828			8431		11145		
COL	COL-ECH	895	925	919																			
COL	COL-GEIH				9268	10491	6154	6237	6171	6045	5942	6074	6072	12914	12628	12755	13550	13664	13864	14025	18001	17900	17520
CRI	CRI-EHPM	3376	4131	4051	4196	4064	6528	6353															
CRI	CRI-ENAHO								8693	8324	8507	8187	8238	8160	8043	8222	8264	7888	7227	7327	7725	7567	7290
DOM	DOM-ENFT	3806	3826	3780	3851	3742	4064	4146	4000	4158	4099	4203	3986	4055	4067								
DOM	DOM-ENCFT															5487	5520	5662	5497	5561	5428	5277	4887
ECU	ECU-ENEMDU	7096	1453	6856	6819	6864	6835	6804	6788	6640	8030	8335	8739	8666	8666	9987	7573	7567	6927	7190	6851	6716	6347
GTM	GTM-ENEI	1961	5438						6701	6632	6966	7091	6380	7663	7488	7292	7242	7307		7510	10231		
GTM	GTM-ENCOVI				10891																		
GTM	GTM-ENEIC																						1059
GUY	GUY-LFS															1395	1384	1360		1372			
HND	HND-EPHPM	4161	4208	3960	4568	4473	4503	4522	3923	4066	3939	3819	5306	5127	5175	5009	5039	4001			3724	8075	7667
HTI	HTI-ECVMAS										5113												
JAM	JAM-LFS	52	793	784	799	784	307	307	766		757	751	748		745			115	112				
MEX	MEX-ENIGH		4017	4254	4165		3906		9319		7719		8895		9836		9907		9962		9968		12190
NIC	NIC-EMNV			5238				3160					6369										
NIC	NIC-ECH								865	856	850												
PAN	PAN-EH	904	904	970	967	979	976	982	958														
PAN	PAN-EHPM									976	976	964	5878	12095	10175	13013	11223	12906			14721		
PAN	PAN-EML																					424	67
PER	PER-ENAHO	4441	4606	4649	4589	4640	4635	4562	4261	4548	8698	8825	11355	11578	11598	12376	13449	13350	13282	13751	13702	15653	14604
PRY	PRY-EPH	5852	5948	4813	4841	4114	6946	6727	6718	6677	6833	4692	6881	5233	8310	7814							
PRY	PRY-EPHC																4601	6543	6602	6327	6587	4568	3671

SLV	SLV-EHPM	6066	12	5933	6665	6595	5404	5534	5460	5872	5894	5905	6131	6088	5246	5237	5228	5186	4784	5138	5088	5015	1693
SUR	SUR-SLC															1855					1757		
TTO	TTO-CSSP	1310	1279	1291	1291	1330	1321	930	1160	1124	1212	1183	1211	1186									
URY	URY-ECH	2213	2189	2229	8710	9682	9609	8924	11172	11140	11017	10844	10899	10911	10985	10894	10467	10556	388	6830	9173	10105	8654
VEN	VEN-ENCOVI														1086	1162	1105		2680	2875			971

Table 5: Temporal coverage of surveys used to generate the data set and their sample sizes.

2. Dataset generation

2.1 Microdata Harmonization and Indicator Generation

Harmonization Process

The dataset is based on microdata from national household surveys harmonized by the IDB's Social Division (SCL), with OLAS collaborating in the definition and harmonization of water and sanitation indicators. The harmonization process standardizes variables across surveys from different countries and years into a common schema, enabling cross-country comparison. The harmonization is documented in detail in the IDB Social Division harmonization documentation and is available in the associated GitHub repositories.

The harmonization process involves the following key steps:

- Variable identification: relevant water and sanitation variables are identified in each country's survey instrument.
- Value mapping: response options from each survey are mapped to the standardized OLAS indicator schema, accounting for differences in question wording and response categories across countries and survey years.
- Weight application: household expansion weights provided in each survey are applied to generate population-representative estimates.
- Dimension extraction: socioeconomic and demographic dimensions (area, income quintile, sex, ethnicity) are extracted and standardized across surveys.

Because survey instruments vary across countries and over time within a country, there are cases where direct mapping is not possible or where the quality of the mapping differs across surveys.

Harmonization Challenges and Water & Sanitation Variables

Harmonizing household surveys across LAC countries is complex due to methodological changes introduced by national statistics institutes over time, differences in thematic and geographic coverage across countries, variation in how databases are structured, and incomplete documentation. These challenges are addressed in detail in *D.1.1. Documentación armonización microdatos encuestas de hogares*, with all code available in the corresponding GitHub repository.

For water and sanitation specifically, the core challenge is that countries ask different questions with different response options, making cross-country comparison difficult. The harmonization strategy addresses this by creating variables broad enough to encompass all possible survey responses. For primary water source, for example, ten categories are defined — from piped network access through to unknown/unclassifiable sources — so that every household in every country can be assigned a category that reflects the true level of detail available in their survey. Countries with detailed response options can be compared directly; countries with cruder options are identifiable, with ambiguous responses absorbed into the unknown category rather than discarded or misclassified.

The same logic applies to less commonly asked topics, such as water sources specifically used for human consumption. Where a survey does not ask the question at all, a dedicated zero/"not asked" value is assigned, making data gaps explicit rather than invisible. This approach allows users to track not just what is known, but also the degree of uncertainty and the limits of survey coverage for any given country and year.

The process for harmonizing the microdata of the household surveys is available on GitHub: github.com/EL-BID/armonizacion_microdatos_encuestas_hogares_scl. See Annex A for the rules used to generate the harmonization variables.

Indicator Generation

Indicators are generated from the harmonized microdata by computing the weighted proportion of households that fall into each category for each combination of country, year, and disaggregation dimension. Each indicator is generated as a:

- value: the proportion of households in the category (between 0 and 1), equivalent to the percent divided by 100.
- level: the total number of households in the category, derived by multiplying the value by the survey expansion factor.

For each country-year-dimension combination, indicators are generated for all available disaggregation strata: area (Total, Urban, Rural), income quintile (Total, Q1–Q5), sex of head of household (Total, Male, Female), and ethnicity of head of household (Total, Indigenous, Afro-descendant, Other), among others. Not all dimensions are available in all countries or survey years.

The process for generating the indicators is available on GitHub: GitHub - BID-DATA/calculo_indicators_R at development · GitHub while the rules for generating the indicators are available in Annex B.

Notes on Indicators

The dataset is designed to encompass all possible survey response options across countries and years, regardless of differences in question structure. This simplifies harmonization but requires users to understand two key design features: contextualization via indicator groups, and representation of data gaps.

Indicator Groups

Indicators are organized into thematic groups that together sum to a measured whole. For example, the general water source group covers all possible primary water sources — when viewed together, these indicators give a complete picture of water source distribution for a given country and year. Viewing indicators within their group over time also helps track changes in survey design: for instance, Brazil's surveys prior to 2016 only distinguished between network water and unknown sources, while from 2016 onward they began categorizing protected wells, rainwater, and other improved sources separately.

Some groups do not sum to a whole but instead measure the same concept in complementary ways, such as the availability/continuity group, where multiple indicators together paint a more complete picture than any single one in isolation.

Data Gaps

The dataset makes data gaps explicit through two types of indicators. First, nulo indicators (e.g. aguafconnulo_ch, aguadispnulo_ch, sanexclusivonulo_ch) flag topics that were not addressed in a given survey — for instance, whether water source information for human consumption specifically was collected. Second, unknown indicators (e.g. aguafdesconocido_ch, sandesconocido_ch) capture responses that could not be classified due to ambiguous response options, such as wells recorded without specifying whether they are protected. Together these allow users to visualize uncertainty in water source types and facilitate contextualized comparison of the indicators in the dataset.

2.2 Data Interpolation

The interpolated dataset fills years between observed survey rounds using linear interpolation, enabling consistent annual time series for trend analysis and as inputs to the regional aggregation. This section describes the interpolation methodology in full.

Interpolation Methodology

The interpolated dataset is generated by applying piecewise linear interpolation between observed survey years independently for each combination of country, indicator, area, income quintile, sex, and ethnicity. The method is consistent with the approach used by the WHO/UNICEF Joint Monitoring Programme (JMP) for SDG 6.

Key methodological decisions:

- Linear interpolation only: values between two observed survey years are estimated using linear interpolation (straight-line). No higher-order smoothing or model-based estimation is applied, to preserve transparency and avoid assumptions about trends.
- No extrapolation: values are not projected beyond the first or last observed survey year for any country-indicator combination. Only the range between the earliest and latest observed year is filled. This is a deliberate departure from some JMP products that include forward projections to the current year.
- Independent by disaggregation: interpolation is performed separately for each combination of area, quintile, sex, and ethnicity. The Total slice is not derived from the disaggregated slices.
- Both value and level interpolated: the household proportion (value) and total household count (level) are interpolated independently. This accounts for the fact that the household population of a stratum may also change over time.
- Standard errors not interpolated: standard errors from the original survey are retained at observed years only. Interpolated years do not carry standard error estimates, as these cannot be meaningfully derived from linear interpolation.

Compositional Renormalization

After interpolation, a renormalization step is applied to compositional groups to ensure they sum to 1. Renormalization is necessary because independent linear interpolation of each group member does not guarantee that the group sum is preserved at exactly 1 in interpolated years.Renormalization is applied only to interpolated years — observed years are left untouched, as they already sum to 1 by construction from the survey microdata.

Renormalization proceeds as follows:

1. members of a compositional group that have non-missing values for a given country-year-stratum are identified.
2. The group sum is computed as the sum of all non-missing members.
3. Each interpolated member is divided by the group sum. Observed members within the same year are not adjusted.
4. Renormalization is applied once, after all interpolation is complete.

Within the no sanitation subgroup, nosandesc_ch is designated as a balancing indicator representing the remainder. It is never independently interpolated — in non-survey years it is set to the remainder after all other group members are accounted for (i.e. sinsan_ch – (nosanvecino_ch + nosanairelibre_ch)). This ensures the group constraint is maintained without artificially smoothing these structurally derived quantities.

Within the three primary groups, the unknown/unclassified categories (aguafdesconocido_ch, sandesconocido_ch) are included as active members of these compositional groups rather than derived as residuals. This is because these categories contain a genuine substantive component — households whose water source or sanitation facility cannot be classified — in addition to reflecting survey instrument variation across years. Renormalization is applied after interpolation and aggregation to ensure group sums remain valid. One indicator, aguafconnulo_ch, is treated as a true structural null: it is never interpolated and is left as NA in non-survey years. This indicator is a pure survey-instrument artifact with no meaningful trend to interpolate. It is retained in the output and included in its compositional group for observed years only.

Status Flags

Each row in the interpolated dataset carries status flags indicating whether the value represents an observed survey year or an interpolated estimate:

- observed: the value comes directly from a survey round for that year.
- interpolated: the value was estimated by linear interpolation between two observed survey years.
- no_data: no value is available for that year (outside the observed range for that country-indicator-stratum combination).

Relationship to JMP Methodology

The OLAS interpolation approach is explicitly aligned with the JMP methodology for SDG 6 monitoring in the following respects: use of linear interpolation between observed points, no higher-order smoothing, and treatment of interpolated values as estimates rather than observations. The primary difference is that OLAS does not perform forward extrapolation to the current year (MAX_EXTRAP_FWD = 0), whereas JMP projects to the current year to enable global monitoring. This reflects the OLAS priority of grounding estimates in observed country data.

2.3 Regional Estimates

The regional dataset provides population-weighted aggregate estimates of indicators for Latin America and the Caribbean overall and for four IDB subregions. These estimates are derived from the interpolated dataset and are accompanied by metadata on the number of contributing countries and the proportion of regional households represented by each estimate.

Geographic Coverage

Table 6 shows the five regional aggregates included in the dataset and their constituent countries.

Region Code	Region Name	Countries	N
LAC	Latin America and the Caribbean	All 26 countries	Overall regional aggregate
Andean	Andean Region	BOL, COL, ECU, PER, VEN	5 countries
Southern Cone	Southern Cone	ARG, BRA, CHL, PRY, URY	5 countries
Central America and Mexico	Central America and Mexico	CRI, GTM, HND, MEX, NIC, PAN, SLV	7 countries
Caribbean	Caribbean	BHS, BLZ, GUY, HTI, JAM, SUR, TTO, DOM	8 countries

Table 6: Regional aggregates and constituent countries

Aggregation Methodology

Regional estimates are computed as household-weighted means of the indicator value across contributing countries. The weight for each country is its total number of households in the relevant disaggregation stratum in a given year, derived from the interpolated dataset using the relationship:

$$\text{denominator} = \text{level\_final} / \text{value\_final}$$

This expression recovers the survey-weighted total number of households in the stratum from the two interpolated fields. A minimum value guard is applied: when value_final < 0.001, the denominator is set to missing to prevent unstable division from near-zero proportions producing implausibly large household counts. For the regional weighted average, each contributing country uses its household total (hh_total) derived as the median denominator across all indicators for that country-stratum-year, rather than the per-indicator level/value. For the household total used in weighting, the median of the denominator across all indicators available for a given country-year-stratum is used. This is more stable than relying on a single indicator's level/value, which can produce outliers due to rounding or near-zero values for rare indicators, which would in turn distort regional estimates. The regional value is then sum(level_final) / sum(hh_total) across countries contributing the given indicator only — not across all countries in the region.

The regional value is then computed as:

$$\text{regional\_value} = \Sigma(\text{level\_final}) / \Sigma(\text{denominator})$$

This is equivalent to a household-weighted mean, where countries with larger household populations contribute proportionally more to the regional estimate. The same formula is applied to all disaggregation strata — Total, Urban, Rural, all quintiles, all sex and ethnicity categories — not only the Total slice.

To preserve the integrity of compositional group sums, countries are only included in a group's regional aggregation if they have at least two members of that group present for the given stratum-year. A country with only one group member present — where that single indicator necessarily equals 1.0 or the group total — cannot be meaningfully combined with countries that have a full breakdown, as it would distort the regional group sum. Countries not belonging to any compositional group are unaffected by this filter. For the nosan subgroup specifically, countries where nosandesc_ch is the only nosan indicator present are excluded from the nosan regional aggregation entirely. Including such countries would create a mismatch between the country sets contributing to sinsan_ch and those contributing to the nosan indicators, which the post-aggregation renormalization step cannot resolve.

Compositional Renormalization

After aggregation, compositional renormalization is applied to regional estimates following a corrected hierarchy that differs slightly from the country-level approach. The detailed water source and sanitary installations in groups 1, 2, and 3 are renormalized directly so their members sum to 1. The corresponding summary/aggregation variables are then recomputed by summing their already-renormalized parent group. This hierarchy avoids double-dividing indicators that are shared across groups. The nosan subgroup (nosanvecino_ch, nosanairelibre_ch, nosandesc_ch) is renormalized to sum to sinsan_ch at the regional level.

Population Coverage Tracking

Because not all countries have data for all indicators in all years, regional estimates can be based on a subset of the subregion's households. The regional dataset includes explicit population coverage metadata so users can assess the representativeness of each estimate.

Coverage is computed at two levels:

- Contributing households (hh_contributing): the total number of households represented by the countries that actually contributed data to a specific indicator regional estimate in a given year.
- Potential households (hh_potential): the total number of households across all subregion member countries that had any household survey data in that year (regardless of indicator availability). This is the best available denominator for the subregion.
- Population coverage (pct_population_covered): hh_contributing / hh_potential × 100, expressed as a percentage.

This metadata is essential for interpreting estimates for indicators that are only available in a subset of countries, or for subregions with highly unequal country population sizes. For example, in the Caribbean subregion, **Haiti represents approximately half of total households**. Regional estimates for the Caribbean that include Haiti will show high population coverage and be strongly influenced by Haiti's values; estimates for years when Haiti lacks data will show low coverage and represent primarily small island states.

No threshold is applied to suppress estimates with low coverage. All estimates are retained but the pct_population_covered field enables users to apply their own coverage threshold appropriate to their analytical purpose.

Status Classification

Each regional estimate is assigned a status flag based on the data status of the contributing country values:

- all_observed: all contributing countries have observed survey values for that year.
- mixed: some contributing countries have observed values and some have interpolated values.
- all_interpolated: all contributing countries have interpolated values (no observed survey in that year).
- all_imputed: applies in edge cases where alternative imputation was used.

3. Data Use

3.1 Recommended Uses

The OLAS Household Survey Dataset is designed for the following uses:

- Trend analysis of water and sanitation access at the country level over time.
- Cross-country comparison within a given year.
- Disaggregated analysis by area, income quintile, sex, and ethnicity to understand equity dimensions of access.
- Regional benchmarking using the regional dataset, with attention to the pct_population_covered field.
- Policy monitoring aligned with SDG 6 targets, particularly for improved water and sanitation access.

3.2 Limitations and Caveats

Users should be aware of the following limitations when using the dataset:

- Interpolated values are estimates, not observations. They should not be treated with the same confidence as values derived from an actual survey round. The status_value field identifies interpolated years.
- No extrapolation is performed. For countries where the most recent survey is several years old, no estimates are available for the intervening years. This is done because linear extrapolation can produce unreliable results when extrapolating over many years.
- Cross-country comparison may be affected by differences in survey instruments. Countries where direct indicator mapping was not possible are documented in the harmonization documentation.
- Regional estimates with low pct_population_covered may not be representative of the region. For subregions with highly unequal country populations (particularly the Caribbean), the absence of data from a major country (e.g. Haiti) substantially reduces the representativeness of the regional estimate.
- The unknown/unclassified categories (aguafdesconocido_ch, sandesconocido_ch) are real categories that should not be ignored. They reflect both genuinely ambiguous facilities and survey instrument variation. Trends in these categories over time can reflect changes in survey design rather than changes in household behavior.
- Income quintiles and ethnicity dimensions are not available for all countries or survey years. Regional estimates for these dimensions will have lower population coverage than Total-level estimates.
- Standard errors are only available for observed years. Interpolated years do not carry uncertainty estimates. A quality-score field (1–5) is included in the final dataset to assist users in assessing reliability. For observed country rows, the score is based on the coefficient of variation (SE / value); for interpolated country rows it is based on the number of years to the nearest observed survey round; for regional estimates it is based on pct_population_covered. A score of 5 indicates highest quality and 1 indicates lowest (see Table 7).

Score	Country observed (CV)	Country interpolated (gap)	Regional (% population covered)
5	CV ≤ 0.05	—	≥ 90%
4	CV ≤ 0.10	1 year	≥ 75%
3	CV ≤ 0.20	2 years	≥ 50%
2	CV ≤ 0.30	3 years	≥ 25%
1	CV > 0.30	≥ 4 years	< 25%
NA	No SE available	—	—

Table 7: Calculation of quality_score

3.3 Accessing the Data

The dataset and associated documentation are available at: https://github.com/olas-lac/olashhs_2026.

Questions or comments about the dataset can be directed to the OLAS team at olasdata@iadb.org.

4. Conclusion

The OLAS Household Survey Dataset is a rich resource that allows users to explore water and sanitation information that comes directly from LAC countries using a schema that is largely comparable across countries. The information is granular, available over decades, and can be disaggregated by several dimensions relevant to policymaking.

The harmonization approach addresses the natural challenges of combining information from surveys with varying instruments and methodologies while the interpolated and regional products extend the analytical utility of the base dataset for trend analysis and regional benchmarking, with full transparency on data status and population coverage.

The dataset is designed to allow users to clearly identify what information is not available — through the null indicators, the status flags and population coverage fields — as well as what is available. Identification of data gaps is a key step in closing them throughout the region.

Users of the dataset are encouraged to read all sections of this methodology document and review the associated code repositories. Comments or questions can be sent to the OLAS at olasdata@iadb.org.

Annex A: Variable Coding

Variable	Description	Type	Format	Categories	
aguared_ch	If household has a connection to the water network on property or in house, regardless of use	Discrete	Numeric	0	No
				1	Yes
aguafconsumo_ch	Source used for drinking water in the household. If the survey does not differentiate between drinking water and water for other uses, assign 0. If it is not clear if a source is improved (ex. "pozo" instead of "pozo protegido"), assign 9.	Discrete	Numeric	0	Survey doesn't ask about drinking water or does not specify that the source is potable
				1	Distribution network, private tap or point of access.
				2	Distribution network, public point of access (standpipe, public tap)
				3	Bottled water
				4	A protected well (covered or protected)
				5	Collected rainwater
				6	Delivered or trucked water
				7	Another improved water source
				8	Directly from a surface water body
				9	Another non-improved water source
				10	A well, spring, or other source without a clear classification.
aguafuente_ch	Water source generally used in the household. If it is not clear if a source is improved (ex. "pozo" instead of "pozo protegido"), assign 9.	Discrete	Numeric	1	Distribution network, private tap or point of access.
				2	Distribution network, public point of access (standpipe, public tap)
				3	Bottled water
				4	A protected well (covered or protected)
				5	Collected rainwater
				6	Delivered or trucked water
				7	Another improved water source
				8	Directly from a surface water body
				9	Another non-improved water source
				10	A well, spring, or other source without a clear classification.
aguadist_ch	Location of water access	Discrete	Numeric	1	Inside the house
				2	Outside of the house but within the property or less than 100 meters from the house.

Variable	Description	Type	Format	Categories	
				3	Outside of the house and property, or more than 100 mts from the home.
				0	Does not specify
aguadisp1_ch	If water is available on a consistent basis. If the country asks if the household has access to water "when needed" fill this filed.	Discrete	Numeric	0	No
				1	Yes
				9	The survey doesn't ask this question
aguadisp2_ch	If water is available on a consistent basis. If water availability/continuity is asked in terms of days and hours of availability, fill this field.	Discrete	Numeric	1	Reports water shutoffs and that water is available less than half the time (3 days or less per week, less than 15 days per month)
				2	Reports water shutoffs but water is available half of the time or more (4 days per week, 15 days per month)
				3	Reports that there hasn't been any cutoff in the specified time period
				9	The survey doesn't ask this question.
aguamala_ch	Drinking water source is not improved. Unknown is for households that gave a response option that is not clearly improved or unimproved	Discrete	Numeric	0	No
				1	Yes
				2	Unknown
aguamejorada_ch	Drinking water/water for human consumption comes from an improved source. Unknown is for households that gave a response option that is not clearly improved or unimproved	Discrete	Numeric	0	No
				1	Yes
				2	Unknown
aguamide_ch	Household has a meter	Discrete	Numeric	0	No
				1	Sí
bano_ch	Type of sanitation facilities	Discrete	Numeric	0	Without instalations
				1	Toilet connected to the sewer network
				2	Toilet connected to a septic system
				3	Improved latrine or other improved toilet

Variable	Description	Type	Format	Categories	
				4	Toilet or latrine that drains to a surface water body or environment
				5	Other unimproved facility
				6	Facility that cannot be clearly classified
banoex_ch	Are sanitation facilities of exclusive use to the household	Discrete	Numeric	0	No
				1	Yes
				9	The question is not addressed
banomejorado_ch	Are sanitation facilities improved (SDG definition).	Discrete	Numeric	0	No
				1	Yes
				2	Cannot specify
sinbano_ch	For those without a sanitation facility, what do they do?	Discrete	Numeric	0	Has toilet
				1	Uses public facilities or those of a friend or neighbor
				2	Open defecation
				3	Does not have a toilet but does not specify what alternatives the household uses.
aguatrat_ch	Do households treat their water before they consume it?	Discrete	Numeric	0	No
				1	Yes
				9	Does not ask/cannot specify

Annex B: Indicator Coding

Indicators	Description	Definition	Topics
aguatrat_ch	Percent of households that treat their drinking water before consumption	if aguatrat_ch = 1	In-home water treatment
aguafconnulo_ch	Percent of households that lack data on the potability or information specifically about water for human consumption	if aguafconsumo_ch == 0	Water source for human consumption
aguafredcon_ch	Percent of households that use the piped water distribution network as their primary source for water used for human consumption (not including public standpipes)	if aguafconsumo_ch == 1	Water source for human consumption
aguafpublicocon_ch	Percent of households that use public standpipes as their primary water source used for human consumption	if aguafconsumo_ch == 2	Water source for human consumption
aguafembotcon_ch	Percent of households that use bottled water as their primary source for water used for human consumption	if aguafconsumo_ch == 3	Water source for human consumption
aguafpozoprotcon_ch	Percent of households that use a protected well as their primary source for water used for human consumption	if aguafconsumo_ch == 4	Water source for human consumption
aguaflluviacon_ch	Percent of households that use rainwater harvesting for human consumption	if aguafconsumo_ch == 5	Water source for human consumption
aguafcamioncon_ch	Percent of households that receive trucked water for human consumption	if aguafconsumo_ch == 6	Water source for human consumption
aguafotramejcon_ch	Percent of households that source water from other improved sources (protected springs) for human consumption	if aguafconsumo_ch == 7	Water source for human consumption
aguafsupercon_ch	Percent of households that source water for human consumption directly from surface water bodies.	if aguafconsumo _ch ==8	Water source for human consumption
aguafotranmcon_ch	Percent of households that source water for human consumption directly from other non-improved sources	if aguafconsumo_ch == 9	Water source for human consumption
aguafdesconcon_ch	Percent of households that use an unclassifiable watersource as their primary source for water used for human consumption	if aguafconsumo_ch == 10	Water source for human consumption And Water source for human

Indicators	Description	Definition	Topics
			consumption - summary
aguafnmcon_ch	Percent of households that use an unimproved water source as their primary source for water used for human consumption	if aguafconsumo_ch %in% c(8, 9)	Water source for human consumption - summary
aguafmejoradacon_ch	Percent of households that use an improved water source as their primary source for water used for human consumption	if aguafconsumo_ch %in% c(1:7)	Water source for human consumption – summary
aguared_ch	Percent of households that have a network connection	aguared_ch == 1	Network connection
aguafred_ch	Percent of households that use an on premises piped water distribution network as their primary water source	if aguafuente_ch == 1	General water source
aguafpublico_ch	Percent of households that use public standpipes as their primary water source	if aguafuente_ch == 2	General water source
aguafembotellada_ch	Percent of households that use bottled water as their primary source for water	if aguafuente_ch == 3	General water source
aguafpozoprot_ch	Percent of households that use protected wells as their primary source for water	if aguafuente_ch == 4	General water source
aguaflluvia_ch	Percent of households that use rainwater harvesting	if aguafuente_ch == 5	General water source
aguafcamion_ch	Percent of households that receive trucked water	if aguafuente_ch == 6	General water source
aguafotramej_ch	Percent of households that source water from other improved sources (protected springs)	if aguafuente_ch == 7	General water source
aguafsuperficial_ch	Percent of households that source water directly from surface water bodies.	if aguafuente_ch ==8	General water source
aguafotranm_ch	Percent of households that source water directly from other non-improved sources	if aguafuente_ch = 9	General water source
aguafdesconocido_ch	Percent of households that use an unclassifiable watersource as their primary source for water	if aguafuente_ch == 10	General water source / General water source - summary
aguafnomejorada_ch	Percent of households that use an unimproved water source as their primary source	if aguamala_ch ==1	General water source – summary

Indicators	Description	Definition	Topics
aguafmejorada_ch	Percent of households that use an improved water source as their primary source	if aguamejorada_ch ==1	General water source – summary
aguadispnulo_ch	Percent of population without information on continuity of water access	if aguadisp1_ch == 9 & aguadisp2_ch ==9	Availablity / Continuity
aguadispcontinuo_ch	Percent of households that report no water shutoffs	aguadisp2_ch %in% c(3)	Availablity / Continuity
aguadisp_ch	Percent of households with water available as needed (more than 50% of the time or explicitly stated) - JMP definition for SDG 6	if aguadisp1_ch == 1 OR if aguadisp1_ch == 9 & aguadisp2_ch %in% c(2,3)	Availablity / Continuity
aguamide_ch	Percent of households with a water meter	if aguamide_ch == 1	Utilities
sinsan_ch	Percent of households without sanitation facilities	if bano_ch == 0	Without sanitation / Sanitation facilities
nosanvecino_ch	Percent of households that rely on public or neighbors' facilities	if sinbano_ch == 1	Without sanitation
nosanairelibre_ch	Percent of households that practice open defecation	if sinbano_ch == 2	Without sanitation
nosandesc_ch	Percent of households without sanitation facilities with unknown alternatives	if sinbano_ch==3	Without sanitation
sanred_ch	Percent of households with a toilet connected to the sewer network	if bano_ch == 1	Sanitation facilities
sanseptic_ch	Percent of households with a toilet connected to a septic system	if bano_ch == 2	Sanitation facilities
sanotramejorado_ch	Percent of households with another improved sanitation facility	if bano_ch == 3	Sanitation facilities
sanotranm_ch	Percent of households with another unimproved sanitation facility	if bano_ch ==5	Sanitation facilities
sanambiente_ch	Percent of households whose sanitation facilities drain to the surface or a surface water body	if bano_ch ==4	Sanitation facilities
sandesconocido_ch	Percent of households with a sanitation facility that cannot be classified	if bano_ch %in% c(6)	Sanitation facilities
sanmejorado_ch	Percent of households with an improved sanitation facility	if bano_ch %in% c(1:3)	Sanitation facilities
sannomejorado_ch	Percent of households with unimproved sanitation facilities	if bano_ch %in% c(4,5)	Sanitation facilities
sanexclusivo_ch	Percent of households with exclusive access to sanitation facilities	if banoex_ch == 1	Exclusivity

Indicators	Description	Definition	Topics
sanexclusivonulo_ch	Percent of households with no information on exclusivity of sanitation facilities	if banoex_ch ==9	Exclusivity

IDB Water and Sanitation Division · Water and Sanitation Observatory for Latin America and the Caribbean (OLAS)

www.olasdata.org