Estimation methods for selected ICT indicators

ICT Data and Analytics Division, Telecommunication Development Bureau, ITU

This version: October 2023

This document presents the methods used to estimate missing country values for selected ICT indicators. Some of the data presented here are provisional, pending the completion of the current estimation exercise. Methods are regularly updated and refined to improve accuracy. The list of indicators for which estimates are produced is revised yearly based on needs, priorities and resource availability. Additionally, data availability varies from year to year, depending on data submissions. These changes in turn may affect the choice of models, results and other decisions. Also, the results reported in this documentation are based on the latest *completed* estimation cycle. For these reasons, this documentation is subject to change.

This version of the documentation covers the following ICT indicators:

Share of households with Internet

Share of individuals using the Internet

Share of individuals who own a mobile phone

Percentage of the population covered by mobile networks (at least 3G, at least LTE/4G)

Active mobile-broadband subscriptions

For more information on the definition of each indicator, refer to ITU Manual for Measuring ICT Access and Use by Households and Individuals and ITU Handbook for the Collection of Administrative Data on Telecommunications/ICT.

For enquiries about this documentation, please write to indicators@itu.int.

The latest version of the documentation is available in the section "Definitions, standards and methodology" of the IDA's website.

Share of households with Internet

Overview

The percentage of households with Internet access is estimated for all ITU Member States[1] for 2020-2022. While some countries have official data available for all years, many do not and require estimation by (1) simple imputation methods such as extrapolation, (2) regression modelling, or (3) using analogies with similar countries. Table 1 summarizes the sources for this indicator.

Table 1: Sources for data on the percentage of households with Internet access for ITU Member States, by region (provisional)

Source	All	Africa	Americas	Arab States	Asia & Pacific	CIS	Europe
Official	**50**	0	7	2	6	3	32
NSO estimate	**1**	1	0	0	0	0	0
Imputation	**37**	2	8	7	8	4	8
Model	**69**	22	16	7	18	2	4
Analogy	**15**	10	2	1	1	0	1
Other source	**6**	3	0	0	3	0	0
Share of total (%)	***90.4***	***86.4***	***94.3***	***77.3***	***87.8***	***100.0***	***97.8***

Overall, data for 178 countries are published using official data, simple imputation, national statistics office (NSO) estimates, modelling results, analogies with nearby countries, or other sources. These countries represented close to 100 per cent of global gross national income (GNI) and over 95 per cent of the world population in 2022 (Table 2). There are 19 countries for which estimation is not possible.[2]

Table 2 Sources for data on the percentage of households with Internet access for ITU Member States, by share of gross national income (GNI) and population (provisional)

Source	All	Share of 2022 GNI	Share of 2022 Population
Official	**50**	28.1	21.7
NSO estimate	**1**	0.0	0.0
Imputation	**37**	16.8	12.7
Model	**69**	53.3	51.4
Analogy	***15***	0.8	6.6
Other sources	**6**	0.5	3.2
Share of total	***90.4***	***99.5***	***95.6***

[1] ITU member states as well as (a) Hong Kong, China, (b) Macau, China, (c) Taiwan, Province of China, and (d) Palestine

[2] Afghanistan, Central African Republic, Comoros, Republic of the Congo, Democratic People's Republic of Korea, Haiti, Lesotho, Liberia, Libya, Mauritania, Micronesia, Myanmar, South Sudan, Syrian Arab Republic, Timor-Leste, Uganda, Ukraine, Venezuela, Yemen

Introduction

The percentage of households with access to the Internet is a core indictor used by the ITU to monitor connectivity. This indicator is normally derived from household surveys, which many countries do not administer, considering the cost and complexity. As a result, many countries do not report data. In addition, when countries submit data for this indicator, it can be for several years beforehand, considering the time it takes to process household surveys. This lag is important in the context of ICT statistics, because of the rapid pace of change in this domain compared to other social statistics derived from surveys.

In this context, estimations are needed to provide a fuller and timelier assessment. This section documents the methodology used to estimate missing data for the years 2020-2022.

Data availability

Questionnaire data is extracted from the database for the last three years (2020, 2021 and 2022) for all ITU Member States[3] using an SQL query. The base indicator used is *HHA4213_HHTotalHousehold*. Shares are calculated by dividing by the household target population (*HHA_HHTotalHousehold*). Data are complete for 50 countries and missing for at least one year in 148 countries.

Table 3 Household Internet access data availability for ITU Member States, 2020-2022

	Data years			Most recent data year	All	LDC
	2020	*2021*	*2022*			
All years	Yes	Yes	Yes	2022	50	0
Missing at least one year	-	-	-	-	147	46
One year	-	-	-	-	*22*	*1*
One year	No	Yes	Yes	2022	1	0
One year	Yes	No	Yes	2022	3	0
One year	Yes	Yes	No	2021	18	1
Two years	-	-	-	-	*29*	*5*
Two years	No	No	Yes	2022	5	1
Two years	No	Yes	No	2021	11	2
Two years	Yes	No	No	2020	13	2
All years	*No*	*No*	*No*	-	*96*	*40*
All years	No	No	No	2019	16	5
All years	No	No	No	2018	7	5
All years	No	No	No	2017	11	5
All years	No	No	No	2016	3	1
All years	No	No	No	2015	0	0
All years	No	No	No	2014	1	0
All years	No	No	No	< 2014	58	24

Note: Data availability as of 6 July 2023. LDC = Least Developed Country. Includes Palestine; Hong Kong, China; Macau, China; and Taiwan, Province of China

[3] Including Hong Kong, China; Macau, China; Palestine; and Taiwan, Province of China..

Data are sparser for Least Developed Countries (LDCs) and countries in Africa, Asia & Pacific and the Americas. However, each of these groups do have some data availability and give representation to these groups of countries in modelling.

Table 4 Household Internet access data availability for ITU Member States, by region 2020-2022

	All	Africa	Arab States	Asia & Pacific	CIS	Europe	Americas
All years	50	0	2	6	3	32	7
Missing at least one year	147	44	20	35	6	14	29
One year	*22*	*1*	*4*	*6*	*3*	*5*	*3*
Two years	*29*	*5*	*6*	*5*	*1*	*4*	*8*
All years	*96*	*38*	*10*	*24*	*2*	*5*	*18*

Note: Includes Palestine; Hong Kong, China; Macau, China; and Taiwan, Province of China

For modelling, data from 2020-2022 are used. Data are available from countries of varied regions and development levels. However, more data are available from higher income countries.

Methodology

Methodologies differed for countries missing data based on their availability of data in other years. For the 50 countries with data already available, data are used as available in surveys and no estimations were performed. In one additional country estimates from countries' national statistics offices are available and used and 6 other countries results from non-governmental surveys are used as estimates – notably data from Multiple Indicator Cluster Surveys (MICS) conducted by UNICEF.

Imputation

In cases where data are available in some but not all years and available data are reasonably consistent over time, imputation methods such as interpolation or extrapolation are used. These methods are used for 37 countries. Where data are available in previous years, data are extrapolated using the weighted average growth over the previous 2 or 3 years[4]. Where there are gaps in data, the geometric mean is used to fill these gaps. For countries with already high levels of individual Internet use, the most recent year data are used to fill gaps. There are also several special cases where imputation is possible and does not fall under these categories – these are cases where there are gaps in previous year data. Resulting estimations are assessed for consistency with previous year data and in the case of decreases or large increases consistency with underlying conditions in the country.

[4] For 3 years, growth in the previous year weighted at 0.5 (one half), 2 years previous at 0.33 (one third) and 3 years previous at 0.17 (one sixth). For 2 years, growth in previous year weighted at 0.6 and 2 years previous at 0.4. Some countries were also estimated using one previous year of stable growth due to a break in series or unavailable data. Countries with high growth were verified against modelled results.

Modelling

For the remaining 103 countries where data are fully unavailable or inconsistent, estimation is attempted through regression modelling[5]. The following variables are identified as potentially explanatory either extracted from the *delhi* SQL database or sourced from databases from other international organizations or private companies. In the case of Ookla data, these are extracted and averaged over all tests performed in a country. Data are excluded where tests are performed on fewer than 100 devices:

- Active mobile-broadband subscriptions per 100 inhabitants (i911mw)
- Fixed-broadband subscriptions per 100 inhabitants (i992b)
- GNI per capita (current US dollars) – the log of GNI also used given the skewed distribution among countries. Source is UN Statistics Division data as ITU database not fully updated. Data only available until 2021 so IMF GDP per capita growth rates (current USD) used to project data for 2022
- Population (i61) (source – UN Population Division) – the log of population also used given the skewed distribution among countries
- International bandwidth per Internet user (bit/s) (i994u) – the log of bandwidth also used given the skewed distribution among countries
- Mobile-cellular subscriptions per 100 inhabitants (i911) and Total telephone subscriptions per 100 inhabitants (i9111)
- Urban population (percentage) (source – World Bank) – ITU data not fully updated
- Affordability (price of 2 GB of mobile data as a percentage of GNI and in PPP)
- Affordability (price of limited voice and data plan as a percentage of GNI and in PPP)
- Affordability (price of voice-only plan as a percentage of GNI and in PPP)
- Labour force participation (15+ and 25+) (source – ILO)
- Median age of population (source – UN Population Division)
- Life expectancy (source – UN Population Division)
- Mean years of schooling (source – UNESCO)
- Mobile and fixed download speed (source – Ookla)
- Mobile and fixed latency (source – Ookla)

Additional variables have been considered but not used for modelling purposes due to poor data availability for countries requiring estimation[6]:

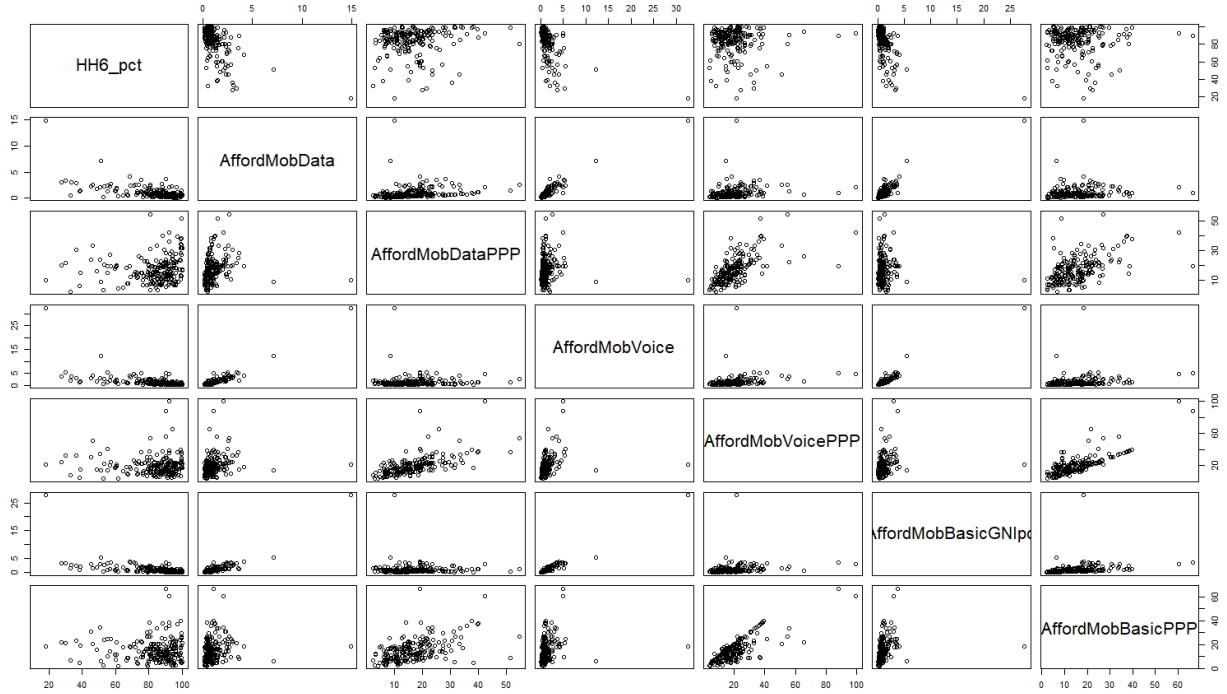
- Percentage of household with Internet access in previous year (HHA4213_HHTotalIndividual – calculated for previous year as share of total target pop)
- Percentage of individuals owning a mobile phone (HHU2710wn_HHTotalIndividual – calculated as share of total target pop)
- Affordability (price of basic device as a percentage of GNI and in USD)
- Government funding per student (source – UNESCO)
- Literacy rate (source – UNESCO)
- Government expenditure on education (source – UNESCO)

[5] Modelling was conducted using R.
[6] Data on number of cell towers per population are not used as an analysis conducted in 2022 revealed that it was a poor driver for similar estimates.

Model preparation

To maximize the dataset and balance against relevance of relationships with explanatory variables, data from 2020-2022 are used. This results in an overall dataset of 223 observations[7] from 101 countries of which six are Least Developed Countries (LDCs). Pairwise analysis of the data in the figures below shows a relatively strong negative relationship with *Affordability* (measured as percent of GNI) and a relatively strong positive relationship with *Active mobile-broadband subscriptions per 100 inhabitants, Fixed-broadband subscriptions per 100 inhabitants,* the log of *International bandwidth per Internet user* and both *Mobile and fixed download speed and latency*. Among non-ICT variables, *life expectancy*, and the log of *GNI per capita (current US dollars)* also appear to have a strong association with *Percentage of households with Internet access*.

Figure 1 Pairwise relationships between proportion of households with access to the Internet and affordability measures, 2020-2022



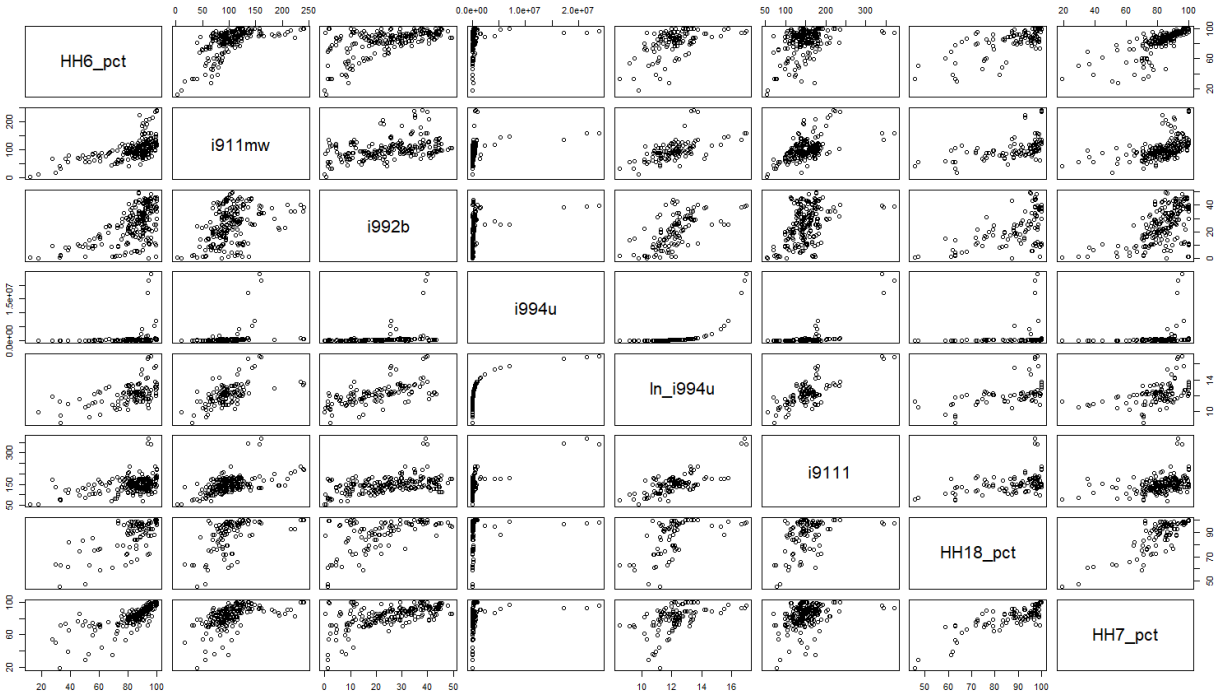
Note: Relationships are shown where pairwise data are available for country-data year combinations. HH6_pct = Households with Internet access (percentage), AffordMobData = price of data only mobile package as share of GNI (2 GB), AffordMobDataPPP = price of data only mobile package at purchasing price parity (PPP) (2 GB), AffordMobVoice = price of a voice-only mobile package as share of GNI, AffordMobVoicePPP = price of a voice-only mobile package at purchasing price parity (PPP), AffordMobBasicGNIpc = price of a limited voice mobile package as share of GNI (70 min + 20 SMS + 500 MB), AffordMobBasicPPP = price of a limited voice mobile package at purchasing price parity (PPP)

As in the estimation for Internet use, cross validation[8] is used. This method provides an alternative which captures the benefits of using the full data set and avoided overfitting. *K-fold* cross-validation was used (k = 10) as an alternative method to identify important explanatory variables. Specifically, this entails dividing the dataset into 10 "folds", running 10 separate regressions with each of the folds removed and testing on the removed fold. The

[7] up to 3 per country for different years. Data as of 1 August 2023
[8] https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation

best performing group of variables is selected based on the lowest root mean square error (RMSE). Traditional stepwise regression was used in previous years. This has been discontinued due to the improved concepts behind cross validation.

Figure 2 Pairwise relationships between proportion of households with Internet access and ITU ICT indicators, 2020-2022



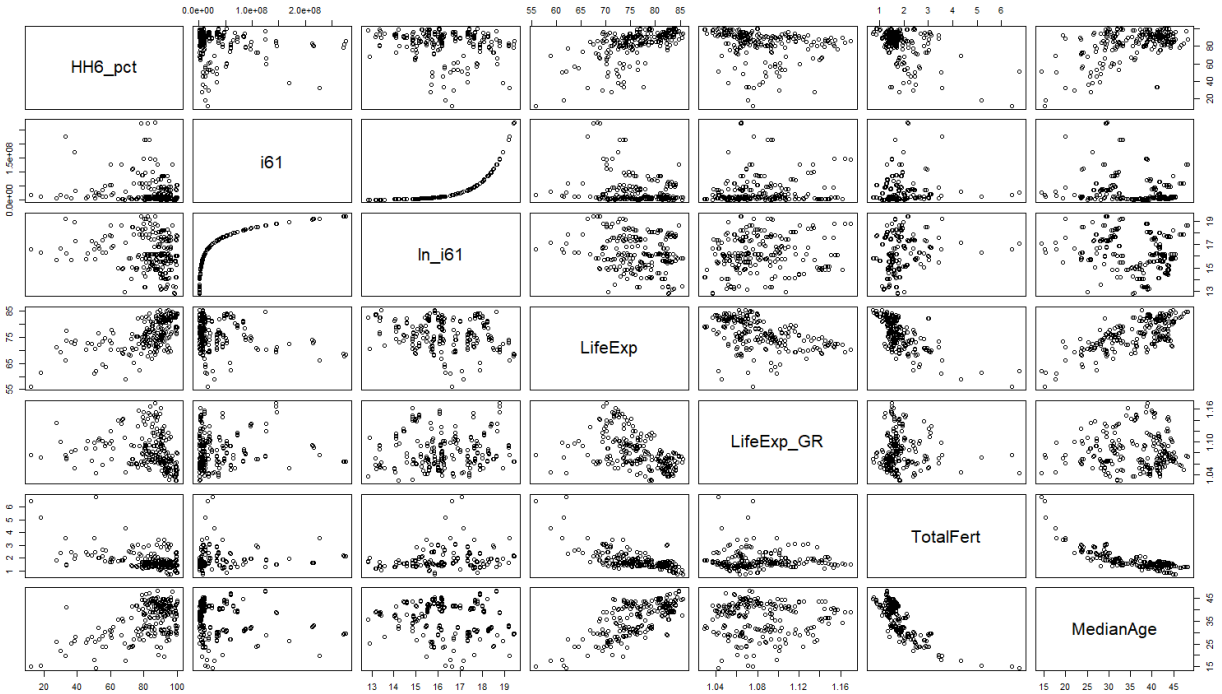
Note: Relationships are shown where pairwise data are available for country-data year combinations. HH6_pct = Households with Internet access (percentage), i911mw = Active mobile-broadband subscriptions per 100 inhabitants, i992b = Fixed-broadband subscriptions per 100 inhabitants, i99H = Internet users (percentage), i994u = International bandwidth per Internet user (bit/s), ln_i994u = International bandwidth per Internet user (log bit/s), i9111 = Total telephone subscriptions per 100 inhabitants, HH18_pct = Share of individuals who own a mobile phone, HH7_pct = Individuals using the Internet (percentage)

Data points are also weighted to avoid overweighting wealthier countries which provide data to the ITU more frequently. The weights are calculated in multiple steps. First, the regional weights are calculated based on the number of countries in each region – this assumes equal weight for all countries regardless of population or economic output. As an example, for the Arab States there are 22 countries or a 11.2 per cent share of the 197 total ITU members and additional economies. This is done in accordance with the broad *one country one vote* principle in the UN system. Second, these regional weights are divided equally among countries reporting data in the region. For the Arab States, 12 countries reported data during the 2020-2022 period so each receives 11.2 / 12 = 0.9 weight. Where countries provided data for more than one year, the weight for each data point is divided by the number of data points provided by the country. United Arab Emirates, for example, provided three years of data and its weight then is 0.9 / 3 = 0.3. This results in a range of weights from 0.2 for one data point from European countries providing three years of data (0.6 full weight for the country) to 4.5 for African countries providing only one year of data.

Two different methods are used to estimate the proportion of individuals using the Internet. In the first, the proportion is estimated directly through ordinary linear regression (OLS). While the simplest method, estimating in this way allows for the possibility of negative

proportions or proportions over 100 per cent. To avoid this problem a logit (log odds) transformation[9] of the share of individuals using the Internet is also used as the dependent variable in OLS. This is particularly useful in estimating countries with very low or high levels of Internet use.

Figure 3 Pairwise relationships between proportion of households with Internet access and possible socio-economic explanatory variables from the UN Population Division, 2020-2022

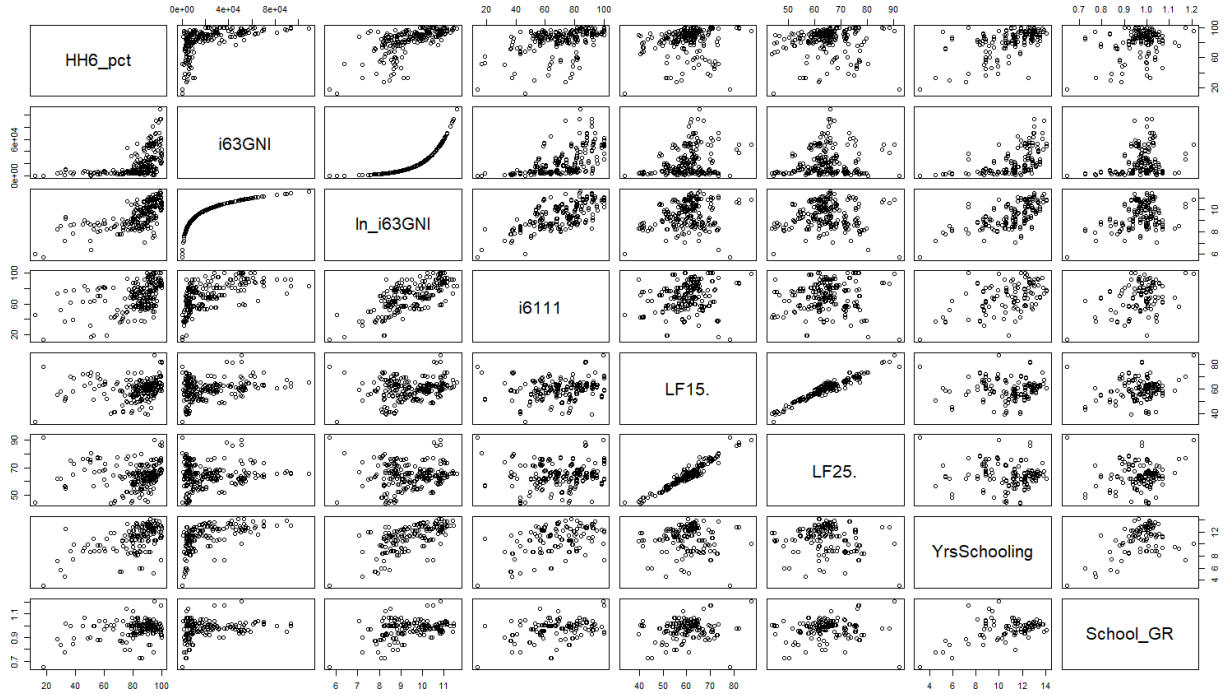


Note: Relationships are shown where pairwise data are available for country-data year combinations. HH6_pct = Households with Internet access (percentage), i61 = population, ln_i61 = log population, LifeExp = life expectancy in years, LifeExp_GR = gender ratio of life expectancy (female/male), TotalFert = total fertility rate (average number of expected for women given current fertility rates), MedianAge = median age of population

Comparing models poses a challenge for multiple reasons. First, different errors are minimized in these two different methods. The logit transformation does not directly minimize the error between predicted and actual percentage of individuals using the Internet as it minimizes the error between the predicted and actual *log odds* of individuals using the Internet. Models estimating the variable in question directly are preferred but are analysed on a case-by-case basis to avoid impossible results outside of the 0 to 100 range of percentages.

[9] logit(p)= ln(p/(1-p)), where p = probability and p/(1-p) is the corresponding odds

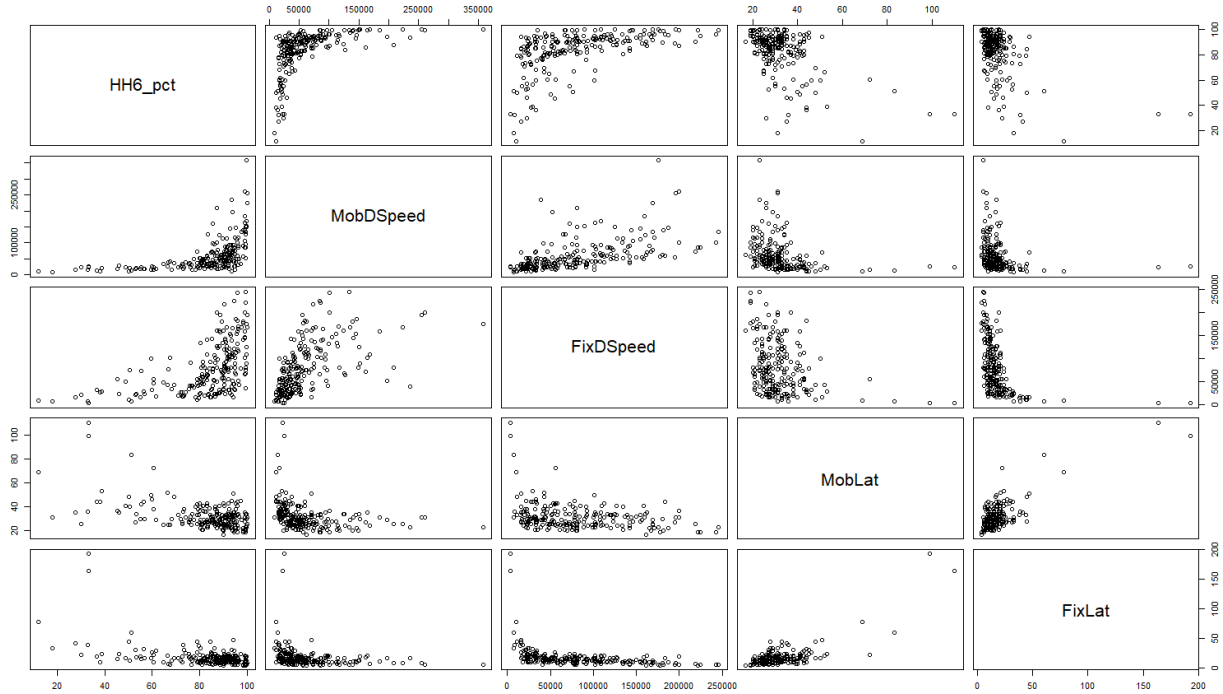
Figure 4 Pairwise relationships between proportion of households with Internet access and possible other socio-demographic explanatory variables, 2020-2022



Note: Relationships are shown where pairwise data are available for country-data year combinations. HH6_pct = Households with Internet access (percentage), i63GNI = Gross national income (GNI) per capita (current US dollars), ln_i63GNI = GNI per capita (log current US dollars), i6111 = Urban population (percentage), LF15 = Share of population 15+ in labour force, LF25 = Share of population 25+ in labour force, YrsSchooling = mean number of years of schooling, School_GR = gender ratio of mean number of years of schooling (female/male).

In addition, there is missingness for many of the explanatory variables. Most notably, of the 223 observations with data on household Internet access, 62 are missing data on *Mean years of schooling* and 105 are missing data on *International bandwidth per Internet user*. On the estimation side, data availability as inputs for eventual models is also considered. The major limitation of mean years of schooling is the lack of data for 2022 – as a result there were 151 observations missing out of 368 country-year combinations requiring estimation. Even more limiting and important to consider are missing data for *International bandwidth per Internet user* (missing 251 observations of those requiring estimation). Other notable variable that limit estimation are *Fixed-broadband subscriptions per 100 inhabitants* (51 missing), *Active mobile-broadband subscriptions per 100 inhabitants* (missing 55), *Share of population aged 25+ in the labour force* (50 missing).

Figure 5 Pairwise relationships between proportion of households with Internet access and possible Internet speed explanatory variables, 2020-2022



Note: Relationships are shown where pairwise data are available for country-data year combinations. HH6_pct = Household with Internet access (percentage), MobDSpeed = Average mobile download speed (Kbps), FixDSpeed = Average fixed network download speed (Kbps), MobLat = Mobile latency (ms), FixLat = Fixed network latency (ms).

In addition, countries providing data on household Internet access tend to have low prices relative to GNI. As a result, many countries missing data on household Internet access also have prices relative to GNI that are well out of scope of models. Price is then a less useful variable for estimating household Internet access for such countries.

To compensate for the issues mentioned above, data sets excluding some or all these variables are also used for model comparison in addition to the full dataset. Other variables dropped were *Share of population 15+ in labour force* due to its strong correlation with *Share of population 25+ in labour force*

Modelling results

Variables are selected based on differing datasets and methods as shown in Table 5. The most consistent variable included in models is mobile-broadband subscriptions p*er 100 inhabitants*. The relationship is positive in all cases – consistent with the idea that mobile phones are often the primary device for Internet access. The *log of GNI per capita* is also selected frequently, including for nearly all linear models. Other variables are selected more frequently for logit models. This includes *life expectancy* and *mobile download speed*. Affordability measures also are key parameters in some but not many models. Other variables selected, but less frequently included *fixed-broadband subscriptions per 100 inhabitants*, *total fertility*, *share of population 25+ in the labour force* and *mobile latency*.

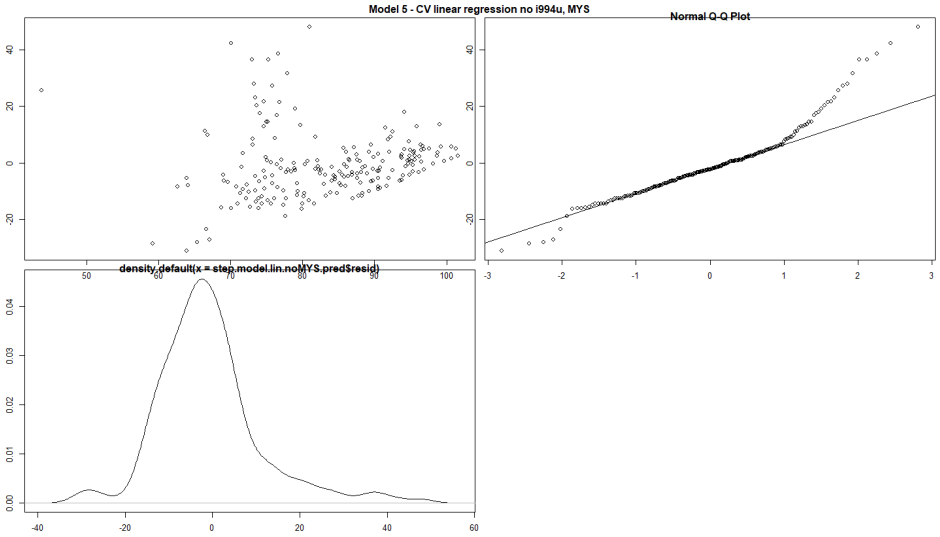
Table 5 Variables included in initial models selected using various variable selection methods

#	Transform	Exclusions	Weighted	n	i911mw	i992b	i63GNI	ln_i63GNI	i9111	i6111	AffordMobDataG	AffordMobDataPP	AffordMobBasicG	TotalFert	LifeExp	LifeExp_GR	LF25.	YrsSchooling	School_GR	MobileDSpeed	MobLat
1			Y	86				x													
2				86				x													
3		a	Y	155				x													
4		a		155	x	x							x								x
5		b	Y	206				x					x								
6		b		206				x			x										
7		c	Y	159	x				x					x		x		x	x		x
8		d	Y	211	x		x	x													x
9		e	Y	214	x		x	x		x				x		x	x				
15	Y		Y	86	x				x			x			x					x	
16	Y			86	x																
17	Y	a	Y	155	x										x					x	
18	Y	a		155	x										x					x	
19	Y	b	Y	206				x												x	
20	Y	b		206	x										x					x	
21	Y	c	Y	159	x										x					x	
22	Y	d	Y	211	x										x					x	
23	Y	e	Y	214	x			x							x	x					

Note: a= Excluding i994u, b = Excluding i994u and mean years of schooling, c = Excluding i994u and affordability measures, d = Excluding i994u, mean years of schooling and affordability measures, e = Reduced model including only mobile subscriptions, GNI per capita, share of pop in labour force, life expectancy, gender ratio of life expectancy, total fertility, median age, and share of urban population. n = number of observations on which the model is based. i911mw = Mobile-broadband subscriptions per 100 inhabitants, i992b = Fixed-broadband subscriptions per 100 inhabitants, i63GNI = Gross national income (GNI) per capita (current US dollars), ln_i63GNI = GNI per capita (log current US dollars), i9111 = Total telephone subscriptions per 100 inhabitants, i6111 = Urban population (percentage), AffordMobData = price of data only mobile package as share of GNI (2 GB), AffordMobDataPPP = price of data only mobile package at purchasing price parity (PPP) (2 GB), AffordMobBasicGNIpc = price of a limited voice mobile package as share of GNI (70 min + 20 SMS + 500 MB), TotalFert = total fertility rate (average number of expected for women given current fertility rates), LifeExp = life expectancy in years, LifeExp_GR = gender ratio of life expectancy (female/male). LF25 = Share of population 25+ in labour force, YrsSchooling = mean number of years of schooling, School_GR = gender ratio of mean number of years of schooling (female/male). MobDSpeed = Average mobile download speed (Kbps), MobLat = Mobile latency (ms).

When regressed against full data sets, three of the above models demonstrated clear evidence of non-normality and were rejected – all three were linear models (models 4, 5, and 6). Residual plots for Model 5 – linear model excluding *international bandwidth per Internet user* and *mean years of schooling* are shown in Figure 6 and demonstrate the violation of regression assumptions.

Figure 6 Residual plots for model 5 – weighted cross validation linear model excluding i994u and mean years of schooling



Certain models are preferred based on the characteristics of the country. For example, non-transformed linear models are not preferred for countries with likely very high or very low level of household Internet access. Unweighted models are not preferred for poorer countries. Such countries are not well-represented in these models as more weight is implicitly given to wealthier countries that have provided more data. Models excluding price are preferred for countries with high prices relative to GNI as the input to such models for these countries would be out of scope.

Estimation

The overall strategy for model selection for use in estimation is to align with previous official data submitted and to use primary models as possible. Models of the log odds (logit) of percentage of individuals using the Internet are preferred for countries with likely low or high values. Many countries are missing data for explanatory variables used in some models. In these cases, the choice of models is limited.

Fifty countries are estimated directly without complication. These estimations are all consistent with previously provided official data or previous estimations. Choice of model is split between models of the log odds of household Internet access (35 countries) and linear (15 countries).

For 19 additional countries models are used partially or indirectly. In 10 of these countries data required for modelling are not available in all years or modelled results are inconsistent for some years. Modelled estimations are used in years with available data and the estimated Internet use growth are used in other years. In nine other countries, estimations do not align with available official data – in these cases the growth in modelled data is used to project or backcast estimations based on these official data.

Other methods

For countries where modelling does yield usable results, analogies with Internet use are used directly. In these cases, the ratio of Internet use share to household Internet access share is

used in the most recent year where official data is available for each and applied to the Internet use estimates.[10]

Remaining countries

There are 19 remaining countries for which estimations are calculated, but not recommended for publication at the individual country-level.

[10] Regional averages as well and as averages for groups such as LDCs, land-locked developing countries (LLDCs) and small island developing states (SIDS) were were also considered. However, this was was not a satisfactory solution, particularly for countries in groups were there were few available data.

Share of individuals using the Internet

Overview

The percentage of individuals using the Internet is estimated for all ITU Member States[11] for 2020-2022. While some countries have official data available for all years, many do not and require estimation by (1) simple imputation methods such as extrapolation, (2) regression modelling, or (3) using analogies with similar countries. Table 6 summarises the sources for this indicator.

Table 6 Sources for data on the share of individuals using the Internet for ITU Member States, by region

Source	All	Africa	Americas	Arab States	Asia & Pacific	CIS	Europe
Official	**51**	0	7	2	8	3	31
NSO estimate	**2**	2	0	0	0	0	0
Imputation	**26**	0	5	7	5	3	6
Model	**89**	32	19	11	21	2	4
Analogy	**4**	0	0	0	0	0	4
Other source	**15**	9	2	0	4	0	0
Share of total (%)	***94.9***	***97.7***	***94.3***	***90.9***	***92.7***	***88.9***	***97.8***

Overall, 187 countries are published using official data, simple imputation, national statistics office (NSO) estimates, modelling results, analogies with nearby countries, or other sources. These countries represent close to 100 per cent of global gross national income (GNI) and world population in 2022. There are 10 countries for which estimations is not feasible.[12]

Table 7 Sources for data on the share of individuals using the Internet for ITU Member States, by share of gross national income (GNI) and population

Source	All	Share of 2022 GNI (%)	Share of 2022 population (%)
Official	**51**	46.3	39.6
NSO estimate	**2**	0.1	1.3
Imputation	**26**	12.0	8.4
Model	**89**	35.1	24.3
Analogy	**4**	0.0	0.0
Other source	**15**	6.0	23.5
Share of total	***94.9***	***99.5***	***97.0***

Introduction

The percentage of individuals using the Internet is a core indictor used by the ITU to monitor connectivity. This indicator is used as a main reference point in publications such as ITU's

[11] ITU member states as well as (a) Hong Kong, China, (b) Macau, China, (c) Taiwan, Province of China, and (d) Palestine

[12] Afghanistan, Central African Republic, Comoros, Democratic People's Republic of Korea, Haiti, Myanmar, Syrian Arab Republic, Turkmenistan, Ukraine, and Venezuela.

annual *Measuring Digital Development: Facts and Figures*. This indicator is normally derived from household surveys, which many countries do not administer, considering the cost and complexity. As a result, many countries do not report data. In addition, when countries submit data for this indicator, it can be for several years beforehand, considering the time it takes to process household surveys. This lag is important in the context of ICT statistics, because of the rapid pace of change in this domain compared to other social statistics derived from surveys.

In this context, estimations are needed to provide a fuller and timelier assessment. This section documents the methodology used to estimate missing data for the years 2020-2022.

Data availability

Questionnaire data was extracted from the database for the last three years (2020, 2021 and 2022) for all ITU Member States[13] using an SQL query. The base indicator used was *HHU4212_HHTotalIndividual*. Shares were calculated by dividing by the individual target population (*HHU_HHTotalIndividual*). Data were complete for 51 countries and missing for at least one year in 147 countries.

Table 8 Individual Internet use data availability for ITU Member States, 2020-2022

	Data years			Most recent data year	All	LDC
	2020	*2021*	*2022*			
All years	Yes	Yes	Yes	2022	51	0
Missing at least one year	-	-	-	-	146	46
One year	-	-	-	-	*20*	*1*
One year	No	Yes	Yes	2022	1	0
One year	Yes	No	Yes	2022	3	0
One year	Yes	Yes	No	2021	16	1
Two years	-	-	-	-	*29*	*5*
Two years	No	No	Yes	2022	4	0
Two years	No	Yes	No	2021	12	3
Two years	Yes	No	No	2020	13	2
All years	*No*	*No*	*No*	-	*97*	*40*
All years	No	No	No	2019	11	3
All years	No	No	No	2018	5	3
All years	No	No	No	2017	10	6
All years	No	No	No	2016	1	1
All years	No	No	No	2015	0	0
All years	No	No	No	2014	2	1
All years	No	No	No	< 2014	68	26

Note: Data availability as of 1 August 2023. LDC = Least Developed Country. Includes Palestine; Hong Kong, China; Macau, China; and Taiwan, Province of China

[13] Including Hong Kong, China; Macau, China; Palestine; and Taiwan, Province of China..

Data are sparser for Least Developed Countries (LDCs) and countries in Africa, Asia & Pacific and the Americas. However, each of these groups have some data availability and give representation to these groups of countries in modelling.

Table 9 Individual Internet use data availability for ITU Member States, by region 2020-2022

	All	Africa	Arab States	Asia & Pacific	CIS	Europe	Americas
All years	51	0	2	8	3	31	7
Missing at least one year	146	44	20	33	6	15	28
One year	*20*	*0*	*4*	*5*	*3*	*5*	*3*
Two years	*29*	*7*	*5*	*5*	*1*	*4*	*7*
All years	*97*	*37*	*11*	*23*	*2*	*6*	*18*

Note: Includes Palestine; Hong Kong, China; Macau, China; and Taiwan, Province of China

For modelling, data from 2020-2022 are used. Data are available from countries of varied regions and development levels. However, more data are available from higher income countries.

Methodology

Methodologies differ for countries missing data based on their availability of data in other years. For the 51 countries with data already available, data are used as available in surveys and no estimations are performed. In two additional countries, estimates from countries' national statistics offices are available and used and 15 other countries results from non-governmental surveys are used as estimates.

Imputation

In cases where data are available in some but not all years and available and data are reasonably consistent over time, imputation methods such as interpolation or extrapolation are used. These methods are used for 26 countries. Where data are available in previous years, data are extrapolated using the weighted average growth over the previous 2 or 3 years[14]. Where there are gaps in data, the geometric mean is used to fill these gaps. For countries with already high levels of individual Internet use, the most recent year data are used to fill gaps. There are also several special cases where imputation is possible and does not fall under these categories – these are cases where there are gaps in previous year data. Resulting estimations were assessed for consistency with previous year data and in the case of decreases or large increases consistency with underlying conditions in the country.

[14] For 3 years, growth in the previous year weighted at 0.5 (one half), 2 years previous at 0.33 (one third) and 3 years previous at 0.17 (one sixth). For 2 years, growth in previous year weighted at 0.6 and 2 years previous at 0.4. Some countries are also estimated using one previous year of stable growth due to a break in series or unavailable data. Countries with high growth are verified against modelled results.

Modelling

For the remaining 103 countries where data are fully unavailable or inconsistent, estimation is attempted through regression modelling[15]. The following variables are identified as potentially explanatory either extracted from the *delhi* SQL database or sourced from databases from other international organizations or private companies. In the case of Ookla data, these are extracted and averaged over all tests performed in a country. Data are excluded where tests are performed on fewer than 100 devices:

- Active mobile-broadband subscriptions per 100 inhabitants (i911mw)
- Fixed-broadband subscriptions per 100 inhabitants (i992b)
- GNI per capita (current US dollars) – the log of GNI also used given the skewed distribution among countries. Source is UN Statistics Division data as ITU database not fully updated. Data only available until 2021 so IMF GDP per capita growth rates (current USD) used to project data for 2022
- Population (i61) (source – UN Population Division) – the log of population also used given the skewed distribution among countries
- International bandwidth per Internet user (bit/s) (i994u) – the log of bandwidth also used given the skewed distribution among countries
- Mobile-cellular subscriptions per 100 inhabitants (i911) and Total telephone subscriptions per 100 inhabitants (i9111)
- Urban population (percentage) (source – World Bank) – ITU data not fully updated
- Affordability (price of 2 GB of mobile data as a percentage of GNI and in PPP)
- Affordability (price of limited voice and data plan as a percentage of GNI and in PPP)
- Affordability (price of voice-only plan as a percentage of GNI and in PPP)
- Labour force participation (15+ and 25+) (source – ILO)
- Median age of population (source – UN Population Division)
- Life expectancy (source – UN Population Division)
- Mean years of schooling (source – UNESCO)
- Mobile and fixed download speed (source – Ookla)
- Mobile and fixed latency (source – Ookla)

Additional variables have been considered but not used for modelling purposes due to poor data availability for countries requiring estimation[16]:

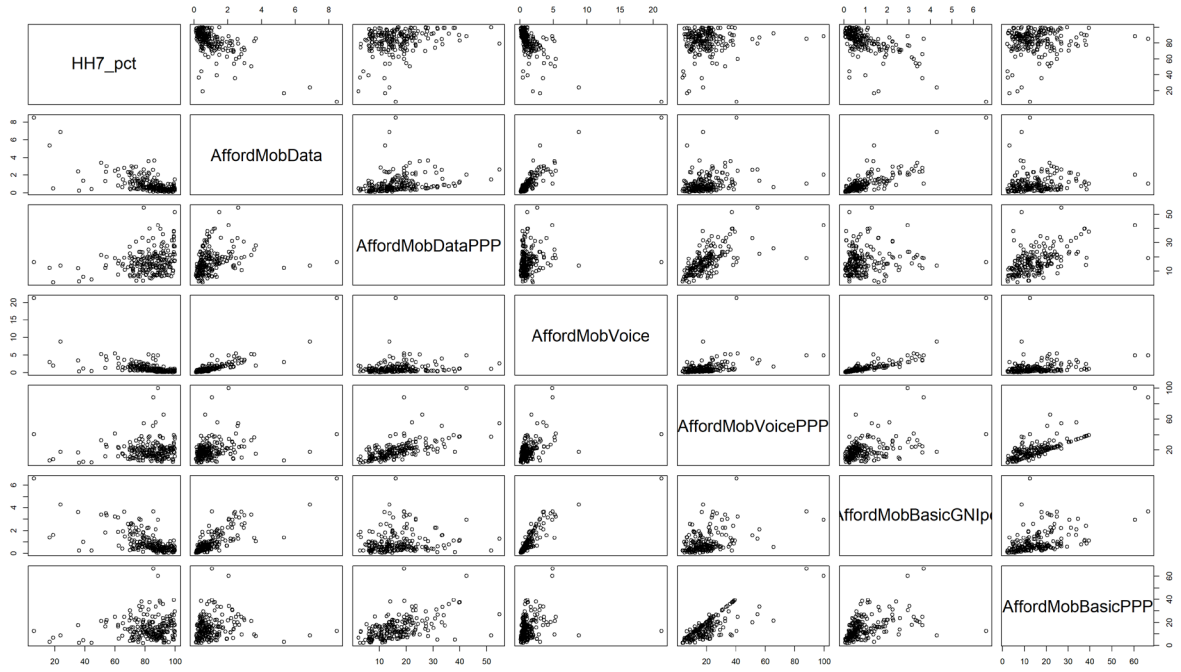
- Percentage of individuals using the Internet in previous year (HHU4212_HHTotalIndividual – calculated for previous year as share of total target pop)
- Percentage of households with Internet access (HHA4213_HHTotalHousehold– calculated as share of total target pop)
- Percentage of individuals owning a mobile phone (HHU271own_HHTotalIndividual – calculated as share of total target pop)
- Affordability (price of basic device as a percentage of GNI and in USD)
- Government funding per student (source – UNESCO)
- Literacy rate (source – UNESCO)
- Government expenditure on education (source – UNESCO)

[15] Modelling conducted using R.
[16] Data on number of cell towers per population are not used as an analysis conducted in 2022 revealed that it was a poor driver for similar estimates.

Model preparation

To maximize the dataset and balance against relevance of relationships with explanatory variables, data from 2020-2022 are used. This results in an overall dataset of 222 observations[17] from 100 countries of which six are Least Developed Countries (LDCs). Pairwise analysis of the data in the figures below shows a relatively strong negative relationship with *Affordability* (measured as percent of GNI) and a relatively strong positive relationship with *Active mobile-broadband subscriptions per 100 inhabitants, Fixed-broadband subscriptions per 100 inhabitants,* the log of *International bandwidth per Internet user* and both *Mobile and fixed download speed*. Among non-ICT variables, *life expectancy*, *total fertility*, *mean years of schooling*, *urban share of population,* and the log of *GNI per capita (current US dollars)* also appear to have a strong association with *Percentage of individuals using the Internet*.

Figure 7 Pairwise relationships between proportion of individuals using the Internet and affordability measures, 2020-2022



Note: Relationships are shown where pairwise data are available for country-data year combinations. HH7_pct = Individuals using the Internet (percentage), AffordMobData = price of data only mobile package as share of GNI (2 GB), AffordMobDataPPP = price of data only mobile package at purchasing price parity (PPP) (2 GB), AffordMobVoice = price of a voice-only mobile package as share of GNI, AffordMobVoicePPP = price of a voice-only mobile package at purchasing price parity (PPP), AffordMobBasicGNIpc = price of a limited voice mobile package as share of GNI (70 min + 20 SMS + 500 MB), AffordMobBasicPPP = price of a limited voice mobile package at purchasing price parity (PPP)

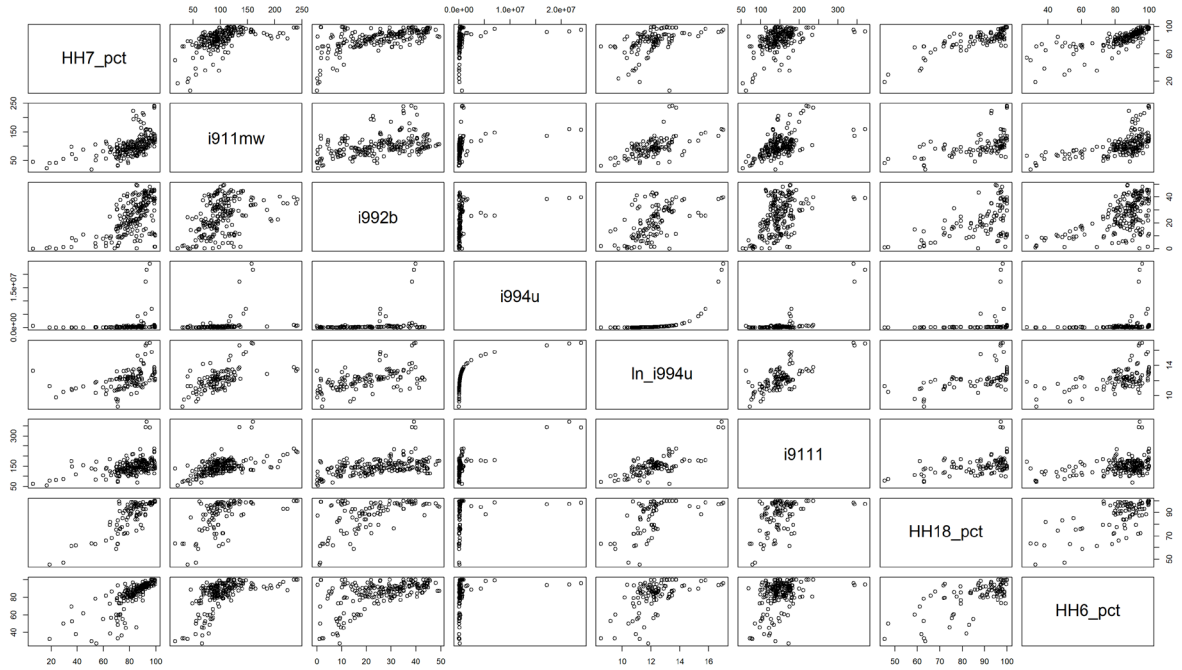
As in the 2022 estimation, cross validation[18] is used. This method provides an alternative which captures the benefits of using the full data set and avoids overfitting. *K-fold* cross-validation was used (k = 10) as an alternative method to identify important explanatory variables. Specifically, this entailed dividing the dataset into 10 "folds", running 10 separate

[17] up to 3 per country for different years. Data as of 1 August 2023

[18] https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation

regressions with each of the folds removed and testing on the removed fold. The best performing group of variables is selected based on the lowest root mean square error (RMSE). Traditional stepwise regression was used in previous years. This has been discontinued due to the improved concepts behind cross validation.

Figure 8 Pairwise relationships between proportion of individuals using the Internet and ITU ICT indicators, 2020-2022



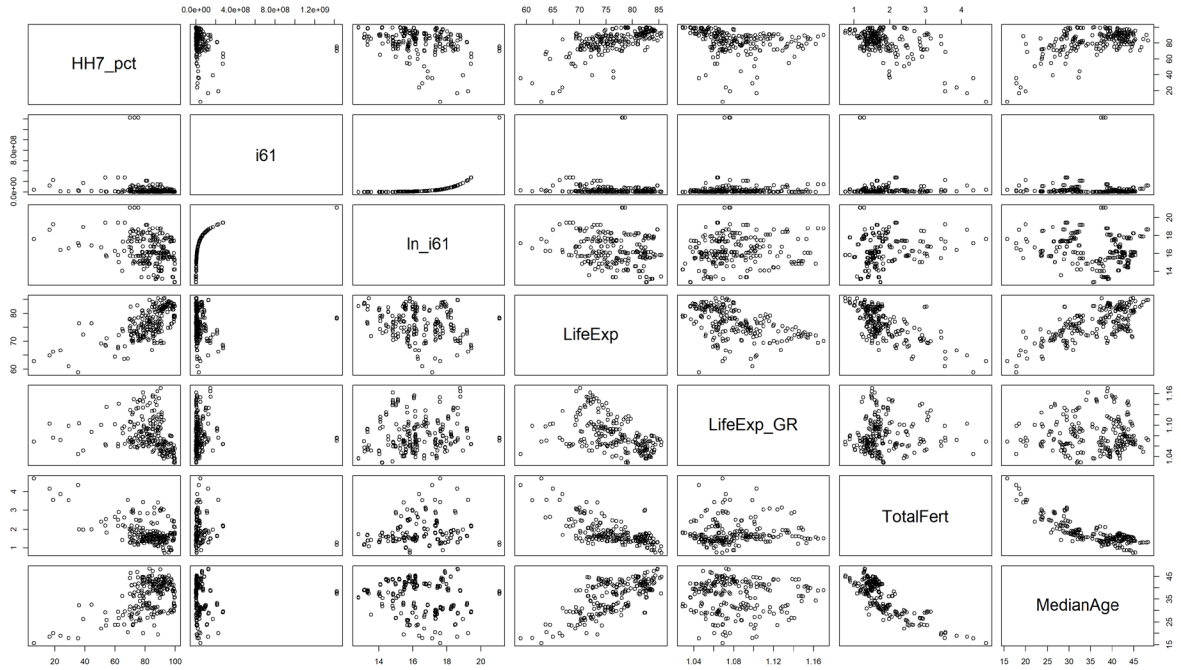
Note: Relationships are shown where pairwise data are available for country-data year combinations. HH7_pct = Individuals using the Internet (percentage), i911mw = Active mobile-broadband subscriptions per 100 inhabitants, i992b = Fixed-broadband subscriptions per 100 inhabitants, i99H = Internet users (percentage), i994u = International bandwidth per Internet user (bit/s), ln_i994u = International bandwidth per Internet user (log bit/s), i9111 = Total telephone subscriptions per 100 inhabitants, HH18_pct = Share of individuals who own a mobile phone, HH6_pct = Share of household with access to the Internet.

Data points are also weighted to avoid overweighting wealthier countries which provide data to the ITU more frequently. The weights are calculated in multiple steps. First, the regional weights are calculated based on the number of countries in each region – this assumes equal weight for all countries regardless of population or economic output. As an example, for the Arab States there are 22 countries or a 11.2 per cent share of the 197 total ITU members and additional economies. This is done in accordance with the broad *one country one vote* principle in the UN system. Second, these regional weights are divided equally among countries reporting data in the region. For the Arab States 11 countries reported data during the 2020-2022 period so each receives 11.2 / 11 = 1.0 weight. Where countries provided data for more than one year, the weight for each data point is divided by the number of data points provided by the country. United Arab Emirates, for example, provided three years of data and its weight then is 1.0 / 3 = 0.3. This results in a range of weights from 0.2 for one data point from European countries providing three years of data (0.6 full weight for the country) to 3.2 for African countries providing only one year of data.

Two different methods are used to estimate the proportion of individuals using the Internet. In the first, the proportion is estimated directly through ordinary linear regression (OLS).

While the simplest method, estimating in this way allows for the possibility of negative proportions or proportions over 100 per cent. To avoid this problem a logit (log odds) transformation[19] of the share of individuals using the Internet is also used as the dependent variable in OLS. This is particularly useful in estimating countries with very low or high levels of Internet use.

Figure 9 Pairwise relationships between proportion of individuals using the Internet and possible socio-economic explanatory variables from the UN Population Division, 2020-2022

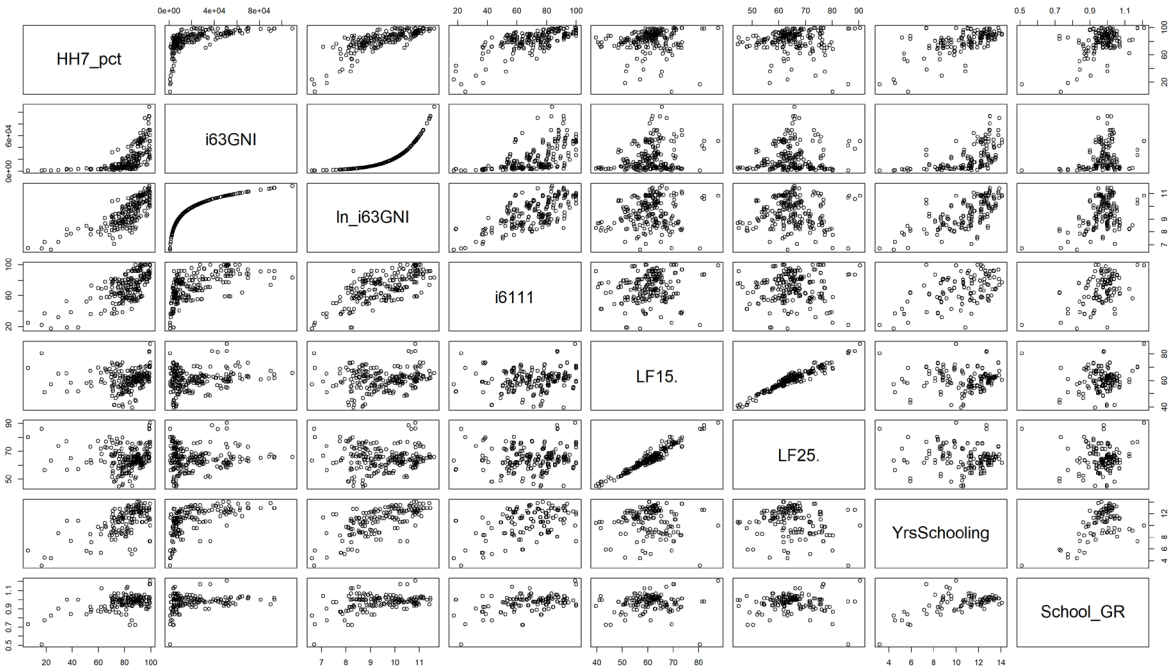


Note: Relationships are shown where pairwise data are available for country-data year combinations. HH7_pct = Individuals using the Internet (percentage), i61 = population, ln_i61 = log population, LifeExp = life expectancy in years, LifeExp_GR = gender ratio of life expectancy (female/male), TotalFert = total fertility rate (average number of expected for women given current fertility rates), MedianAge = median age of population.

Comparing models poses a challenge for multiple reasons. First, different errors are minimized in these two different methods. The logit transformation does not directly minimize the error between predicted and actual percentage of individuals using the Internet as it minimizes the error between the predicted and actual *log odds* of individuals using the Internet. Models estimating the variable in question directly are preferred but are analysed on a case-by-case basis to avoid impossible results outside of the 0 to 100 range of percentages.

[19] logit(p)= ln(p/(1-p)), where p = probability and p/(1-p) is the corresponding odds

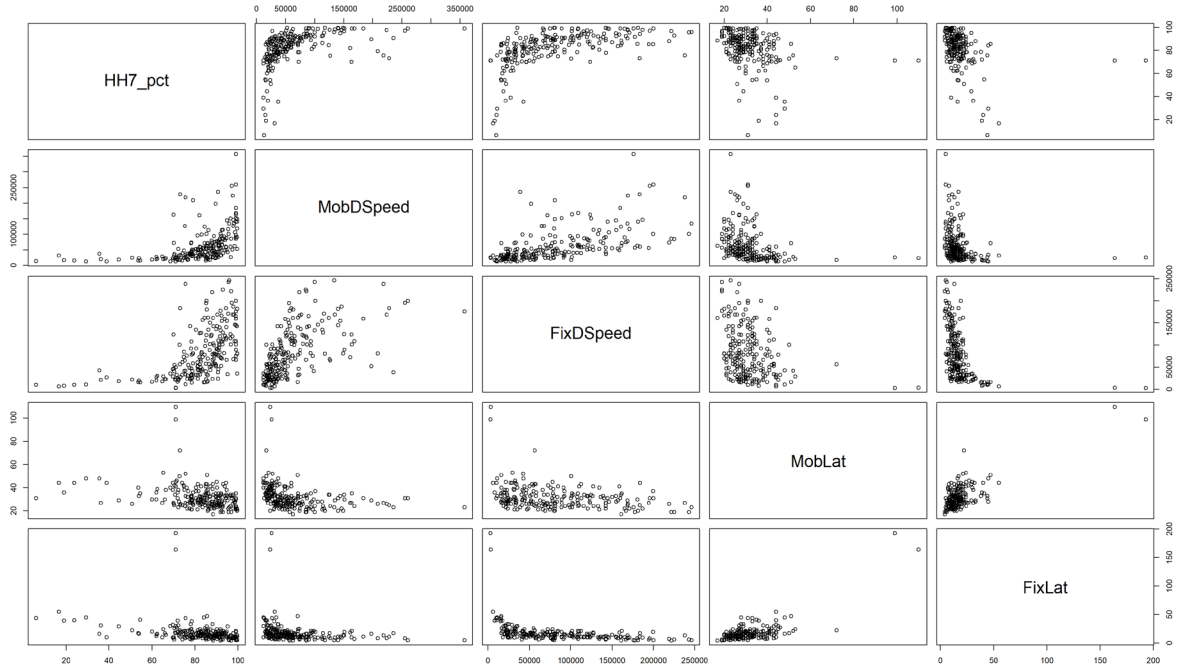
Figure 10 Pairwise relationships between proportion of individuals using the Internet and possible other socio-demographic explanatory variables, 2020-2022



Note: Relationships are shown where pairwise data are available for country-data year combinations. HH7_pct = Individuals using the Internet (percentage), i63GNI = Gross national income (GNI) per capita (current US dollars), ln_i63GNI = GNI per capita (log current US dollars), i6111 = Urban population (percentage), LF15 = Share of population 15+ in labour force, LF25 = Share of population 25+ in labour force, YrsSchooling = mean number of years of schooling, School_GR = gender ratio of mean number of years of schooling (female/male).

In addition, there is missingness for many of the explanatory variables. Most notably, of the 222 observations with data on individual Internet use, 60 are missing data on *Mean years of schooling* and 102 are missing data on *International bandwidth per Internet user*. On the estimation side, data availability as inputs for eventual models is also considered. The major limitation of mean years of schooling is the lack of data for 2022 – as a result there were 154 observations missing out of 370 country-year combinations requiring estimation. Even more limiting is *International bandwidth per Internet user* where 255 observations are missing of those requiring estimation. Somewhat less limiting, but still important to consider are missing data for *Fixed-broadband subscriptions per 100 inhabitants* (53 missing), *Active mobile-broadband subscriptions per 100 inhabitants* (missing 57), *Share of population 25+ in the labour force* (50 missing).

Figure 11 Pairwise relationships between proportion of individuals using the Internet and possible Internet speed explanatory variables, 2020-2022



Note: Relationships are shown where pairwise data are available for country-data year combinations. HH7_pct = Individuals using the Internet (percentage), MobDSpeed = Average mobile download speed (Kbps), FixDSpeed = Average fixed network download speed (Kbps), MobLat = Mobile latency (ms), FixLat = Fixed network latency (ms).

In addition, countries providing data on Internet use tend to have low prices relative to GNI. As a result many countries missing data on Internet use also have prices relative to GNI that are well out of scope of models. Price is then a less useful variable for estimating Internet use for such countries.

To compensate for the issues mentioned above, data sets excluding some or all these variables are also used for model comparison in addition to the full dataset. Other variables dropped are *Share of population 15+ in labour force* due to its strong correlation with *Share of population 25+ in labour force*

Modelling results

Variables are selected based on differing datasets and methods as shown in Table 10. The most consistent variable included in models was *percentage of urban population (i6111)*. The relationship is positive in all cases – consistent with the idea that urban areas tend to have higher rates of Internet use. Affordability measures also are key parameters for estimating Internet use. Interestingly, the *price of limited voice and data plans (as a percent of GNI)* are selected in more models than that of the *price of basic data plans (as a percent of GNI)*. Other variables are selected more frequently for logit models. This includes the *percentage of individuals aged 25+ in the labour force*, *mobile download speed* and *mobile latency*. *GNI per capita* or the *log of GNI per capita* are also often selected. Perhaps surprisingly *mobile-broadband subscriptions per 100 inhabitants* is selected in few models.

Table 10 Variables included in initial models selected using various variable selection methods

#	Transform	Exclusions	Weighted	n	DataYear	ln_i994u	i911mw	i992b	i63GNI	ln_i63GNI	i61	ln_i61	i911	i9111	i6111	AffordMobVoiceGNIpc	AffordMobVoicePPP	AffordMobDataGNIpc	AffordMobDataPPP	AffordMobBasicGNIpc	AffordMobBasicPPP	MedianAge	TotalFert	LifeExp	LifeExp_GR	LF25.	YrsSchooling	School_GR	FixDSpeed	MobileDSpeed	FixLat	MobLat
1			Y	87	x	x			x	x		x			x					x												
2				87	x	x						x			x					x	x				x	x	x					
3		a	Y	156	x		x	x	x		x		x																			
4		a		156							x				x	x	x										x	x	x	x		x
5		b	Y	207								x		x	x			x	x				x									
6		b		207						x					x	x	x															x
7		c	Y	160											x								x									
8		d	Y	211	x		x		x	x	x		x	x	x							x	x									x
9		e	Y	214			x		x	x					x							x	x	x								
15	Y		Y	87		x					x				x	x			x	x				x	x	x				x		x
16	Y			87		x					x				x				x	x	x				x	x						x
17	Y	a	Y	156				x		x		x			x	x											x			x		
18	Y	a		156				x		x	x	x			x	x								x	x	x				x		
19	Y	b	Y	207				x		x		x			x	x														x		
20	Y	b		207				x	x		x		x		x			x	x	x					x					x		x
21	Y	c	Y	160				x		x		x			x															x	x	
22	Y	d	Y	211				x		x	x				x								x							x	x	
23	Y	e	Y	214			x			x					x							x	x		x	x						

Note: a= Excluding i994u, b = Excluding i994u and mean years of schooling, c = Excluding i994u and affordability measures, d = Excluding i994u, mean years of schooling and affordability measures, e = Reduced model including only mobile subscriptions, GNI per capita, share of pop in labour force, life expectancy, gender ratio of life expectancy, total fertility, median age, and share of urban population. n = number of observations on which the model is based. DataYear = year of data, ln_i994u = log of international bandwidth per Internet user, i911mw = Active mobile-broadband subscriptions per 100 inhabitants, i992b = Fixed-broadband subscriptions per 100 inhabitants, i63GNI = Gross national income (GNI) per capita (current US dollars), ln_i63GNI = GNI per capita (log current US dollars), i61 = population, ln_i61 = log population, i911 = Mobile-cellular subscriptions per 100 inhabitants, i9111 = Total telephone subscriptions per 100 inhabitants, i6111 = Urban population (percentage), AffordMobVoice = price of a voice-only mobile package as share of GNI, AffordMobData = price of data only mobile package as share of GNI (2 GB), AffordMobDataPPP = price of data only mobile package at purchasing price parity (PPP) (2 GB), AffordMobBasicGNIpc = price of a limited voice mobile package as share of GNI (70 min + 20 SMS + 500 MB), AffordMobBasicPPP = price of a limited voice mobile package at purchasing price parity (PPP), MedianAge = median age of population, TotalFert = total fertility rate (average number of expected for women given current fertility rates), LifeExp = life expectancy in years, LifeExp_GR = gender ratio of life expectancy (female/male). LF25 = Share of population 25+ in labour force, YrsSchooling = mean number of years of schooling, School_GR = gender ratio of mean number of years of schooling (female/male). MobDSpeed = Average mobile download speed (Kbps), FixDSpeed = Average fixed network download speed (Kbps), MobLat = Mobile latency (ms), FixLat = Fixed network latency (ms).

When regressed against full data sets, none of the above models demonstrate clear evidence of non-normality and all are retained.

Certain models were preferred based on the characteristics of the country. For example, non-transformed linear models were not preferred for countries with likely very high or very low level of Internet use. Unweighted models are not preferred for poorer countries. Such countries are not well-represented in these models as more weight is implicitly given to wealthier countries that have provided more data. Models excluding price are preferred for countries with high prices relative to GNI as the input to such models for these countries would be out of scope.

Estimation

The overall strategy for model selection for use in estimation is to align with previous official data submitted and to use primary models as possible. Models of the log odds (logit) of percentage of individuals using the Internet are preferred for countries with likely low or high values. Many countries are missing data for explanatory variables used in some models. In these cases, the choice of models is limited.

Sixty-six countries are estimated directly without complication. These estimations are all consistent with previously provided official data or previous estimations. Choice of model is split between models of the log odds of individual Internet use (39 countries) and linear (27 countries).

For 26 additional countries models are used partially or indirectly. In 12 of these countries data required for modelling are not available in all years or modelled results are inconsistent for some years. Modelled estimations are used in years with available data and imputed growth or analogies with the growth of similar countries are used in other years. In 17 other countries, estimations do not align with available official data – in these cases the growth in modelled data is used to project or backcast estimations based on these official data. In two countries a combination of modelling and ratio of individual Internet use to available official data on household Internet access is used for estimation.

Among these countries requiring the use of models indirectly are two cases where official data are not used in favor of alternative methods. In one case, official data have been provided frequently in recent years, but show a very unlikely decrease in Internet use. As such a trend is unlikely given technological trends for other countries in the region these data are overwritten and replaced with modelled data. In the second case a very low Internet use share is reported. Compared to other countries in the region this does not appear to be a reasonable estimate. As such, modelled estimates are used in place of these data after consulting with countries.

Other methods

For some smaller countries analogies with neighbors are used. Andorra estimates were set equal to Spain, San Marino and Vatican estimates were set equal to Italy and Liechtenstein estimates are set equal to Switzerland.

Regional averages as well as averages for groups such as LDCs, land-locked developing countries (LLDCs) and small island developing states (SIDS) are also considered. However, this was not a satisfactory solution, particularly for countries in groups where there are few available data in the first place.

Remaining countries

There are 10 countries for which estimates are calculated, but not recommended for publication at the individual country-level. These estimates can be used to calculate aggregates for the *Facts and Figures* publication.

Share of individuals who own a mobile phone

Overview

The percentage of individuals who own a mobile phone are estimated for all ITU Member States[20] for 2020-2022. While some countries have official data available for all years, many do not and require estimation by (1) simple imputation methods such as extrapolation, (2) regression modelling, or (3) using analogies with similar countries. Table 11 summarizes the sources for data on the percentage of individuals who own a mobile phone.

Table 11 Source for data on the percentage of individuals who own a mobile phone for ITU Member States, by region (provisional)

Source	All	Africa	Americas	Arab States	Asia & Pacific	CIS	Europe
Official	**22**	0	4	2	7	3	6
NSO estimate	**1**	1	0	0	0	0	0
Imputation	**18**	1	1	7	3	3	3
Model	**135**	35	27	11	28	3	31
Analogy	**5**	0	1	0	0	0	4
Other source	**3**	3	0	0	0	0	0
Share of total	***93.4***	***90.9***	***94.3***	***90.9***	***92.7***	***100.0***	***95.7***

Overall, data for 184 countries are published using official data, simple imputation, national statistics office (NSO) estimates, modelling results, analogies with nearby countries, or other sources. These countries represented close to 100 per cent of global gross national income (GNI) and over 95 per cent of the world population in 2022. There are 13 countries where estimations were not feasible[21].

Table 12 Source for data on the percentage of individuals who own a mobile phone for ITU Member States, by share of gross national income (GNI) and population (provisional)

Source	All	Share of 2022 GNI	Share of 2022 Population
Official	**22**	13.6	15.6
NSO estimate	**1**	0.0	0.0
Imputation	**18**	9.0	7.0
Model	**135**	68.8	66.4
Analogy	**5**	7.6	3.4
Other source	**3**	0.5	3.7
Share of total	***93.4***	***99.5***	***96.1***

[20] ITU member states as well as (a) Hong Kong, China, (b) Macau, China, (c) Taiwan, Province of China, and (d) Palestine

[21] Afghanistan, Central African Republic, Comoros, Croatia, Democratic People's Republic of Korea, Haiti, Lesotho, Myanmar, Niger, Syrian Arab Republic, Uganda, Ukraine, Venezuela

Introduction

The percentage of individuals who own a mobile phone is a core indictor used by the ITU to monitor connectivity. This indicator is used as a main reference point in publications such as ITU's annual *Measuring digital development: Facts and figures*. This indicator is normally derived from household surveys, which many countries do not administer, considering the cost and complexity. As a result, many countries do not report data. In addition, when countries submit data for this indicator, it can be for several years beforehand, considering the time it takes to process household surveys. This lag is important in the context of ICT statistics, because of the rapid pace of change in this domain compared to other social statistics derived from surveys.

In this context, estimations are needed to provide a fuller and timelier assessment. This section documents the methodology used to estimate missing data for the years 2020-2022.

Data availability

Questionnaire data are extracted from the database for the last three years (2020, 2021 and 2022) for all ITU Member States[22] using an SQL query. The base indicator used is *HHU271own_HHTotalIndividual*. Shares are calculated by dividing by the individual target population (*HHU_HHTotalIndividual*). Data are complete for 22 countries and missing for at least one year in 176 countries.

Table 13 Individual mobile phone ownership data availability for ITU Member States, 2020-2022

	Data years			Most recent data year	All	LDC
	2020	*2021*	*2022*			
All years	Yes	Yes	Yes	2022	22	0
Missing at least one year	-	-	-	-	176	46
One year	-	-	-	-	*17*	*0*
One year	No	Yes	Yes	2022	2	0
One year	Yes	No	Yes	2022	3	0
One year	Yes	Yes	No	2021	12	0
Two years	-	-	-	-	*26*	*5*
Two years	No	No	Yes	2022	5	1
Two years	No	Yes	No	2021	9	2
Two years	Yes	No	No	2020	12	2
All years	*No*	*No*	*No*	-	*133*	*41*
All years	No	No	No	2019	12	2
All years	No	No	No	2018	12	3
All years	No	No	No	2017	16	6
All years	No	No	No	2016	20	6
All years	No	No	No	2015	13	1
All years	No	No	No	2014	11	2

[22] Including Hong Kong, China; Macau, China; Palestine; and Taiwan, Province of China..

	Data years			Most recent data year	All	LDC
	2020	*2021*	*2022*			
All years	No	No	No	< 2014	49	21

Note: Data availability as of 6 July 2023. LDC = Least Developed Country. Includes Palestine; Hong Kong, China; Macau, China; and Taiwan, Province of China

Data are sparser for Least Developed Countries (LDCs) and countries in Africa, Asia & Pacific and the Americas. However, each of these groups do have some data availability and give representation to these groups of countries in modelling.

Table 14 Individual mobile phone ownership data availability for ITU Member States, by region 2020-2022

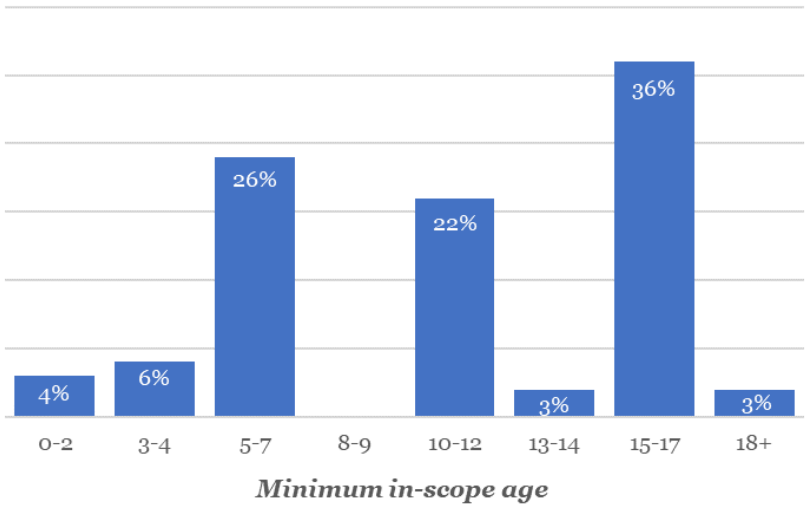
	All	Africa	Arab States	Asia & Pacific	CIS	Europe	Americas
All years	22	0	2	7	3	6	4
Missing at least one year	176	44	20	34	6	40	32
One year	*17*	*0*	*4*	*3*	*2*	*5*	*2*
Two years	*26*	*6*	*4*	*5*	*1*	*6*	*4*
All years	*133*	*38*	*12*	*26*	*3*	*29*	*26*

Note: Includes Palestine; Hong Kong, China; Macau, China; and Taiwan, Province of China

For modelling, data from 2020-2022 are used. *HHU2710wn_HHTotalIndividual* are complemented by official data received through other sources or provided only in percentages for several additional countries – available in the indicator *xHH18_IDI*. Data are available from countries of varied regions and development levels. However, more data are available from higher income countries.

A challenge in comparing data on mobile phone ownership is the varying in-scope ages for data provided by countries. While for Internet use it is a feasible (though unlikely) assumption that the share of users may not vary enough to strongly affect comparability, mobile phone ownership cannot be assumed for young children – ownership among young children is likely to be close to zero. Consequently, the variation in minimum in-scope age among countries providing data (Figure 12) creates difficulties in maintaining comparability.

Figure 12 Proportion of countries with various minimum in-scope ages of those providing data on mobile phone ownership, 2019 or later



Note: Data as of June 2022

To address this issue, it is assumed that modelling results are the share of individuals aged 10+. This is the median minimum age of those providing data – the best simple approximation available. Aggregates are also calculated on the basis of the 10+ population rather than the full population as done for Internet use estimates.

Methodology

Methodologies differ for countries missing data based on their availability of data in other years. For the 22 countries with data already available, data are used as available in surveys and no estimations are performed. In one additional country estimates from countries' national statistics offices are available and used and 3 other countries results from non-governmental surveys are used as estimates. Lastly, in four countries with high (or likely high) mobile phone ownership, an analogy was made with official data available on mobile phone use. In these countries it is assumed that mobile phone use is similar to ownership.

Imputation

In cases where data are available in some but not all years and available data are reasonably consistent over time, imputation methods such as interpolation or extrapolation are used. These methods are used for 18 countries. Where data are available in previous years, data are extrapolated using the weighted average growth over the previous 2 or 3 years[23]. Where there are gaps in data, the geometric mean is used to fill these gaps. For countries with already high levels of mobile phone ownership, the most recent year data are used to fill gaps. There are also several special cases where imputation was possible and do not fall under these categories – these are cases where there were gaps in previous year data.

[23] For 3 years, growth in the previous year weighted at 0.5 (one half), 2 years previous at 0.33 (one third) and 3 years previous at 0.17 (one sixth). For 2 years, growth in previous year weighted at 0.6 and 2 years previous at 0.4. Some countries are also estimated using one previous year of stable growth due to a break in series or unavailable data. Countries with high growth are verified against modelled results.

Resulting estimations are assessed for consistency with previous year data and in the case of decreases or large increases consistency with underlying conditions in the country.

Modelling

For the remaining 149 countries where data are fully unavailable or inconsistent, estimation is attempted through regression modelling[24]. The following variables are identified as potentially explanatory either extracted from the *delhi* SQL database or sourced from databases from other international organizations or private companies. In the case of Ookla data, these are extracted and averaged over all tests performed in a country. Data were excluded where tests were performed on fewer than 100 devices:

- Active mobile-broadband subscriptions per 100 inhabitants (i911mw)
- Fixed-broadband subscriptions per 100 inhabitants (i992b)
- GNI per capita (current US dollars) – the log of GNI also used given the skewed distribution among countries. Source is UN Statistics Division data as ITU database not fully updated. Data only available until 2021 - IMF GDP per capita growth rates (current USD) to project data for 2022
- Population (i61) (source – UN Population Division) – the log of population also given the skewed distribution among countries
- International bandwidth per Internet user (bit/s) (i994u) – the log of bandwidth also given the skewed distribution among countries
- Mobile-cellular subscriptions per 100 inhabitants (i911) and Total telephone subscriptions per 100 inhabitants (i9111)
- Urban population (percentage) (source – World Bank) – ITU data not fully updated
- Affordability (price of 2 GB of mobile data as a percentage of GNI and in PPP)
- Affordability (price of limited voice and data plan as a percentage of GNI and in PPP)
- Affordability (price of voice-only plan as a percentage of GNI and in PPP)
- Labour force participation (15+ and 25+) (source – ILO)
- Median age of population (source – UN Population Division)
- Life expectancy (source – UN Population Division)
- Mean years of schooling (source – UNESCO)
- Mobile and fixed download speed (source – Ookla)
- Mobile and fixed latency (source – Ookla)
- Percentage of individuals using the Internet (HHU4212_HHTotalIndividual – calculated as share of total target pop)

Additional variables have been considered but not used for modelling purposes due to poor data availability for countries requiring estimation[25]:

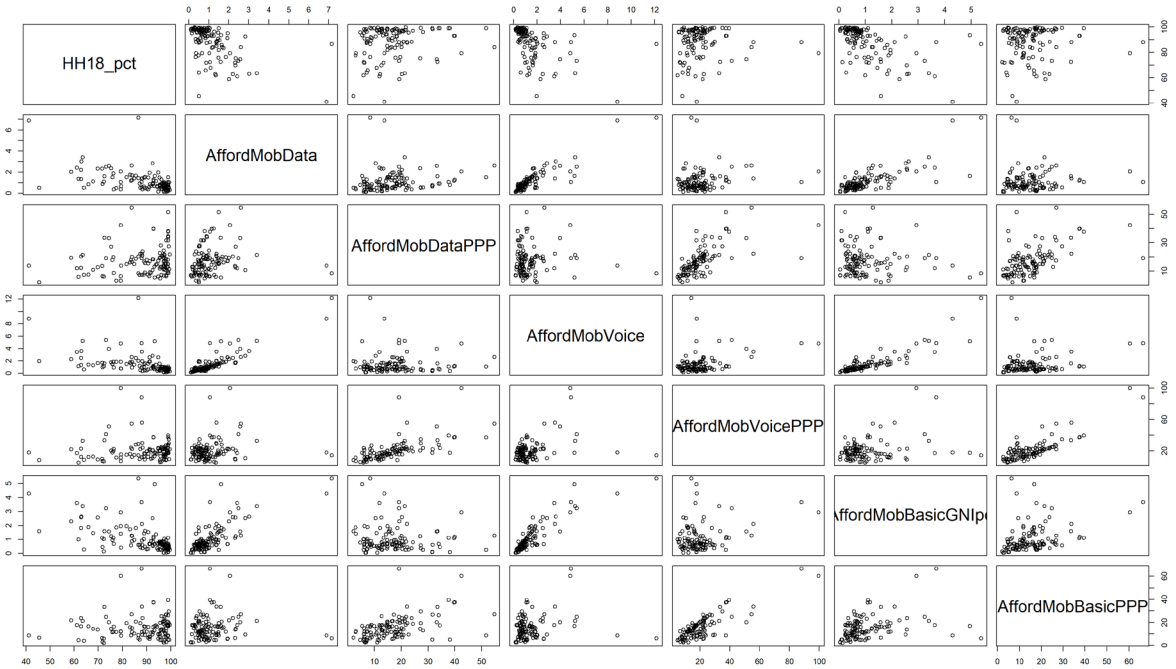
- Percentage of households with Internet access (HHA4213_HHTotalHousehold– calculated as share of total target pop)
- Affordability (price of basic device as a percentage of GNI and in USD)
- Government funding per student (source – UNESCO)
- Literacy rate (source – UNESCO)
- Government expenditure on education (source – UNESCO)

[24] Modelling conducted using R.
[25] Data on number of cell towers per population are not used as an analysis conducted in 2022 revealed that it was a poor driver for similar estimates.

Model preparation

To maximize the dataset and balance against relevance of relationships with explanatory variables, data from 2020-2022 are used. This results in an overall dataset of 124 observations[26] from 65 countries of which five are Least Developed Countries (LDCs). Pairwise analysis of the data in the figures below shows a relatively strong negative relationship with *Affordability* (measured as percent of GNI) and a relatively strong positive relationship with *Active mobile-broadband subscriptions per 100 inhabitants, Fixed-broadband subscriptions per 100 inhabitants,* the log of *International bandwidth per Internet user*. Among non-ICT variables, *life expectancy*, *median age*, *mean years of schooling*, and the log of *GNI per capita (current US dollars)* also appear to have a strong association with *Percentage of individuals who own a mobile phone.*

Figure 13 Pairwise relationships between proportion of individuals who own a mobile phone and affordability measures, 2020-2022



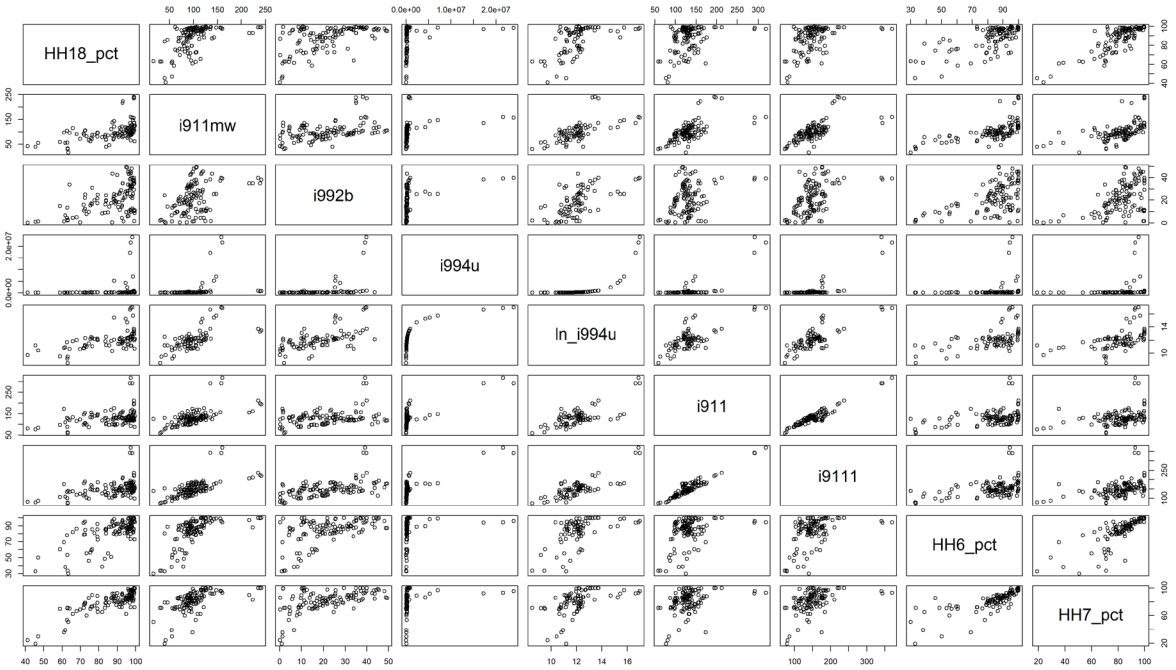
Note: Relationships are shown where pairwise data are available for country-data year combinations. HH18_pct = Individuals who own a mobile phone (percentage), AffordMobData = price of data only mobile package as share of GNI (2 GB), AffordMobDataPPP = price of data only mobile package at purchasing price parity (PPP) (2 GB), AffordMobVoice = price of a voice-only mobile package as share of GNI, AffordMobVoicePPP = price of a voice-only mobile package at purchasing price parity (PPP), AffordMobBasicGNIpc = price of a limited voice mobile package as share of GNI (70 min + 20 SMS + 500 MB), AffordMobBasicPPP = price of a limited voice mobile package at purchasing price parity (PPP)

As in the 2022 estimation, cross validation[27] is used. This method provides an alternative which captures the benefits of using the full data set and avoided overfitting. *K-fold* cross-validation is used (k = 10) as an alternative method to identify important explanatory variables. Specifically, this entails dividing the dataset into 10 "folds", running 10 separate regressions with each of the folds removed and testing on the removed fold. The best

[26] up to 3 per country for different years. Data as of 7 July 2023
[27] https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation

performing group of variables was selected based on the lowest root mean square error (RMSE). Traditional stepwise regression was used in previous years. This has been discontinued due to the improved concepts behind cross validation.

Figure 14 Pairwise relationships between proportion of individuals who own a mobile phone and ITU ICT indicators, 2020-2022



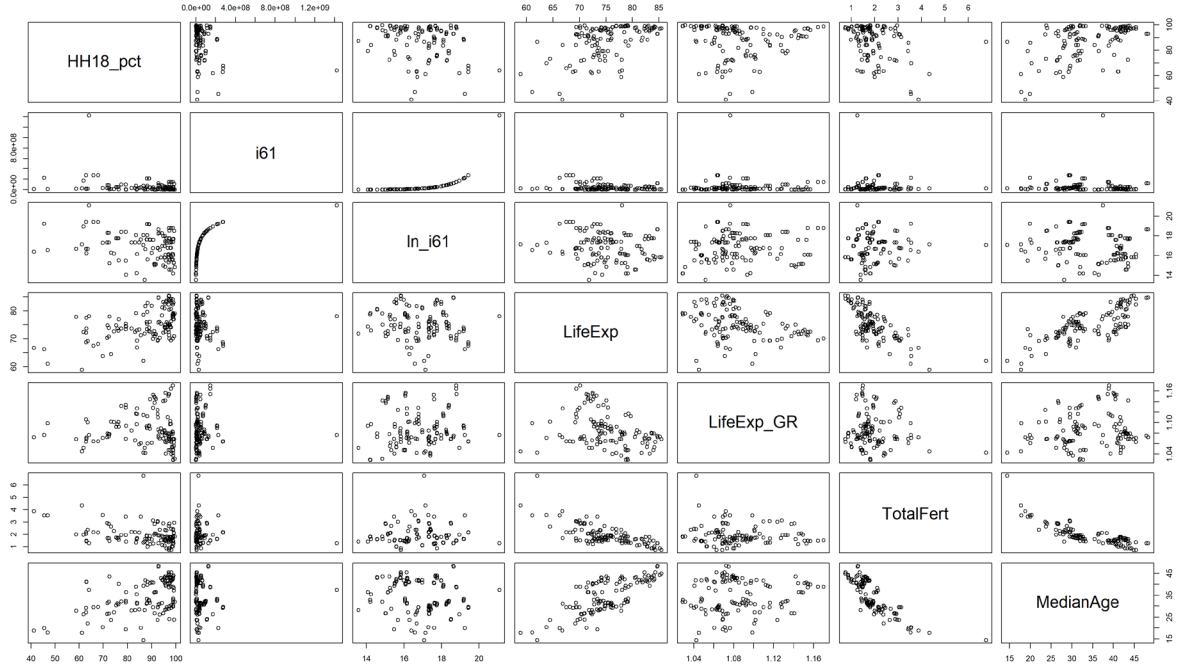
Note: Relationships are shown where pairwise data are available for country-data year combinations. HH18_pct = Share of individuals who own a mobile phone, i911mw = Active mobile-broadband subscriptions per 100 inhabitants, i992b = Fixed-broadband subscriptions per 100 inhabitants, i99H = Internet users (percentage), i994u = International bandwidth per Internet user (bit/s), ln_i994u = International bandwidth per Internet user (log bit/s), i9111 = Total telephone subscriptions per 100 inhabitants, HH7_pct = Individuals using the Internet (percentage), HH6_pct = Share of household with access to the Internet.

Data points are also weighted to avoid overweighting wealthier countries which provide data to the ITU more frequently. The weights are calculated in multiple steps. First, the regional weights are calculated based on the number of countries in each region – this assumes equal weight for all countries regardless of population or economic output. As an example, for the Arab States there are 22 countries or a 11.2 per cent share of the 197 total ITU members and additional economies. This is done in accordance with the broad *one country one vote* principle in the UN system. Second, these regional weights are divided equally among countries reporting data in the region. For the Arab States 11 countries reported data during the 2020-2022 period so each receives 11.2 / 9 = 1.1 weight. Where countries provided data for more than one year, the weight for each data point is divided by the number of data points provided by the country. United Arab Emirates, for example, provided three years of data and its weight then is 1.1 / 3 = 0.4. This results in a range of weights from 0.3 for one data point from CIS countries providing three years of data (0.6 full weight for the country) to 3.7 for African countries providing only one year of data.

Two different methods are used to estimate the proportion of individuals who own a mobile phone. In the first, the proportion is estimated directly through ordinary linear regression (OLS). While the simplest method, estimating in this way allows for the possibility of

negative proportions or proportions over 100 per cent. To avoid this problem a logit (log odds) transformation[28] of the share of individuals who own a mobile phone is also used as the dependent variable in OLS. This is particularly useful in estimating countries with very low or high levels of mobile phone ownership.

Figure 15 Pairwise relationships between proportion of individuals who own a mobile phone and possible socio-economic explanatory variables from the UN Population Division, 2020-2022

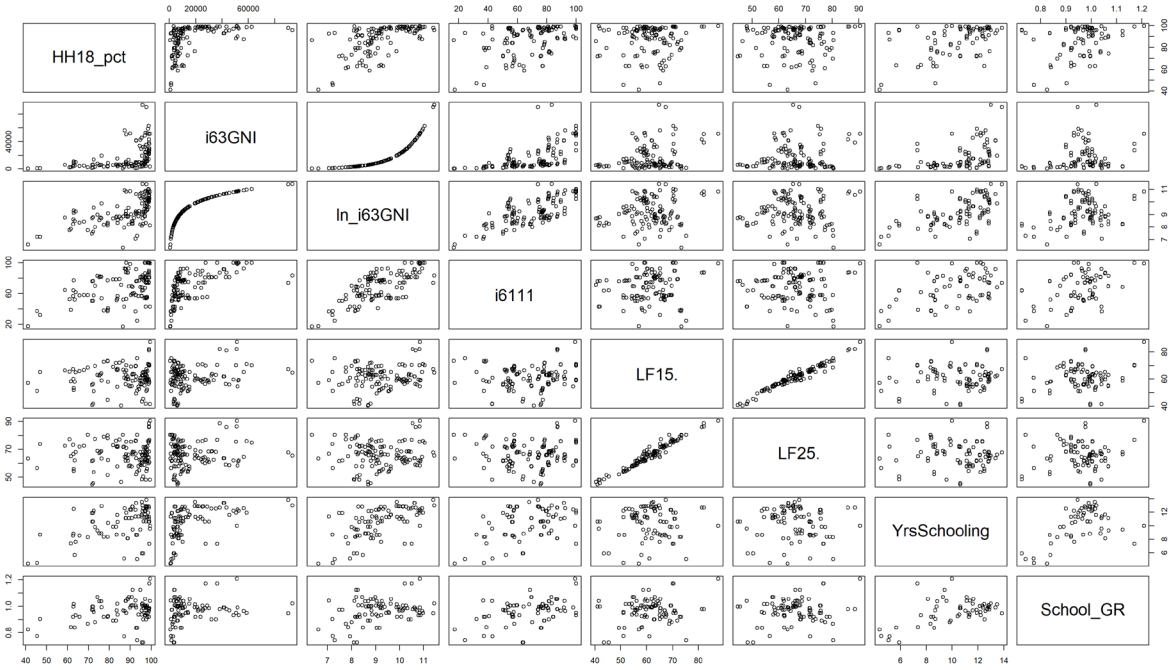


Note: Relationships are shown where pairwise data are available for country-data year combinations. HH18_pct = Individuals who own a mobile phone (percentage), i61 = population, ln_i61 = log population, LifeExp = life expectancy in years, LifeExp_GR = gender ratio of life expectancy (female/male), TotalFert = total fertility rate (average number of expected for women given current fertility rates), MedianAge = median age of population.

Comparing models poses a challenge for multiple reasons. First, different errors are minimized in these two different methods. The logit transformation does not directly minimize the error between predicted and actual percentage of individuals who own a mobile phone as it minimizes the error between the predicted and actual *log odds* of the percentage of individuals who own a mobile phone. Models estimating the variable in question directly are preferred but are analysed on a case-by-case basis to avoid impossible results outside of the 0 to 100 range of percentages.

[28] logit(p)= ln(p/(1-p)), where p = probability and p/(1-p) is the corresponding odds

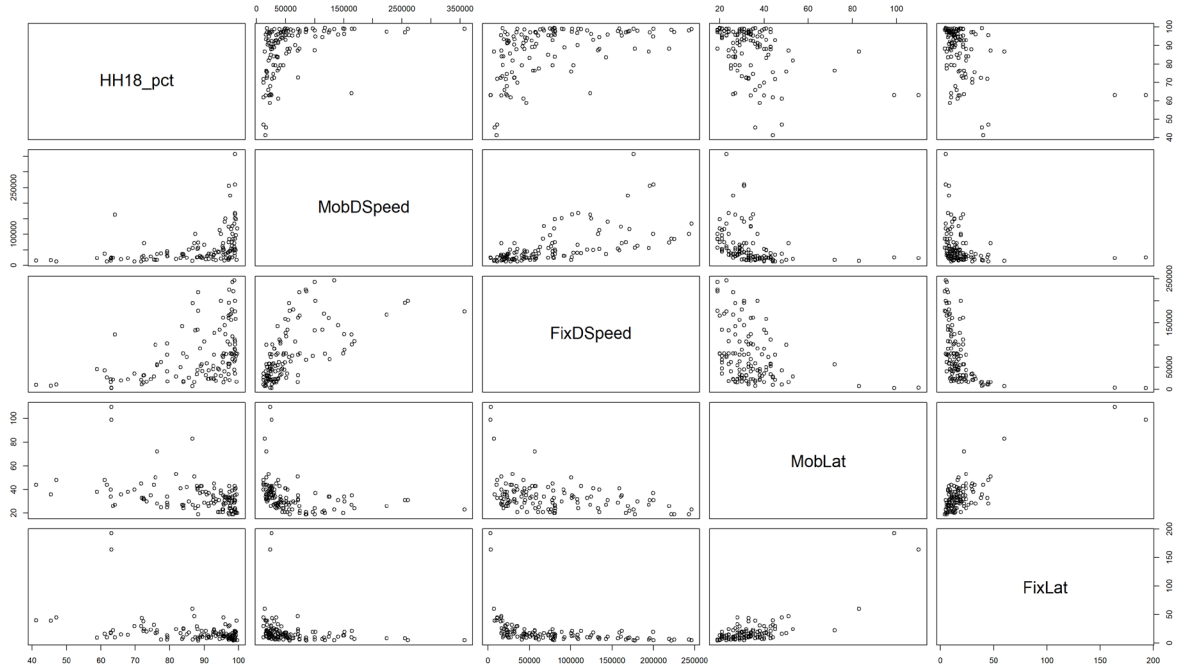
Figure 16 Pairwise relationships between proportion of individuals who own a mobile phone and possible other socio-demographic explanatory variables, 2020-2022



Note: Relationships are shown where pairwise data are available for country-data year combinations. HH18_pct = Individuals who own a mobile phone (percentage), i63GNI = Gross national income (GNI) per capita (current US dollars), ln_i63GNI = GNI per capita (log current US dollars), i6111 = Urban population (percentage), LF15 = Share of population 15+ in labour force, LF25 = Share of population 25+ in labour force, YrsSchooling = mean number of years of schooling, School_GR = gender ratio of mean number of years of schooling (female/male).

In addition, there is missingness for many of the explanatory variables. Most notably, of the 124 observations with data on individual mobile phone ownership, 31 are missing data on *Mean years of schooling* and 39 are missing data on *International bandwidth per Internet user*. On the estimation side, data availability as inputs for eventual models is also considered. The major limitation of mean years of schooling is the lack of data for 2022 – as a result there are 183 observations missing out of 468 country-year combinations requiring estimation. *International bandwidth per Internet user* is a more extreme case where 318 of those requiring estimation are missing data. Somewhat less limiting, but still important to consider were missing data for *Fixed-broadband subscriptions per 100 inhabitants* (53 missing), *Active mobile-broadband subscriptions per 100 inhabitants* (missing 58), *Share of population 25+ in the labour force* (51 missing).

Figure 17 Pairwise relationships between proportion of individuals who own a mobile phone and possible Internet speed explanatory variables, 2020-2022



Note: Relationships are shown where pairwise data are available for country-data year combinations. HH18_pct = Individuals who own a mobile phone (percentage), MobDSpeed = Average mobile download speed (Kbps), FixDSpeed = Average fixed network download speed (Kbps), MobLat = Mobile latency (ms), FixLat = Fixed network latency (ms).

In addition, countries providing data on mobile phone ownership tend to have low prices relative to GNI. As a result many countries missing data on mobile phone ownership also have prices relative to GNI that are well out of scope of models. Price is then a less useful variable for estimating mobile phone ownership for such countries.

To compensate for the issues mentioned above, data sets excluding some or all these variables are also used for model comparison in addition to the full dataset. Other variables dropped were *Share of population 15+ in labour force* due to its strong correlation with *Share of population 25+ in labour force*

Modelling results

Variables are selected based on differing datasets and methods as shown in Table 15. *Percentage of individuals using the Internet* was a dominant parameter in models – serving as the sole parameter in several models. For this reason, several models are also tried excluding Internet use to assess the importance of other variables. In such models, affordability measures, *mobile-broadband subscriptions per 100 residents (i911mw)* and *GNI per capita* are often significant variables. Somewhat surprisingly, socio-economic variables such as *urban share of population*, *median age*, *total fertility*, *life expectancy* and *labour force share* are selected in very few usable models.

Table 15 Variables included in initial models selected using various variable selection methods

#	Transform	Exclusions	Weighted	n	i911mw	i992b	i63GNI	ln_i63GNI	i61	AffordMobData	AffordMobBasic	TotalFert	LifeExp	LifeExp_GR	LF25.	MobileDSpeed	FixLat	MobLat	HH7_pct
1			Y	60					x		x								x
2				60					x		x								x
3		a	Y	86					x		x								x
4		a		86															x
5		b	Y	110															x
6		b		110					x		x								x
7		c	Y	90													x		x
8		d	Y	114	x				x						x		x		x
9		e	Y	111	x				x	x			x	x					
10		f	Y	115			x	x	x									x	
15	Y		Y	60															x
16	Y			60															x
17	Y	a	Y	86															x
18	Y	a		86							x								x
19	Y	b	Y	110															x
20	Y	b		110					x		x					x			x
21	Y	c	Y	90															x
22	Y	d	Y	114	x														x
23	Y	e	Y	111		x		x	x	x		x	x	x		x			
24	Y	f	Y	115	x		x	x	x									x	

Note: a= Excluding i994u, b = Excluding i994u and mean years of schooling, c = Excluding i994u and affordability measures, d = Excluding i994u, mean years of schooling and affordability measures, e = Excluding HH7_pct, i994u and mean years of schooling, f = Excluding HH7_pct, affordability, i994u and mean years of schooling. n = number of observations on which the model is based. i911mw = Active mobile-broadband subscriptions per 100 inhabitants, i992b = Fixed-broadband subscriptions per 100 inhabitants, i63GNI = Gross national income (GNI) per capita (current US dollars), ln_i63GNI = GNI per capita (log current US dollars), i61 = population, AffordMobData = price of data only mobile package as share of GNI (2 GB), AffordMobBasicGNIpc = price of a limited voice mobile package as share of GNI (70 min + 20 SMS + 500 MB), TotalFert = total fertility rate (average number of expected for women given current fertility rates), LifeExp = life expectancy in years, LifeExp_GR = gender ratio of life expectancy (female/male). LF25 = Share of population 25+ in labour force, MobDSpeed = Average mobile download speed (Kbps), MobLat = Mobile latency (ms), FixLat = Fixed network latency (ms), HH7_pct = Share of individuals using the Internet.

When regressed against full data sets, none of the above models demonstrate clear evidence of non-normality and all are retained.

Certain models are preferred based on the characteristics of the country. For example, non-transformed linear models are not preferred for countries with likely very high or very low level of Internet use. Unweighted models are not preferred for poorer countries. Such

countries are not well-represented in these models as more weight is implicitly given to wealthier countries that have provided more data. Models excluding price are preferred for countries with high prices relative to GNI as the input to such models for these countries are out of scope.

Estimation

The overall strategy for model selection for use in estimation is to align with previous official data submitted and to use primary models as possible. Models of the log odds (logit) of percentage of individuals who own a mobile phone are preferred for countries with likely low or high values. Many countries are missing data for explanatory variables used in some models. In these cases, the choice of models is limited.

116 countries are estimated directly without complication. These estimations are all consistent with previously provided official data or previous estimations. Choice of model is split between models of the log odds of individual mobile phone ownership (73 countries) and linear models (43 countries).

For 19 additional countries models are used partially or indirectly. In one country data required for modelling are not available in all years. Modelled estimations are used in years with available data and an analogy with the growth of a similar country is used in other years. In 18 other countries, estimations do not align with available official data – in these cases the growth in modelled data is used to project or backcast estimations based on these official data.

Other methods

For one smaller country, an analogy with a neighbor is used. Specifically, Vatican estimates are set equal to Italy.

Regional averages as well as averages for groups such as LDCs, land-locked developing countries (LLDCs) and small island developing states (SIDS) are also considered. However, this is not a satisfactory solution, particularly for countries in groups where there are few available data.

Remaining countries

For the 13 remaining countries estimations are calculated, but not recommended for publication at the individual country-level. These estimates can be used to calculate aggregates for the Facts and Figures publication.

Percentage of the population covered by mobile networks (at least 3G, at least LTE/4G)

Data on *percentage of the population covered by mobile networks (at least 3G, at least LTE/4G)* are one of the core ICT indicators collected annually through the ITU World Telecommunication/ICT Indicators questionnaires. Administrative sources (mobile network operators) provide these data, which are submitted to ITU by regulatory bodies and ministries. The majority of the ITU Member States[29] report data for this indicator (see Table 16).

Table 16 ITU Member States not submitting data for *Percentage of the population coverage by mobile networks*

	Data year		
	2020	2021	2022
Member States that **did not submit** data for:			
At least 3G (count/share*)	17/9%	22/11%	48/24%
At least LTE/4G (count/share)	16/8%	21/11%	51/26%

Note: Data availability as of 29 September 2023.
* out of 196. Includes Palestine; Hong Kong, China; and Macau, China.

To fill the data gaps for countries that do not submit data, two methods are employed: 1) estimation using published data; and – if the first method is not successful – 2) estimation using trends.

Estimation using published data

Data on coverage are sometimes available in the report and/or on the website of regulators and/or operators. Based on this information, it may be possible to estimate network coverage for the population of a country. This process involves the following steps:

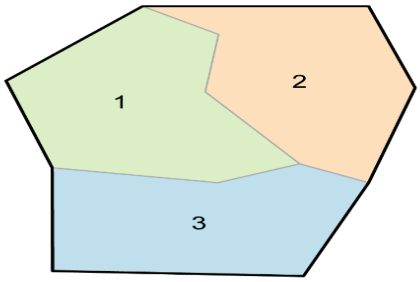
1. **Identifying market players**: It is necessary to determine the number of, and obtain information on mobile network operators and primary market providers of the service in each country. This is achieved by checking regulatory information from the country's regulator's website, the ITU's regulatory database, and third-party sources such as data provided by GSMA Intelligence, which offers valuable insights into the operators within each country.
2. **Annual report analysis**: Once the MNOs are identified, it is necessary to research and download their annual reports from operator's website. It is crucial to scrutinize and compile data from the annual reports meticulously since operators may report on operations outside of the country (or across multiple countries for foreign operators). In cases where one operator is owned by another significant or foreign operator, it is necessary to determine the market share specifically for the country where data is being estimated.

[29] Including Hong Kong, China; Macau, China; and Palestine.

3. **Alternative sources of annual reports**: If annual reports are unavailable on operators' websites, annual reports from national stock exchange commissions or international exchange commissions (e.g., the Security and Exchange Commission) can be consulted. These reports, often in the form of Form 20-F and Form 6-F filings, provide comprehensive information about the company, including subscriptions data, tariff details, staffing information, financial data, and more.
4. **Press releases and other media reports**: If data are unavailable from company annual reports or regulators' websites, the data of an operator can be approximated through industry analysis and news articles provided by various sources. These sources may include operator press releases, official statements from regulators or ministries, and reputable newspapers within the country. Such data sources may include absolute subscription numbers, market shares, penetration rates, and growth rates, population coverage, which can be used to derive estimates for the country.

Once the data from all operators are available, the total percentage of the population covered can be calculated as illustrated in Figure 18.

Figure 18 Example of aggregation with three different regions and operator's population coverage



	Coverage (% of population)		
	Operator 1	Operator 2	Operator 3
Region 1 (25% pop)	70%	**80%**	70%
Region 2 (25% pop)	50%	60%	**70%**
Region 3 (50% pop)	**80%**	70%	70%

- **Country coverage:** 80% * 25% + 70% * 25% + 80% * 50% = **77.5%**

Estimation using trends

When data are not available from annual nor industry reports, data are estimated by analysing trends from the previous five years. This estimation process is facilitated through the utilization of the Expert Modeller forecasting functionality in SPSS, employing the following syntax:

```
* Time Series Modeler.
TSMODEL
   /MODELSUMMARY  PRINT=[MODELFIT]
   /MODELSTATISTICS  DISPLAY=YES MODELFIT=[ SRSQUARE
RSQUARE MAE]
   /SERIESPLOT OBSERVED FORECAST
   /OUTPUTFILTER DISPLAY=ALLMODELS
   /SAVE  PREDICTED(Predicted)
```

```
/AUXILIARY  CILEVEL=95 MAXACFLAGS=24
/MISSING USERMISSING=EXCLUDE
/MODEL DEPENDENT= "countries to be modelled"
   PREFIX='Model'
/EXPERTMODELER TYPE=[ARIMA EXSMOOTH]
/AUTOOUTLIER  DETECT=OFF.
```

The Expert Modeller determines the best-fitting model for a country, either ARIMA or exponential smoothing model. This functionality automatically tests the data for seasonality, intermittency, and missing values, and selects appropriate model.

Active mobile-broadband subscriptions

Active mobile-broadband subscriptions is one of the core ICT indicators collected and disseminated by the ITU. Data are gathered annually through the ITU World Telecommunication/ICT Indicators questionnaires. Administrative sources (mobile network operators) provide this data, which are compiled and submitted to ITU by regulators and ministries. The majority of Member States[30] (Table 17) submit data for this indicator.

Table 17 ITU Member States not submitting data for *Active mobile-broadband subscriptions*

	Data year		
	2020	2021	2022
Countries that **did not submit** data (count/share)	17/9%	21/11%	42/21%

Note: Data availability as of 29 September 2023. Includes Palestine; Hong Kong, China; and Macau, China.

The following methods are employed to address the data gaps for countries that do not submit data:

Estimation based on published data:

Administrative data sources are sometimes accessible on the websites and/or in reports of regulators and/or operators . Missing data can be approximated by compiling published data or estimating from data reported by operators. This compilation process requires the following steps:

1. **Identifying operators:** To approximate market size, it is necessary to determine the number of, and obtain information on mobile network operators (MNOs) and primary market providers of the service in each country. This is achieved by checking regulatory information from the country's regulator's website, the ITU's regulatory database, and third-party sources such as data provided by GSMA Intelligence, which offers valuable insights into the operators within each country.
2. **Annual report analysis:** Once the MNOs are identified, it is necessary to research and download their annual reports from their website. It is crucial to scrutinize and compile data from the annual reports meticulously since operators may report on operations outside of the country (or across multiple countries for foreign operators). In cases where an MNO is owned by another significant or foreign operator, it is necessary to determine the market share specifically for the country where data is being estimated.
3. **Alternative sources of annual reports:** If annual reports are unavailable on operators' websites, annual reports from national stock exchange commissions or international exchange commissions (e.g., the Security and Exchange Commission) can be consulted. These reports, often in the form of Form 20-F and Form 6-F filings, provide comprehensive information about the company, including subscriptions data, tariff details, staffing information, financial data, and more.
4. **Press releases and other media reports:** If data are unavailable from company annual reports or regulatory websites, subscriptions data can sometimes be estimated through industry analysis and news articles provided by various sources. These sources may include operator press releases, official statements from regulatory bodies or ministries, and reputable newspapers within the country. Such data sources

[30] Including Hong Kong, China; Macau, China; and Palestine.

may include absolute subscription numbers, market shares, penetration rates, and growth rates, which can be used to derive estimates for the country.

Once the data from all operators have been collected, the number of total active mobile-broadband subscriptions can be calculated by summing the number of active mobile-broadband subscriptions of each operator.

Estimation using market share

In case data from all operators are not available, data can be estimated by dividing the number of mobile-broadband subscriptions reported by the dominant operator (in terms of subscriptions) by its market share.

Estimation using trends

When data cannot be retrieved through the methods above, subscriptions can still be estimated by analysing trends from previous years. This estimation process is facilitated through the utilization of the Expert Modeller forecasting function in SPSS. The syntax is the following:

```
* Time Series Modeler.
TSMODEL
   /MODELSUMMARY  PRINT=[MODELFIT]
   /MODELSTATISTICS  DISPLAY=YES MODELFIT=[ SRSQUARE
RSQUARE MAE]
   /SERIESPLOT OBSERVED FORECAST
   /OUTPUTFILTER DISPLAY=ALLMODELS
   /SAVE  PREDICTED(Predicted)
   /AUXILIARY  CILEVEL=95 MAXACFLAGS=24
   /MISSING USERMISSING=EXCLUDE
   /MODEL DEPENDENT= "countries to be modelled"
      PREFIX='Model'
   /EXPERTMODELER TYPE=[ARIMA EXSMOOTH]
   /AUTOOUTLIER  DETECT=OFF.
```

Expert Modeller determines the best-fitting model for a country, either ARIMA or exponential smoothing model. This functionality automatically tests the data for seasonality, intermittency, and missing values, and selects appropriate model.