Imputation of IDI indicators

Daniela Benavente, Consultant

Geneva, December 11, 2018

Context and scope of the imputation

- Imputation is a crucial step in the computation of the ICT Development Index (IDI)
- 14 IDI indicators, 6 of these new indicators, 5 of these computed from subindicators
- In total: 8 + 1 + 18 = 27 indicators up for imputation

The 2018 IDI **(Indicators added in 2018)**

Access sub-index

1.1 Households with a computer (%)

1.2 Households with Internet access (%)

1.3 International Internet bandwidth (bit/s/Internet user)

1.4 Population covered by 3G mobile networks

- At least 3G (%) – G
- At least LTE/WiMAX (%) – L

1.5 Fixed-broadband subscriptions by speed tiers

- 256 kbit/s to 2Mbit/s (% of total) – S
- 2 to 10 Mbit/s (% of total) – M
- Equal to or above 10 Mbit/s (% of total) – F

Use sub-index

2.1 Individuals using the Internet (%)

2.2 Active mobile-broadband subscriptions (/100 inhabitants)

2.3 Mobile-broadband Internet traffic (/subscription) – T

2.4 Fixed-broadband Internet traffic (/subscription) – T

2.5 Phone ownership (%)

Skills sub-index

3.1 Mean years of schooling

3.2 Secondary gross enrollment ratio (%)

3.3 Tertiary gross enrollment ratio (%)

3.4 Individuals with ICT skills (%)

1. Copying or moving a file or folder
2. Using copy and paste tools to duplicate or move information within a document
3. Sending e-mails with attached files
4. Using basic arithmetic formula in a spreadsheet
5. connecting and installing new devices
6. Creating electronic presentations with presentation software
7. Finding, downloading, installing and configuring software
8. Transferring files between a computer and other devices
9. Writing a computer program using a specialized programming language

Missingness

			Years								% missing			
			2010	2011	2012	2013	2014	2015	2016	2017	2015	2016	2017	
IDI	**3 out of 3 (100%)**	**Index**	70	71	72	73	84	90	112	115	54.1%	42.9%	41.3%	
	3 out of 5 (60%)	**Access**	83	84	85	87	94	113	156	190	42.3%	20.4%	3.1%	
	3 out of 5 (60%)	**Use**	10	15	26	33	62	73	108	117	62.8%	44.9%	40.3%	
	3 out of 4 (75%)	**Skills**	123	125	119	108	107	111	67	4	43.4%	65.8%	98.0%	
	Average		74	74	73	73	83	95	105	108	51.6%	46.7%	44.9%	
	New indicators		6	9	14	17	39	60	73	90	69.3%	62.9%	54.3%	
Access	**1.1 HH with a computer (%)**	**xHH4_IDI**	91	93	87	92	67	92	134	189	53.1%	31.6%	3.6%	ITU
	1.2 HH with Internet access (%)	**xHH6_IDI**	89	90	90	89	93	85	150	190	56.6%	23.5%	3.1%	ITU
	1.3 Int'l Internet bandwidth (bit/s/user)	**i994u**	186	187	186	185	185	189	190	191	3.6%	3.1%	2.6%	ITU
	1.4 Pop covered by 3G mobile networks	**i271IDI**	0	0	12	18	60	124	135	193	36.7%	31.1%	1.5%	ITU
	Mobile	i271pop	133	135	145	129	142	181	192	195	7.7%	2.0%	0.5%	ITU
	Mobile3G	i271G	89	104	107	94	108	142	158	194	27.6%	19.4%	1.0%	ITU
	MobileLTE/WiMAX	i271GA	0	0	14	21	72	150	145	194	23.5%	26.0%	1.0%	ITU
	1.5 Fixed-broadband subscriptions by speed tier	**i992bIDI**	23	27	31	34	65	84	99	108	57.1%	49.5%	44.9%	
	Slow S (%)	i992b_256to2p	73	83	88	82	98	125	134	129	36.2%	31.6%	34.2%	
	Medium M (%)	i992b_2to10p	71	82	90	84	102	128	135	130	34.7%	31.1%	33.7%	
	Fast F (%)	i992b_G10p	65	77	87	81	99	121	134	130	38.3%	31.6%	33.7%	
	Slow S (subscriptions)	i4213_256to2	74	84	91	85	100	130	137	133	33.7%	30.1%	32.1%	
	Medium M (subscriptions)	i4213_2to10	71	84	93	87	105	133	137	134	32.1%	30.1%	31.6%	
	Fast F (subscriptions)	i4213_G10	66	80	88	83	101	126	136	134	35.7%	30.6%	31.6%	
	Fixed-broadband subscriptions	i4213tfbb	165	166	165	164	161	165	168	185	15.8%	14.3%	5.6%	ITU

			Years								% missing			
			2010	2011	2012	2013	2014	2015	2016	2017	2015	2016	2017	
Use	**2.1 Individuals using the Internet (%)**	**i99H**	119	103	82	83	83	100	178	193	49.0%	9.2%	1.5%	ITU
	2.2 Active mobile-broadband subsc (/100 inhab)	**i911mw**	188	178	172	183	191	195	195	195	0.5%	0.5%	0.5%	ITU
	2.3 Mobile-broadband Internet traffic	**i136mwi_psub**	11	16	26	27	46	65	88	103	66.8%	55.1%	47.4%	
	Mobile-broadband Internet traffic (exabytes)	i136mwi	11	16	27	28	48	68	90	104	65.3%	54.1%	46.9%	
	Active mobile-broadband subscriptions	i271mw	116	138	142	154	156	164	164	193	16.3%	16.3%	1.5%	ITU
	2.4 Fixed-broadband Internet traffic (/subsc)	**i135tfb_psub**	4	8	13	13	26	36	52	70	81.6%	73.5%	64.3%	
	Fixed-broadband Internet traffic (exabytes)	i135tfb	5	9	13	13	26	38	52	72	80.6%	73.5%	63.3%	
	2.5 Phone ownership	**xHH18_IDI**	0	0	0	9	32	37	48	43	81.1%	75.5%	78.1%	
Skills	**3.1 Mean years of schooling**	**iMYS**	58	56	50	42	52	50	2	0	74.5%	99.0%	100.0%	UNESCO
	3.2 Secondary gross enrollment ratio (%)	**isec**	139	144	138	130	133	137	97	9	30.1%	50.5%	95.4%	UNESCO
	3.3 Tertiary gross enrollment ratio (%)	**iter**	128	134	131	119	121	119	81	7	39.3%	58.7%	96.4%	UNESCO
	3.4 Individuals with ICT skills (%)	**xHH15IDI**	0	0	0	3	7	15	14	20	92.3%	92.9%	89.8%	
	1. Copying or moving a file or folder	xHH15c	0	30	33	4	42	48	49	49	75.5%	75.0%	75.0%	
	2. Using copy and paste tools to	xHH15d	0	29	32	5	42	42	34	28	78.6%	82.7%	85.7%	
	3. Sending e-mails with attached files	xHH15e	31	30	0	38	19	22	24	26	88.8%	87.8%	86.7%	
	4. Using basic arithmetic formula in a spreadsheet	xHH15f	0	30	33	5	42	47	46	44	76.0%	76.5%	77.6%	
	5. connecting and installing new devices	xHH15i	0	30	33	5	43	20	22	24	89.8%	88.8%	87.8%	
	6. Creating electronic presentations	xHH15s	0	0	0	6	12	46	44	42	76.5%	77.6%	78.6%	
	7. Finding, downloading, installing and configuring software	xHH15p	0	30	33	6	44	49	49	48	75.0%	75.0%	75.5%	
	8. Transferring files between a computer and other devices	xHH15t	0	30	33	7	43	49	49	48	75.0%	75.0%	75.5%	
	9. Writing a computer program using a specialized programming language	xHH15w	0	30	33	4	43	48	46	46	75.5%	76.5%	76.5%	

Missingness

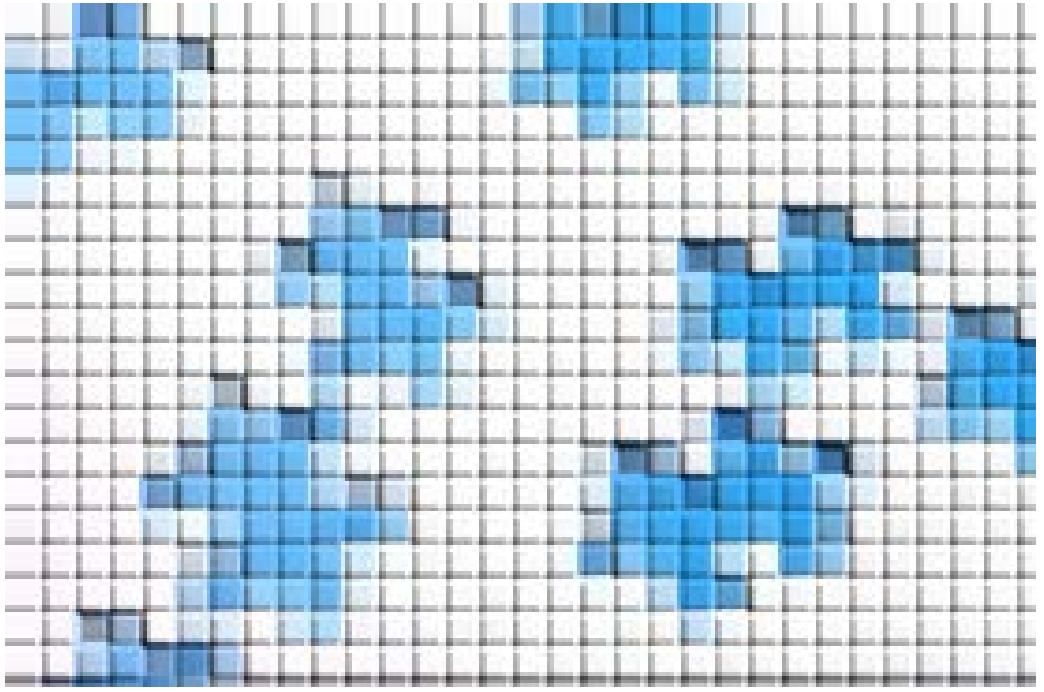
- Decreases in recent years, but 66.9% in 2017
- Decreases with income, 84.0% for low-income economies
- UN vulnerability indicators:
  - 77.7% for developing countries
  - 83.2% for LDCs
  - 77%for LLDCs
  - 83.7% for SIDS

Field	Criteria	Missing	Total	% of total
All	Total	32,928		100.0
All	Not missing	9,193	32,928	27.9
Database	Missing	23,735	32,928	72.1
Year	2017	2,752	4,116	66.9
Year	2016	2,692	4,116	65.4
Year	2015	2,721	4,116	66.1
Year	2014	2,865	4,116	69.6
Year	2013	3,232	4,116	78.5
Year	2012	3,092	4,116	75.1
Year	2011	3,075	4,116	74.7
Year	2010	3,306	4,116	80.3
Income group (WB)	Low income	4,726	5,628	84.0
Income group (WB)	Lower-middle income	6,630	8,463	78.3
Income group (WB)	Upper-middle income	6,373	8,904	71.6
Income group (WB)	High income	5,721	9,639	59.4
Country type (UN)	Developing	19,047	24,528	77.7
Country type (UN)	Developed	4,520	8,232	54.9
Least Developed Country	1	6,571	7,896	83.2
Non-LDC	0	17,164	25,032	68.6
Landlocked Developing Country	1	4,142	5,376	77.0
Non-LLDC	0	19,593	27,552	71.1
Small Island Devloping State	1	5,065	6,048	83.7
Non-SIDS	0	18,670	26,880	69.5

- Decreases with population
- Decreases with GNI per capita
- Missingness by region:
  - 54.6% in Europe
  - 52.3% in Northern America
  - Between 60 and 80% in Latin America, Asia, Northern Africa, Australia and New Zealand
  - Above 80% in the Caribbean, rest of Africa, rest of Oceania

Field	Criteria	Missing	Total	% of total
Population	>=1 billion people	18,224	26,166	69.6
Population	>=100 million people	3,753	4,788	78.4
Population	>=50 million people	839	966	86.9
Population	>=30 million people	442	504	87.7
Population	>=10 million people	309	336	92.0
Population	>=2 million people	-	-	NA
Population	>=0 people	168	168	100.0
GNI per capita (Atlas method)	>=100,000 current US$	19	42	45.2
GNI per capita (Atlas method)	>=50,000 current US$	1,033	1,890	54.7
GNI per capita (Atlas method)	>=30,000 current US$	1,562	2,772	56.3
GNI per capita (Atlas method)	>=10,000 current US$	3,423	5,649	60.6
GNI per capita (Atlas method)	>=2,000 current US$	9,210	12,327	74.7
GNI per capita (Atlas method)	>=0 current US$	6,859	8,400	81.7
M49SubRegion	Europe (M49)	3,027	5,544	54.6
M49SubRegion	Northern America (M49)	1,054	2,016	52.3
M49SubRegion	Caribbean	1,795	2,184	82.2
M49SubRegion	Latin America	2,533	3,360	75.4
M49SubRegion	Central Asia (M49)	638	840	76.0
M49SubRegion	Southern Asia (MDG=M49)	1,152	1,512	76.2
M49SubRegion	Eastern Asia (M49)	824	1,176	70.1
M49SubRegion	South-eastern Asia (MDG=M49)	1,368	1,848	74.0
M49SubRegion	Northern Africa (M49)	681	1,008	67.6
M49SubRegion	Western Asia (M49)	2,185	3,192	68.5
M49SubRegion	Eastern Africa	2,475	3,024	81.8
M49SubRegion	Middle Africa	1,289	1,512	85.3
M49SubRegion	Southern Africa	681	840	81.1
M49SubRegion	Western Africa	2,159	2,688	80.3
M49SubRegion	Australia and New Zealand (M49)	219	336	65.2
M49SubRegion	Oceania (M49) excl AUS & NZL	1,655	1,848	89.6

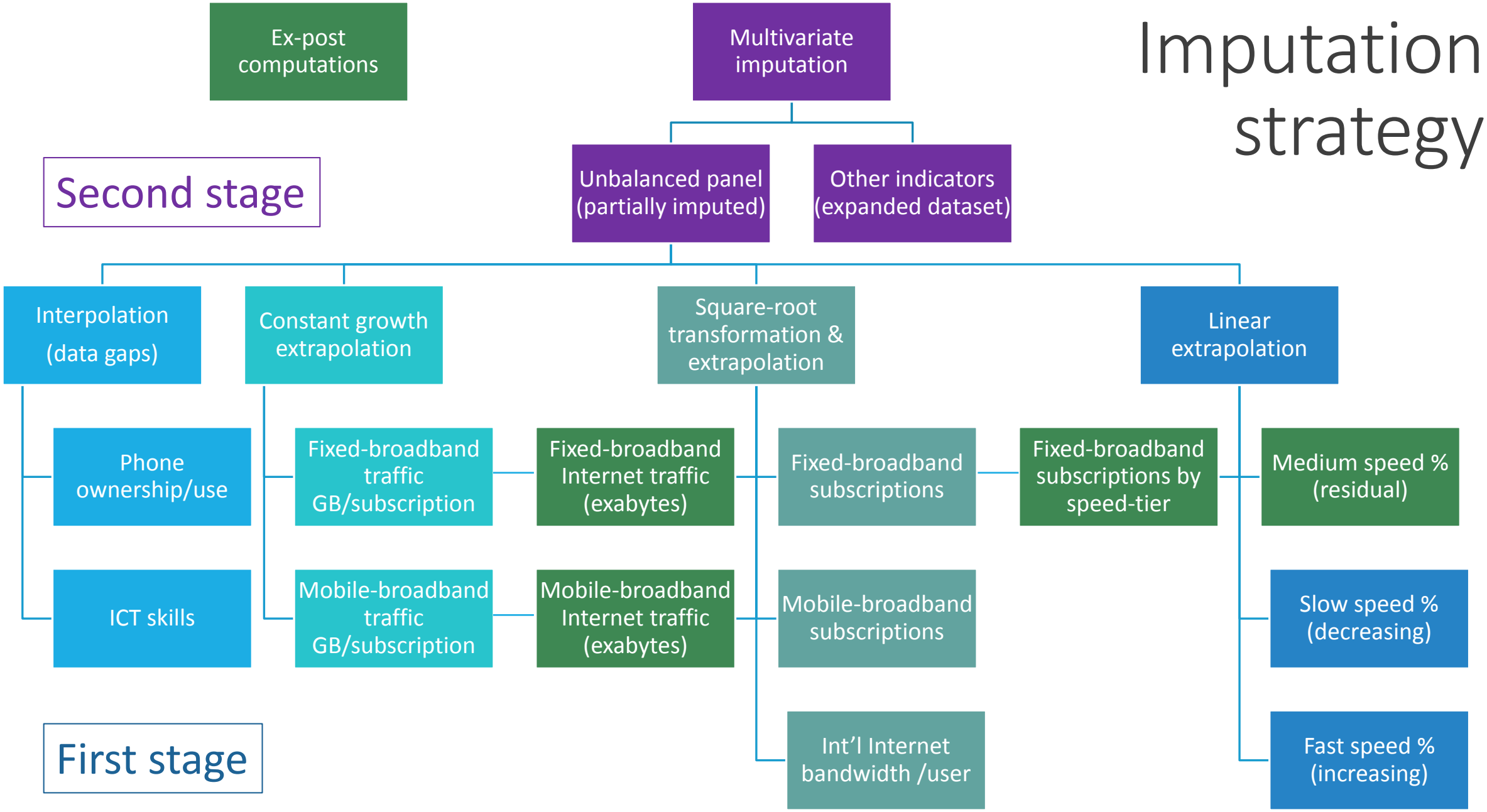
“The idea of imputation is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all”(Dempster & Rubin, 1983)



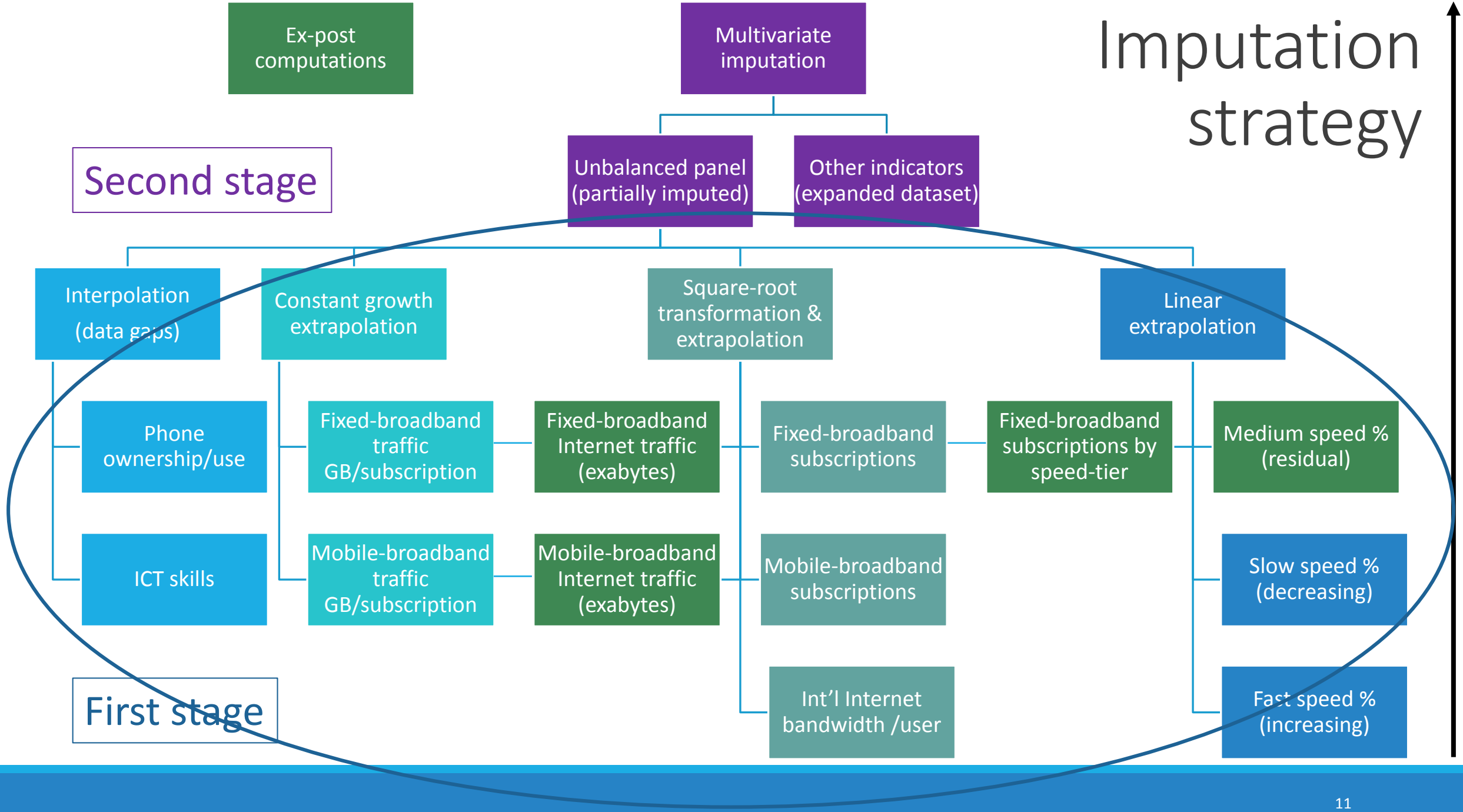
Imputation strategy

- First stage: Imputation for country series with **some observations** with **interpolation** and **extrapolation**
- Second stage: **Joint-multivariate multiple imputation** in the context of the ICT Development Index
- Strategies discarded:
  - LOCF and NOCB (last/next observation carried forward/backward): trends are not respected
  - Mean/mode imputation: generates bias and lowers variability in data
  - Regression-based models: lack of theoretical or empirical econometric model
  - Nearest neighbors: few donors for hot/cold deck imputation
  - Statistical packages that are cross-sectional and/or that cannot be specified to respect constraints or trends

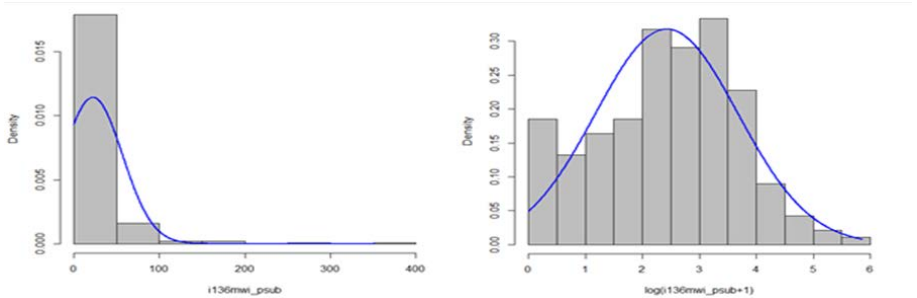
Imputation strategy



Imputation strategy



Summary for three series

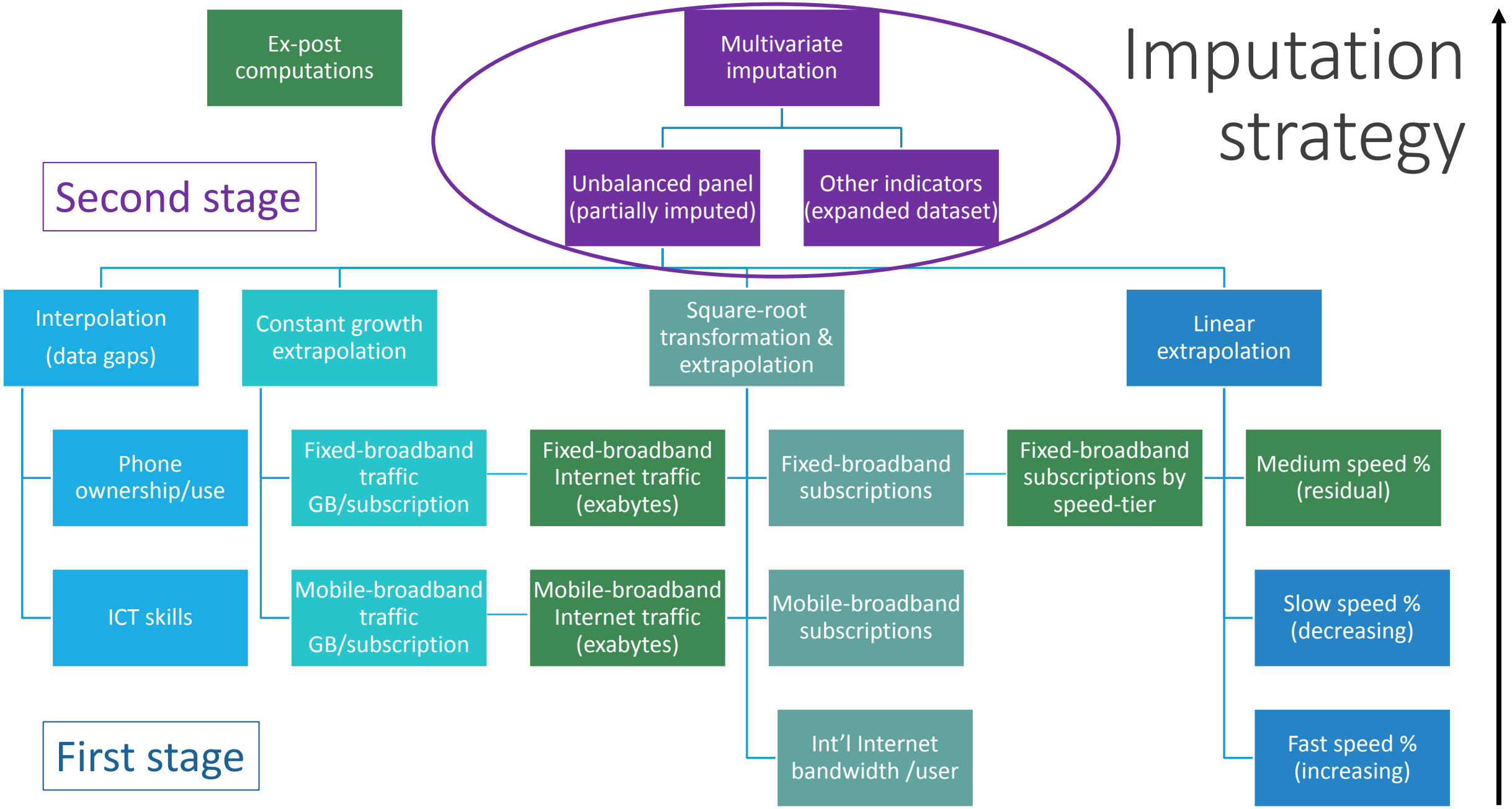


Indicator	3G mobile networks coverage	Fixed-broadband subscriptions by speed tier	Mobile/fixed-broadband Internet traffic per subscription
Coherence and aggregation constraints	Mobile-cellular network ≥ at least 3G (G) ≥ at least LTE-WiMax (L)	Subscriptions by speed tier = total subscriptions, so that S + M + F = 1	If subscriptions are positive, traffic must be missing or positive, but not zero
Series analysis, imputation of data gaps	Imputed by ITU	Linear interpolation of S (decreasing) and F (increasing), M as residual	Constant growth extrapolation T-1 = 0.75 T
Computation for the IDI	Weighted average 0.5 G + 0.5 L	Weighted sum 0.1 S + 0.35 M + F	Log-transformed (figure)

Outcomes of the first stage

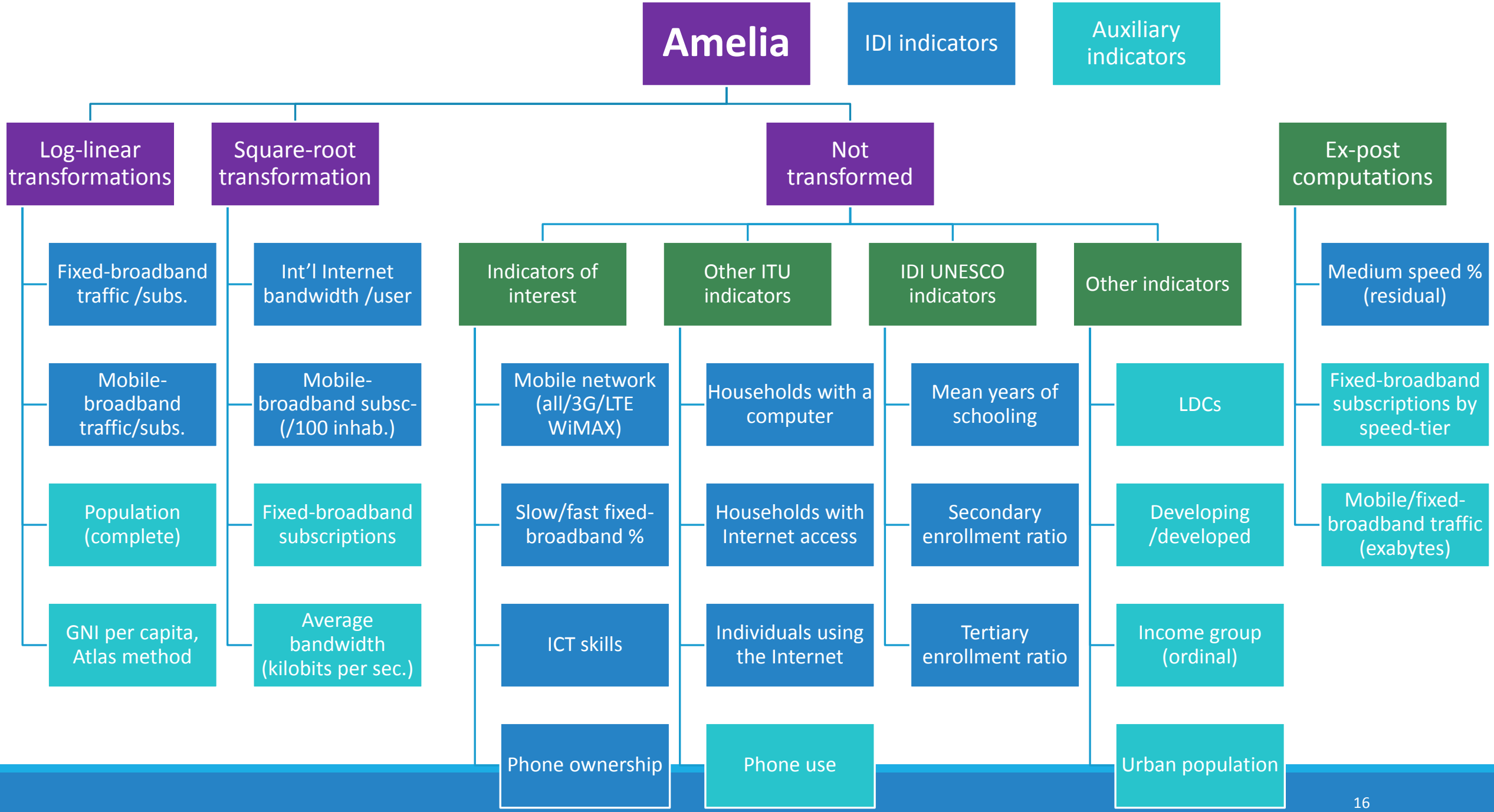
- Quantitative outcomes:
  - Missingness down to 59.8% (from 62.9%)
  - Imputation of 423 missing values for the 16 indicators of interest
- Qualitative outcomes:
  - Better knowledge of the dataset
  - Optimal use of the information contained in observed data
  - Impose coherence in trends, growth patterns and relationships between indicators
- All outcomes combined facilitate convergence in the second stage

Imputation strategy



4. Country series **without** observations
The Amelia package in R

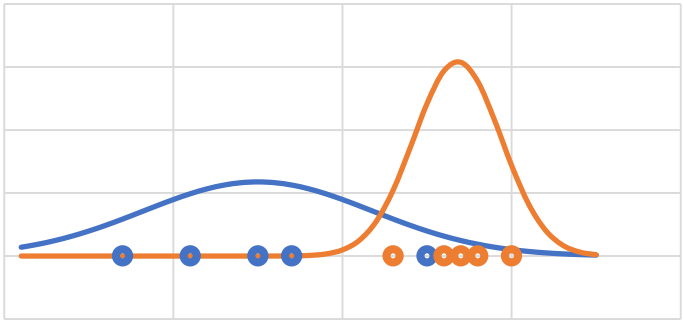
- Theoretical framework: **the ICT Development Index**
- Compatible with **panel data** (cross-sectional data and time effects)
- **Expectation Maximimization algorithm** (a non-analytical model)
- **Multivariate joint modelling approach**, as opposed to specifying a model for each indicator
- Seed value for **replicability**
- **A single estimate is computed for each missing value** (deterministic estimates)
- Stochastic multiple imputation to obtain **standard errors and confidence intervals for estimates**
- **Assumptions: data are missing-at-random, indicators are normally distributed**

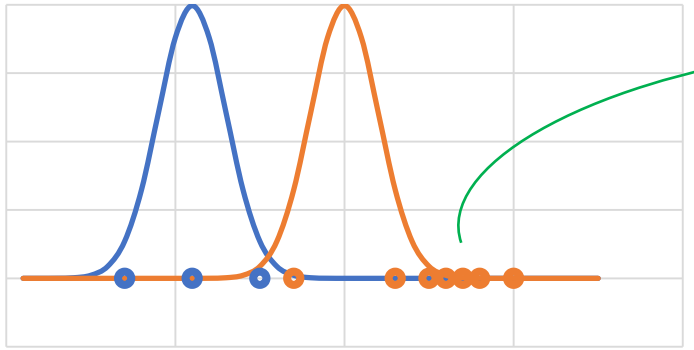

Amelia
IDI indicators
Auxiliary indicators
Log-linear transformations
Square-root transformation
Not transformed
Ex-post computations
Fixed-broadband traffic /subs.
Mobile-broadband traffic/subs.
Population (complete)
GNI per capita, Atlas method
Int'l Internet bandwidth /user
Mobile-broadband subsc-(/100 inhab.)
Fixed-broadband subscriptions
Average bandwidth (kilobits per sec.)
Indicators of interest
Other ITU indicators
IDI UNESCO indicators
Other indicators
Mobile network (all/3G/LTE WiMAX)
Slow/fast fixed-broadband %
ICT skills
Phone ownership
Households with a computer
Households with Internet access
Individuals using the Internet
Phone use
Mean years of schooling
Secondary enrollment ratio
Tertiary enrollment ratio
LDCs
Developing /developed
Income group (ordinal)
Urban population
Medium speed % (residual)
Fixed-broadband subscriptions by speed-tier
Mobile/fixed-broadband traffic (exabytes)

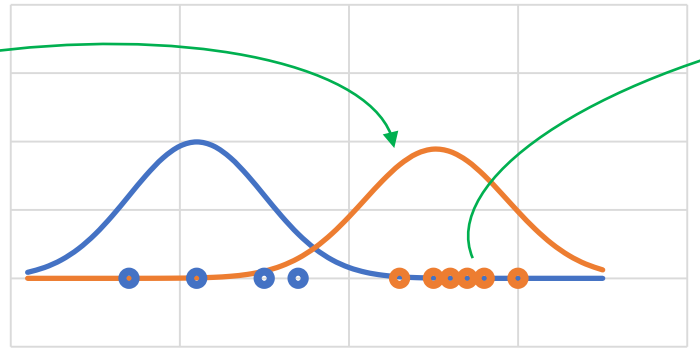
EM algorithm with one indicator

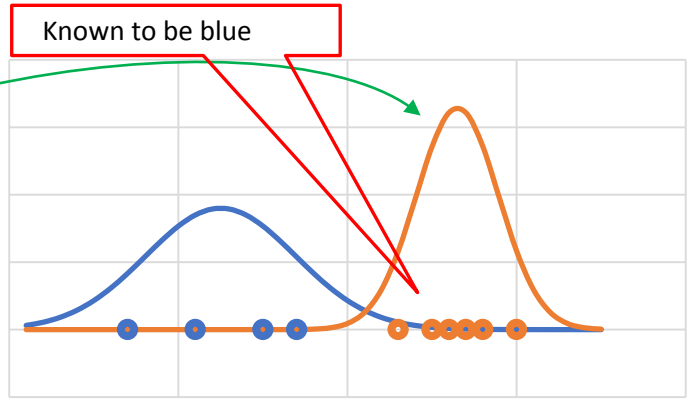
- First iteration: two random distributions
  - Expectation-step: likelihood (probability) of coming from each group, 3 mostly blue, 7 mostly orange (in k-means colours fully blue, fully orange, based on distance to mean only)
  - Maximization-step: mean and variances are computed again (maximum likelihood estimates for mean and variance), weighted by posterior probabilities (no weights with k-means)
- Start again the E-step (assign observations) and so on until convergence

The true models (not known)







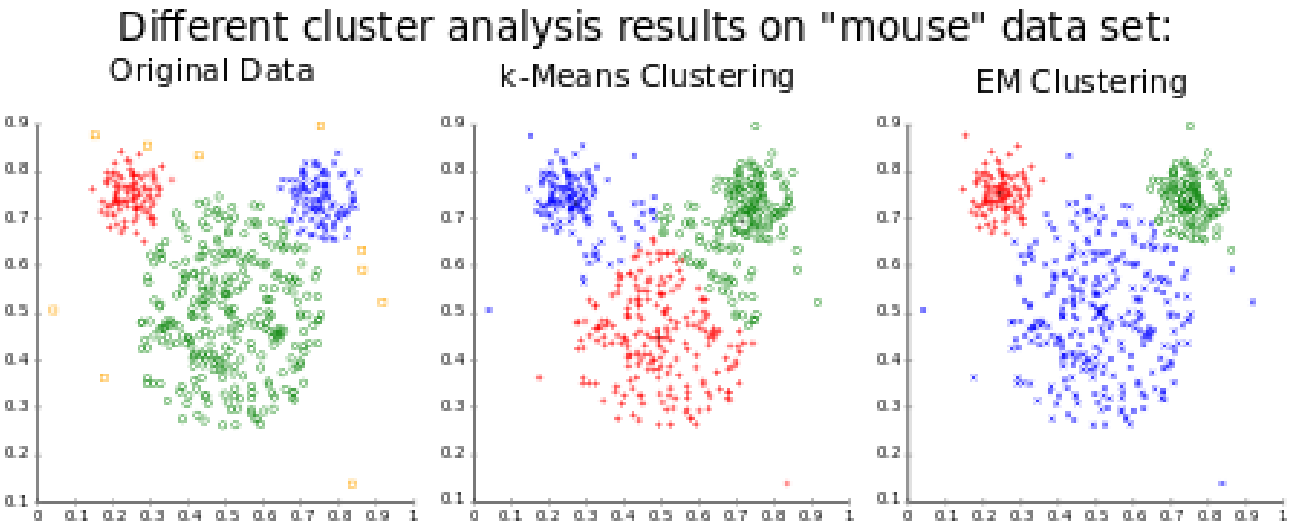


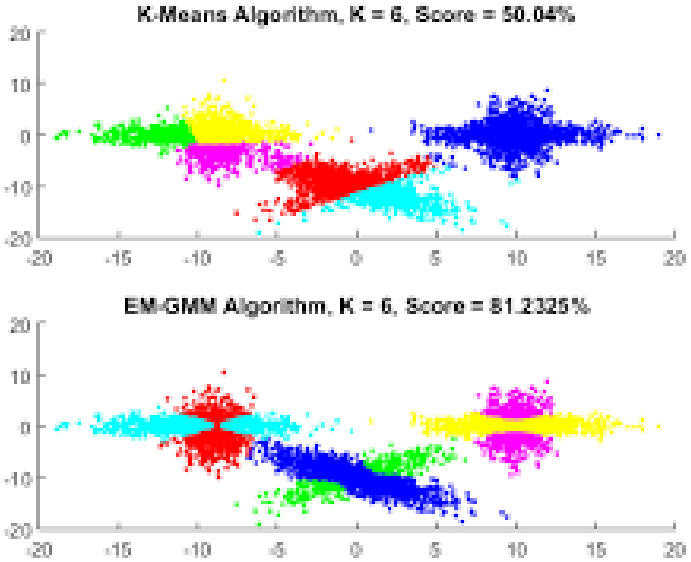
First iteration, random distributions for two groups, soft-assigns observations into groups (E-step)

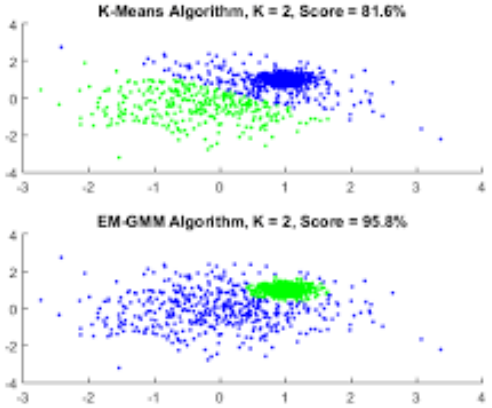
Second iteration, computes means and variances from 1st it. (M-step), re-assigns observations to groups (E-step)

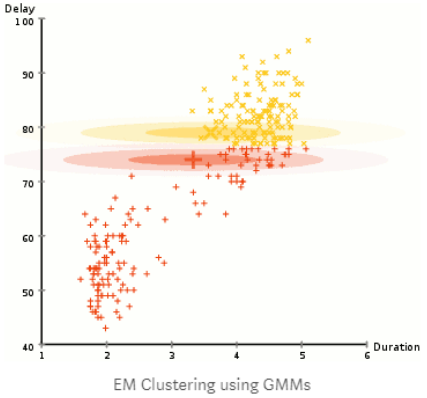
Third iteration, computes means and variances, assignments don't change, so iteration is final, one mistake, and true blue distribution is flatter

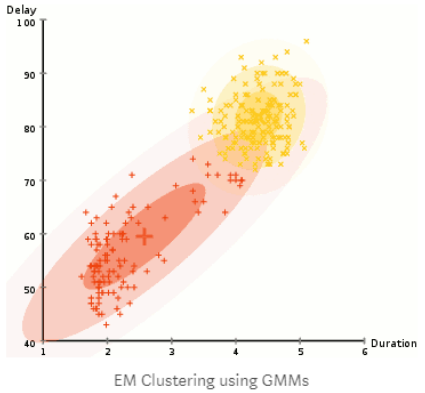
EM algorithm with two indicators

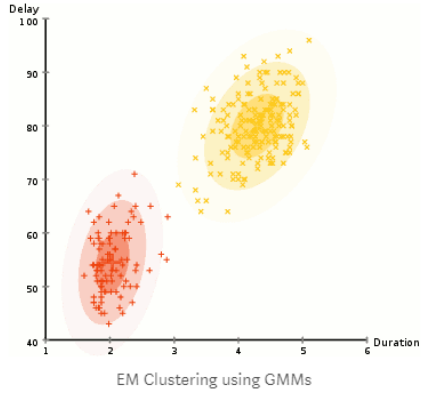












Outcomes of the second stage

- The algorithm achieved normal EM convergence after **1,935 iterations** over **116 patterns of missingness**
- **Quantitative outcomes:**
  - **Complete data** for 196 countries over the 2015-2017 period
  - 6,298 estimates overall (and even more with expost computations)
- Qualitative outcomes:
  - Multivariate EM imputation provides estimates that are **jointly likely given observed data**
  - Estimates computed on the basis of the **normality assumption**
  - Estimates based on the **conceptual framework of the IDI**
- Therefore estimates are particularly suited for **aggregation into the ICT Development Index**

The ICT Development Index

Daniela Benavente, Consultant

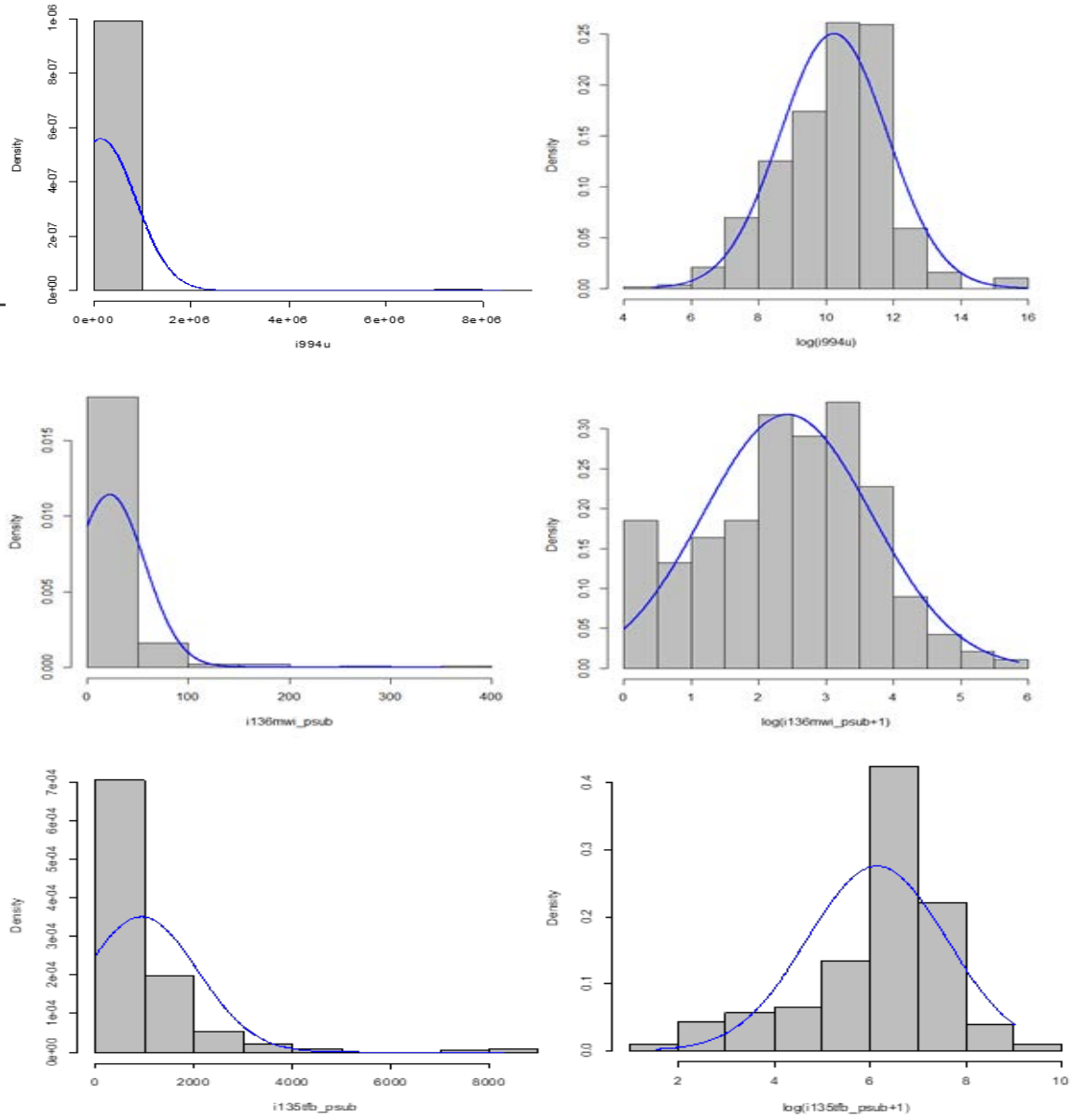
Geneva, October 17, 2018

Ultimate goals of modelling choices

- Modelling choices involved:
  - Transformations,
  - Reference values,
  - Normalization,
  - Weights and aggregation
- Objectives:
  - Continuity with past IDI editions
  - Ensure statistical coherence in addition to the conceptual coherence
  - Achieve balance in scores, so that outliers at the top are not unduly biasing results
  - Stability over time (through goalposts)
- The JRC assessed the **robustness** of scores and rankings to changes in modelling choices

Transformations to approach normality

- For the purposes of aggregation, indicator distributions should approach normality
- Three indicators had to be **log-transformed** for that purpose:
  - 1.3 International Internet bandwidth (bit/s/Internet user)
  - 2.3 Mobile-broadband Internet traffic (/subscription)
  - 2.4 Fixed-broadband Internet traffic (/subscription)
- Transformation used in past editions for 1.3



Normalization and scoring

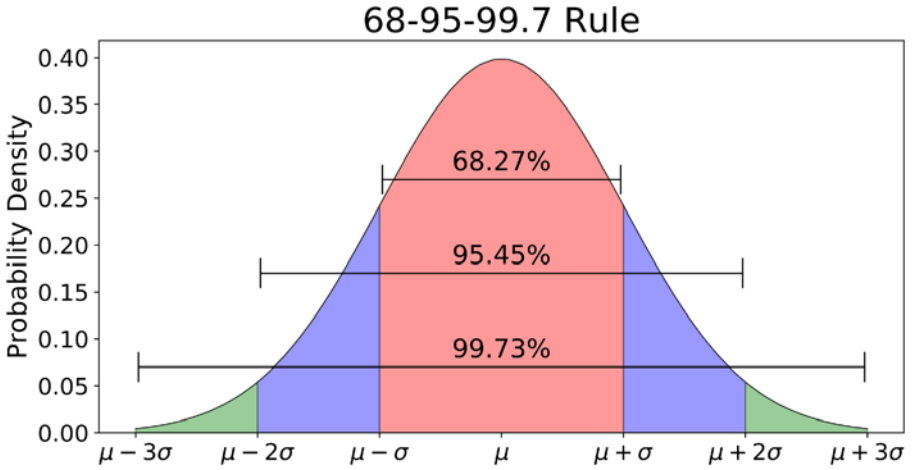
- **Scores from 0 to 10**
- **Normalization: distance to reference value**

$$score = \frac{value}{reference\ value} * 10$$

- For log-transformed indicators, the reference value is the **sample mean plus two standard deviations**

$$score = \frac{\log(X)}{mean + 2stdev} * 10$$

  - Formula used for 1.3 in past editions



- For a normal distribution, 95.45% of data is within two standard deviations σ of the mean μ
- Implication of the reference value $\mu + 2\sigma$:
  - Min at 0, mean at 5, max at 10,
  - 68.27% of scores between 2.5 and 7.5,
  - 50% of scores between 3.325 and 6.675 (interquartile range),
  - For the 196 countries, 5 highest values are winsorized (outlier frequency of 1/44 at each tail)

1.4 Population covered by 3G mobile networks

Mobile networks		At least 3G	At least LTE/WiMAX	Sum	Score (division by 10)
Weight		0.5	0.5	1.0	
Country A	Value (%)	60.7	7.0	-	
	Value * weight	30.3	3.5	33.8	3.4
Country B	Value (%)	97.0	50.0	-	
	Value * weight	48.5	25.0	73.5	7.4
Country C	Value (%)	100.0	20.0	-	
	Value * weight	50.0	10.0	60.0	6.0
Country D	Value (%)	100.0	100.0	-	
	Value * weight	50.0	50.0	100.0	10.0

Fixed-broadband subscriptions by speed tier as a % of total fixed-broadband subscriptions

Fixed-broadband subscriptions by speed tier		Slow (%)	Medium (%)	Fast (%)	Sum	Score (division by 10)
Weight		0.1	0.35	1	-	
Country A	Value (%)	12.3	39.8	48	100	
	Value * weight	1.2	13.9	48	63.1	6.3
Country B	Value (%)	0	10.8	89.2	100	
	Value * weight	0	3.8	89.2	93	9.3
Country C	Value (%)	53.5	38.4	8.1	100	
	Value * weight	5.4	13.4	8.1	26.9	2.7
Country D	Value (%)	0	100	0	100	
	Value * weight	0	100	0	35	3.5

3.4 Individuals with ICT skills – Scoring example

Group	Skill		Value (%)		Normalization		Average	Score
Basic	1	69.75	Highest value	77.98	No change	77.98	59.49	5.95
	2	65.29						
	3	**77.98**						
	8	55.84						
Standard	4	**44.39**	Highest value	44.39	Division by 0.8	55.49		
	5	44.27						
	6	41.29						
	7	33.60						
Advanced	9	**9.00**	No change	9.00	Division by 0.2	44.99		

The ICT Development Index
Reference values (RV) and weights (W)

Access	RV	W	Use	RV	W	Skills	RV	W
	100	40		100	40		100	20
1.1. Households with a computer (%)	100	20	2.1 Individuals using the Internet (%)	100	20	3.1 Mean years of schooling (years)	15	25
1.2 Households with Internet access (%)	100	20	2.2 Active mobile-broadband subscriptions (/100 inhabitants)	120	20	3.2 Secondary gross enrollment ratio (%)	100	25
1.3 International Internet bandwidth (bit/s/Internet user)	mean +2 sd	20	2.3 Mobile-broadband Internet traffic (/subscription)	Mean + 2 sd	20	3.3 Tertiary gross enrollment ratio (%)	100	25
1.4 Population covered by mobile networks (%)	100	20	2.4 Fixed-broadband Internet traffic (/subscription)	Mean + 2 sd	20	3.4 Individuals with ICT skills (%)	100	25
						Basic	100	33.3
1.5 Fixed-broadband subscriptions by speed tiers (% of total)	100	20	2.5 Individuals who own a mobile phone (%)	100	20	Standard	80	33.3
						Advanced	20	33.3

Thank you!

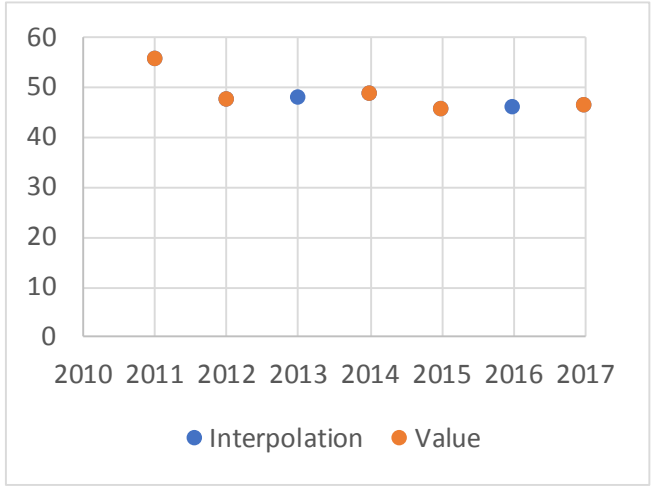
Backup slides

3. Country series **with** observations Interpolation (data gaps)

- Simple linear interpolation between two data points to fill in data gaps used for:
  - Mobile phone ownership (IDI 2.5)
  - Mobile phone use (auxiliary)
  - Nine ICT skills (IDI 3.4)
- No extrapolation for two reasons:
  - Some countries present decreasing trends or zigzagging
  - Some countries show increasing trends coupled with high growth or jumps, forecasted values easily exceeded 100%

Year	Data	Interpolation
2010	NA	NA
2011	55.56	55.56
2012	47.41	47.41
2013	NA	48.04
2014	48.68	48.68
2015	45.54	45.54
2016	NA	45.99
2017	46.43	46.43

Not imputed



3. Country series **with** observations Linear extrapolation

- Used for fixed-broadband data by speed tier (% of total)
- General description of imputation
  - Rescaling if sum is not 100
  - Calculation of differences to sum of 100
  - Linear extrapolation of **slow speed tier data** (decreasing trend)
  - Linear extrapolation of **fast speed tier data** (increasing trend)
  - **Medium speed tier data** computed as a residual
  - Fixes on a case-by-case basis

	Country A Original data				Country B Original data			
Year	Slow	Medium	Fast	Sum	Slow	Medium	Fast	Sum
2010	11.10	NA	NA	11.10	NA	NA	NA	-
2011	10.38	NA	NA	10.38	83.64	NA	NA	83.64
2012	11.97	NA	NA	11.97	NA	NA	NA	-
2013	NA	NA	NA	-	NA	NA	NA	-
2014	9.42	NA	NA	9.42	NA	11.35	1.45	12.80
2015	5.47	1.15	NA	6.62	81.08	21.77	2.86	105.71
2016	NA	NA	NA	-	59.79	34.73	5.48	100.00
2017	9.35	3.27	87.38	100.00	39.64	47.82	12.54	100.00
	Imputed data				Imputed data			
Year	Slow	Medium	Fast	Sum	Slow	Medium	Fast	Sum
2010	11.10	-	88.90	100.00	100.00	-	-	100.00
2011	10.38	-	89.62	100.00	83.64	16.36	-	100.00
2012	11.97	-	88.03	100.00	85.00	15.00	-	100.00
2013	9.70	-	90.30	100.00	86.00	14.00	-	100.00
2014	9.42	-	90.58	100.00	87.20	11.35	1.45	100.00
2015	5.47	1.15	93.38	100.00	76.70	20.60	2.70	100.00
2016	8.16	1.46	90.38	100.00	59.79	34.73	5.48	100.00
2017	9.35	3.27	87.38	100.00	39.64	47.82	12.54	100.00

Problem, not equal to 100

3. Country series **with** observations
Square root transformation & extrapolation

- Concerns three indicators **right-skewed with high growth:**
  - International Internet bandwidth (bits/s), per Internet user
  - Fixed-broadband subscriptions
  - Active mobile-broadband subscriptions
- Square root transformation and extrapolation:
  - These indicators were **transformed by taking square roots**
  - Imputation is computed by **linear forecasting of transformed indicators**
  - Fixes on an ad-hoc basis

Bad fit and negative estimate (-3.3 million)

		Forecasts			
Year	Original	Linear	Square root	Logarithm	Final
2010	2,175,300	NA	2,426,027	3,229,477	2,175,300
2011	NA	6,437,455	6,497,301	5,280,285	6,497,301
2012	NA	16,226,596	12,534,360	8,633,414	12,534,360
2013	23,522,000	26,015,737	20,537,205	14,115,875	23,522,000
2014	28,001,960	35,804,879	30,505,834	23,079,853	28,001,960
2015	42,330,000	45,594,020	42,440,249	37,736,208	42,330,000
2016	56,606,000	55,383,161	56,340,449	61,699,761	56,606,000
2017	71,983,153	65,172,303	72,206,435	100,880,843	71,983,153

Best fit for most countries

Bad fit (big increase)

3. Country series **with** observations
Constant growth extrapolation (traffic data)

- Applied to mobile/fixed-broadband traffic per subscription:
  - Subscriptions imputed by square root forecasting (previous slide)
  - Constant growth of 33.3% applied to traffic per subscription
    - Each value between 0.6 to 0.8 times the value for the next year; 0.75 on average
    - This ratio of 0.75 corresponds to 33.3% growth
    - Stable, particularly for fixed-broadband traffic
  - Traffic in exabytes computed from the other two estimates
  - Ad-hoc fixes
- Advantages:
  - The three series are reconciled with expected trends
  - Best use of existing traffic data (one datapoint is required)

Fixed-broadband			
Year	Subsc.	GB/ subscription	Traffic in petabytes
2010	5,391	237.5	1.28
2011	6,852	316.7	2.17
2012	31,077	422.2	13.12
2013	37,988	563.0	21.39
2014	24,835	750.6	18.64
2015	23,694	1,000.9	23.71
2016	26,385	1,334.5	35.21
2017	25,062	1,939.5	48.61
1 petabyte = 0.001 exabyte			
Imputed data			

1.4 Population covered by 3G mobile networks

- **Policy objective: universal coverage by LTE/WiMAX** or more advanced technologies
- A ***weighted average*** achieves these objectives, if:
  - G is percentage of the population covered by at least 3G (i271G)
  - L is percentage of the population covered by at least LTE/WiMAX (i271GA), then

$$score = gG + (1 - g)L = g(G - L) + L$$

  - A country with 100% LTE/WiMAX, i.e. $G = L = 100$, scores the maximum
  - A country with only 2G coverage, i.e. $G = L = 0$ scores 0
- A lower weight is assigned to lesser technologies:
  - $g < 1$ ($G - L$ which is the percentage of the population covered by 3G but not by LTE/WiMax)
  - 1 for coverage by at least LTE/WiMAX

Descriptive statistics - Population covered by mobile networks, a composite (%)								
Parameter g	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
Ratio LTE/WiMAX to simple 3G	**10**	**5**	**3.33**	**2.5**	**2**	**1.66**	**1.43**	**1.25**
Mean	63.7	65.8	68.0	70.1	72.3	74.5	76.6	78.8
Standard deviation	34.6	32.7	30.9	29.3	27.8	26.5	25.4	24.5
Skewness	-0.5	-0.5	-0.6	-0.7	-0.8	-1.0	-1.2	-1.4
Kurtosis - 3	-1.2	-1.2	-1.0	-0.8	-0.5	0.0	0.5	1.1
1st quartile	36.4	41.7	43.2	49.0	52.4	57.0	65.0	70.0
Median	74.3	75.0	76.9	78.4	81.0	83.1	86.0	88.0
3rd quartile	96.4	96.6	97.0	97.3	97.6	97.9	98.0	98.2

- Impossibility of aiming for a median value of 50% for indicator 4
- No value of parameter g would ensure that the distribution approached normality
- Qualitative judgment to derive a value of g of 0.5 so that the formula becomes:

$$score = 0.5G + 0.5L = 0.5(G - L) + L$$

  - LTE/WiMAX (L) weights twice as much as simple 3G ($G - L$)
  - A country with 100% 3G and no LTE/WiMAX scores 5 out of 10

1.5 Fixed-broadband subscriptions by speed tiers as a % of total fixed-broadband subscriptions

- **Policy objective: universal coverage of the fastest speed**
- A ***weighted sum*** achieves these objectives, if
  - S: 256 kbit/s to 2 Mbit/s subscriptions as a % of total fixed-broadband subscriptions (i4213_256to2)
  - M: 2 to 10 Mbit/s subscriptions as a % of total fixed-broadband subscriptions (i4213_2to10)
  - F: Equal to or above 10 Mbit/s subscriptions as a % of total fixed-broadband subscriptions (i4213_G10)

$$score = sS + mM + F$$

- Lesser technologies get lower weights, i.e. $s < m < 1$
  - A country with 100% fast speed subscriptions, i.e. $S = M = 0$ and $F = 100$, scores the maximum
  - A country with no broadband subscriptions, i.e. $S = M = F = 0$ scores 0
- Equivalent to second formula for mobile networks: $score = g(G - L) + L$
  - "at least" percentages v. percentages adding up to 1

Descriptive statistics - Fixed-broadband subscriptions by speed tiers (% of total)								
Parameter s (slow speed)	0.10	0.10	0.10	0.20	0.30	0.40	0.50	0.60
Parameter m (medium speed)	0.20	0.35	0.50	0.30	0.40	0.50	0.60	0.70
Ratio medium to slow	2.00	3.50	5.00	1.50	1.33	1.25	1.20	1.17
Ratio fast to medium	5.00	2.86	2.00	3.33	2.50	2.00	1.67	1.43
Mean	53.97	58.25	62.92	59.42	64.86	70.31	75.75	81.20
Standard deviation	29.79	27.95	26.79	26.34	22.90	19.46	16.03	12.60
Skewness	0.13	0.02	-0.20	0.12	0.11	0.09	0.07	0.03
Kurtosis - 3	-1.44	-1.37	-1.23	-1.44	-1.44	-1.43	-1.42	-1.40
1st quartile	27.81	34.87	40.42	36.26	44.73	53.09	61.34	70.00
Median	49.67	56.18	62.09	55.95	62.23	68.36	74.49	80.28
3rd quartile	83.72	86.02	88.85	85.72	87.72	89.72	91.72	93.71

- The most appropriate values for parameters s and m are 0.1 and 0.35

$$score = 0.1S + 0.35M + F$$

  - The medium speed tier weights 3.5 times as much as the slow speed tier
  - The fast speed tier weights 10 times as much as the slow speed tier
  - A country with no fast speed subscriptions scores between 1 and 3.5

3.4 Individuals with ICT skills

- The EGTI/EGH Sub-group on Reviewing the IDI Indicators assigned the nine skills to three groups:

Group	Computation	Reference value
Basic	The highest value among indicators 1, 2, 3 and 8	100
Standard	The highest value among indicators 4, 5, 6 and 7	80
Advanced	The value of indicator 9	20

- Reference values are set according to policy objectives:
  - All individuals should achieve basic skills, the reference value (goalpost) is therefore kept at 100
  - For standard skills, a reference value is set at 80 (the highest observed value in 2015-2017 is 78.23)
  - For skill 9 (advanced), a reference value is set at 20 (the highest observed value in 2015-17 is 18.28)

Reference values

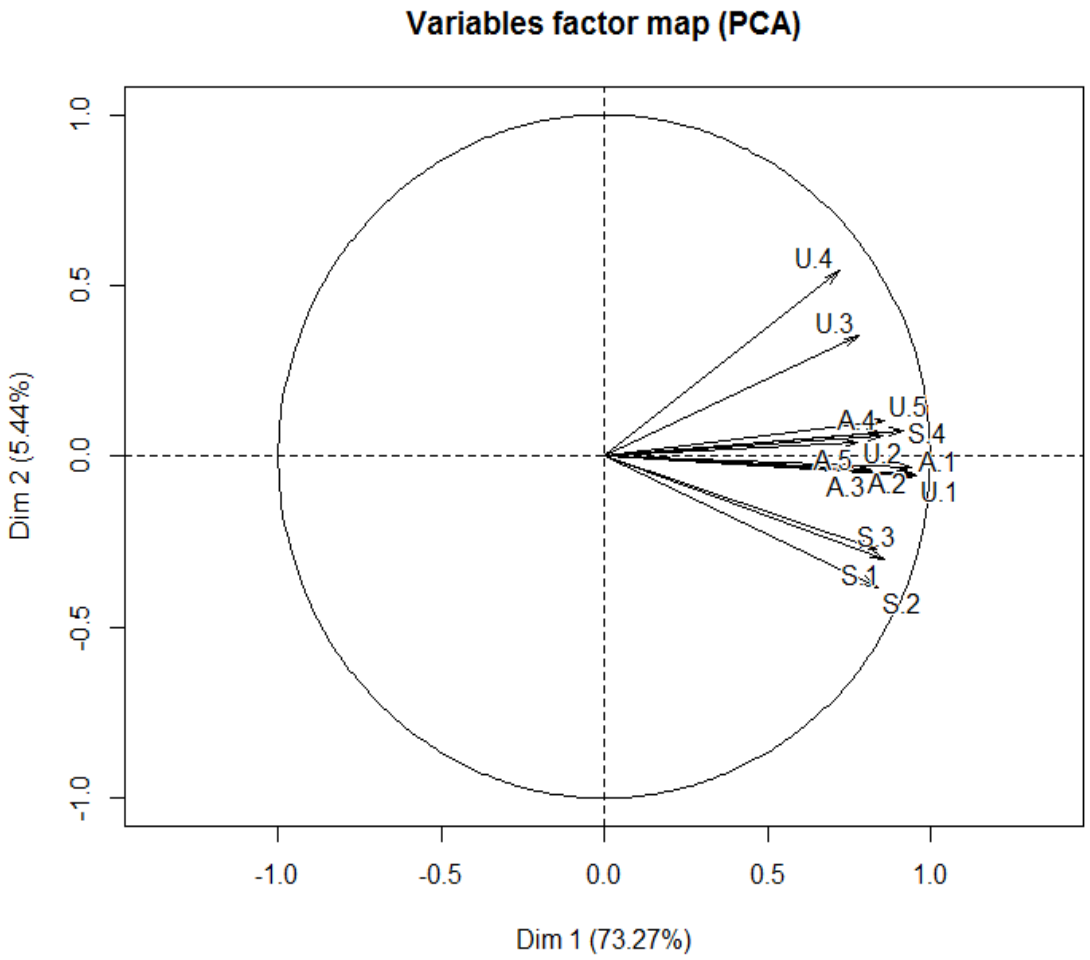
- Objectives:
  - **Winsorization of outliers at the top**
  - Policy **goalpost,** and
  - **Allow relative and absolute comparisons over time**
- Reference values of past editions are maintained:
  - 100 for percentages 1.1, 1.2, 2.1, 3.2 and 3.3
  - 120 for 2.2 Active mobile-broadband subscriptions per 100 inhabitants (2015-17: interquartile range of 20.24 to 80.06 with median value of 49.88, no reason to increase it)
  - 15 years for 3.1 Mean years of schooling, which is the projected years of schooling by 2025 by UNDP
  - Sample mean plus two standard deviations for 1.3
- New indicators:
  - 100 for composites 1.4, 1.5 and for indicator 2.5 (phone ownership)
  - Sample mean plus two standard deviations for indicators 2.3 and 2.4 (on broadband traffic)
  - Composite 3.4 (ICT skills): 100 for basic skills, 80 for standard skills and 20 for advanced skills

Weights and aggregation

- Aggregation is based on **weighted arithmetic averages**
- At the index level, weights of past editions are maintained:
  - 40 for the Access sub-index
  - 40 for the Use sub-index
  - 20 for the Skills sub-index
- Simple weights for the aggregation of indicators into sub-indices
- Simple weights for the aggregation of ICT skills groups:
  - Basic
  - Standard
  - Advanced

Equal weights

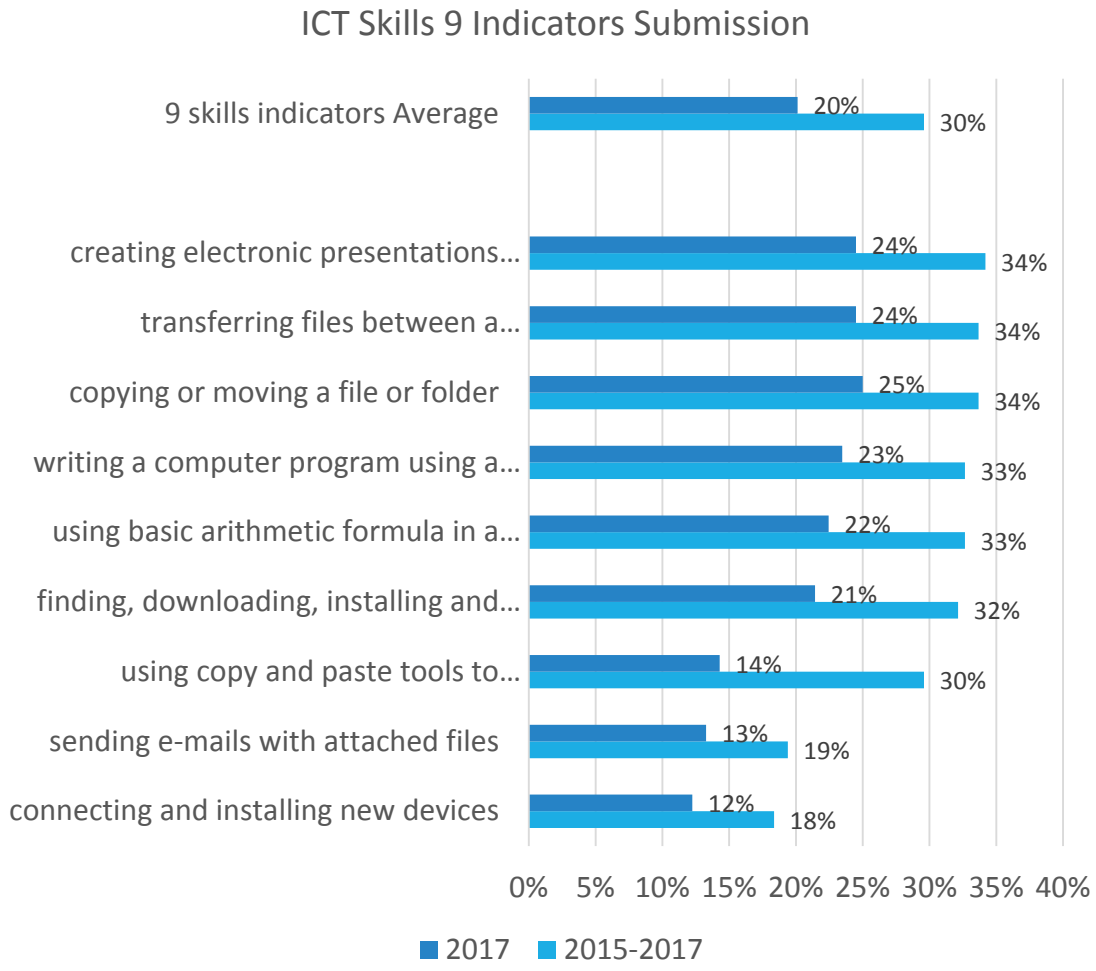
- Equal weights has the benefit of simplicity.
- The statistical audit performed by the JRC confirmed that using equal weights had good statistical properties.
- PCA analysis confirms the presence of a single statistical dimension at the index level (capturing 73% of the variance) and at the sub-index levels (capturing between 75% and 82% of the variance)



Source: Joint Research Center Audit, October 2018

Why is the 20% weight for the Skills Sub-index maintained?

- ICT skills indicators have the lowest data submission rates: 20% in 2017
- With equal weights, compared to Access and Use indicators, Skills indicators would get a higher contribution to the IDI
  - This is because Access and Use have five indicators each, compared to four for Skills
  - With 40-40-20, indicator final weights (contributions) are 8% for Access and Use and 5% for Skills
  - With equal weights (33-33-33), these contributions would be 6.67% and 8.33% respectively



Amelia assumptions and specifications

- The first assumption is that data are missing at random (MAR):
  - Missingness and data are functions of observed data
  - Estimation based on an **expanded dataset** including indicators correlated with incomplete series (IDI indicators for example), and other indicators that could explain missingness itself (such as income group or development status)
- The second assumption is that data are multivariate normally distributed
  - Some indicators are transformed to approach normality
  - Among others, square-root transformations in the first stage
  - Distributions approaching normality is also desirable for composite indicators
- Bounds set for all indicators

Mobile/fixed broadband traffic per subscription
Original, log-transformation, QQ-plot