WP/19/210

IMF Working Paper

Digital Connectivity in Sub-Saharan Africa: A Comparative Perspective

by C. Emre Alper and Michal Miktus

INTERNATIONAL MONETARY FUND

WP/19/210

IMF Working Paper

Digital Connectivity in Sub-Saharan Africa: A Comparative Perspective

by C. Emre Alper and Michal Miktus

INTERNATIONAL MONETARY FUND

IMF Working Paper

African Department

Digital Connectivity in Sub-Saharan Africa:

A Comparative Perspective[1]

Prepared by C. Emre Alper and Michal Miktus

Authorized for distribution by Benedict Clements

September 2019

Abstract

Higher digital connectivity is expected to bring opportunities to leapfrog development in sub-Saharan Africa (SSA). Experience within the region demonstrates that if there is an adequate digital infrastructure and a supportive business environment, new forms of business spring up and create jobs for the educated as well as the less educated. The paper first confirms the global digital divide through the *unsupervised machine learning* clustering K-means algorithm. Next, it derives a composite digital connectivity index, in the spirit of De Muro-Mazziotta-Pareto, for about 190 economies. Descriptive analysis shows that majority of SSA countries lag in digital connectivity, specifically in infrastructure, internet usage, and knowledge. Finally, using fractional logit regressions we document that better business enabling and regulatory environment, financial access, and urbanization are associated with higher digital connectivity.

JEL Classification Numbers: C43, O33, O57

Keywords: Digitalization, unsupervised machine learning, fractional logit model

Author's E-Mail Address: EAlper@imf.org, MMiktus@imf.org

[1] The authors wish to express their gratitude to Aidar Abdychev, Ben Clements, Jean Philippe Gillet, Clement Ncuti, Laure Redifer, Axel Schimmelpfennig, Murat Yavuz, and participants of the IMF seminar on July 18, 2019 for useful comments and suggestions.

I. INTRODUCTION

1. **Large uncertainties characterize the major trends determining the future of work in sub-Saharan Africa (SSA) and connectivity is a key policy area.** With its population projected to reach about 1.7 billion by 2040 from 1.0 billion currently, the United Nations projects a net increase in the working-age population (15–64 years) in SSA of about 20 million people per year. The need to generate 20 million jobs per year during the next two decades will be the key challenge facing policy makers in the SSA. The October 2018 Regional Economic Outlook for sub-Saharan Africa (IMF, 2018) identifies connectivity as a key policy area to promote job creation and yield dramatic improvements in living conditions. Connectivity goes beyond the need for traditional physical infrastructure of roads, railways, and ports, which is currently the focus of most country investment plans.

2. **SSA countries also need to be digitally connected to take advantage of technical change and growth opportunities.** Higher digital connectivity, coupled with an improved business climate, strong investment in people's education and health, and good governance would deliver digital dividends, deemed critical for the twenty-first century workplace (World Bank, 2016). Experience within the region demonstrates that if there is an adequate digital infrastructure and a supportive business environment, new forms of business spring up and create jobs for the educated as well as the less educated.[2] The region has been investing heavily in Information, Communications and Technology (ICT) infrastructure, including most recently, internet and mobile-cellular signal coverage.

3. **Nonetheless, the quantity of infrastructure per se is only a part of the challenge**. It is also important to consider the quality of infrastructure and its costs to users. Ongoing efforts to reform the policy and regulatory frameworks to make broadband access more affordable, accessible and universal, needs to be accompanied by skills development to fully exploit the technological advancement benefits. Finally, the population's capacity to access the Internet, including cultural acceptance, supporting policy, and availability of smartphones and computers at the household level are all necessary factors.

4. **We investigate the current state of play in digital connectivity in the SSA from a comparative perspective and analyze the drivers of heterogeneity across countries.** The paper improves upon previous work by (i) assessing a significantly higher number of ICT indicators; (ii) using the most recent (2016–17) available data for a comprehensive set of countries based on data availability (193 economies); and (iii) applying several methodologies, including machine learning techniques to investigate the existence of global digital divide and to formulate a composite index

[2] Among others, see Hjort and Poulsen (2019).

across countries.[3] Using fractional logit regressions with the aforementioned index as the dependent variable, the paper assesses the relative importance of various factors on digital connectivity, including SDG indicators, variables that characterize countries' business and regulatory environment, risk, transparency and corruption perceptions, as well as the usual set of macroeconomic indicators.

5. **We first employ unsupervised machine learning algorithms: (i) clustering technique of k-means to assess the existence of global digital divide; and (ii) dimensionality reduction, via principle components analysis (PCA), to investigate variation in digital connectivity.** In this initial step, we do not impose any modelling structure on the available ICT variables. To put this differently, we let the data speak for itself and determine the optimal number cluster(s) and the countries in each cluster based on elbow technique, silhouette method, and gap statistics. The intention of the principal components analysis (PCA) is to motivate the composite index of digital connectivity by checking if a quasi-linear technique of dimensionality reduction would be a good approximation.

6. **The paper constructs a digital connectivity index by imposing a modelling structure on the ICT variables and grouping them under five fundamental categories.** These are key sub-indices summarizing a country's ability to access ICT in line with those used by International Telecommunication Union (ITU). The five categories are (i) infrastructure; (ii) knowledge; (iii) affordability; (iv) quality; and (v) actual internet usage. These sub-indices are then aggregated in a single composite index through the Mazziotta-Pareto methodology (De Muro and others, 2011), allowing us to summarize a set of individual indicators that are assumed to be not fully substitutable. We construct the composite digital connectivity index, Enhanced Digital Access Index (EDAI), by using an improved aggregation technique, expand the variables which inform the index, and use the most recent available data for a larger number of countries relative to the Digital Access Index (DAI), launched by International Telecommunication Union, ITU, in 2003. Next, we use the EDAI to explore the drivers of digital connectivity variation in the world, including in SSA, by checking relative strengths and weaknesses across the five dimensions. The index can be used by policymakers to assess the level of preparedness for their countries for the Fourth Industrial Revolution.

7. **Finally, we estimate fractional logit models for the full sample and other country groupings, including SSA, to assess the relationship of various factors with digital connectivity.** We use over 100 independent variables and use step-wise regressions to reduce the number of explanatory variables by minimizing the quasi Akaike Information Criterion. Besides macroeconomic indicators, explanatory variables include various indicators related to the Sustainable Development Goals (SDGs), ease of doing business, regulatory environment, transparency, country risk, employment, climate, and corruption perceptions. Upon narrowing the number of independent

[3] The term digital divide or digital split owes to Norris (2001). Drivers of digital divide include socio-economic factors, geographical factors, educational, attitudinal and generational factors, or through physical disabilities (Cullen, 2001).

variables, we estimate the same model for Advanced Economies (AEs), Emerging Market and Middle-Income Countries (MICs), Low Income Developing Countries (LIDCs), and the SSA economies to verify whether these variables are robust across different country groupings. Finally, only for SSA economies, we also check which components of the Country Policy and Institutional Assessment (CPIA) variables affect digital connectivity, while controlling for per capita income.

8. **Our results indicate the existence of global digital divide and significant correlations between the business and regulatory environment and digital connectivity**. The results suggest there is room for policy action to improve connectivity by addressing these. Specifically, the results indicate that there is a global digital divide, with a clustering of countries into three main groups; (ii) the variation in digital connectivity across countries can be broadly approximated by the first principal component, motivating a quasi-linear index to construct the index of digital connectivity; (iii) there is a significant heterogeneity in digital connectivity across different analytical country groupings based on income and geography; (iv) the majority of SSA countries lag behind in digital connectivity, with the exception of Botswana, Cabo Verde, Gabon, Lesotho, Mauritius, Seychelles, and South Africa and LIDCs such as Ghana and Rwanda; (v) among the five dimensions, SSA countries on average perform well in terms of affordability and quality, but do less well on infrastructure, internet usage, and knowledge; and finally (vi) fractional logit regressions underscore the importance of the regulatory and business enabling environment, higher urbanization and urban access to electricity for digital connectivity. Estimation results for SSA indicate that better business enabling and regulatory environment, financial access, urbanization, and availability of postal services are associated with higher digital connectivity. Specifically, we find that leveling the playing field for female entrepreneurs and reducing property registration costs are positively related to higher digital connectivity.

9. **The structure of the paper is as follows:** Section II introduces the unsupervised machine learning algorithm of k-means and its implementation on the ICT data, both for the world and the SSA. Section III details the construction of the composite EDAI through the Mazziotta-Pareto methodology, and provides robustness verifications in the form of equal and progressive weighting schemes. Section IV presents the results based on the fractional logit regressions to explore the factors which correlate strongly with digital connectivity across countries. Section V concludes.

II. Estimating the Digital Divide: Unsupervised Machine Learning

10. **In the following section we describe the ICT dataset used in estimations and the methodology to investigate patterns across countries**. Specifically, we describe the data used to construct the composite index of digital connectivity and the rationale for using the k-means algorithm for the investigation of the ICT adoption. **The primary data source is the World Telecommunication/ICT Indicators Database, augmented by the UN E-Government Survey and UNESCO Institute for Statistics (UIS) database.** The list of variables, their detailed definitions, and sources are described in Appendix I. Three-digit ISO codes of the 193 countries, based on data availability and their analytical groupings based on income, are presented in Appendix II. Throughout the paper all estimations are done using R software and averages for various analytical groups are calculated as the weighted averages using as weights PPP GDP shares based on the World Economic Outlook database (IMF, 2019a).

12. **Unsupervised machine learning algorithms infer patterns from a dataset without imposing labels.** Therefore, unlike their supervised equivalents, they cannot be directly applied to a regression or a classification problem. Nevertheless, they are useful to help discover the underlying structure of the data and thus they are often performed as part of an exploratory data analysis. Two of the main techniques implemented in unsupervised learning are principal component and cluster analysis. The latter serves to group or segment datasets with shared attributes in order to extrapolate algorithmic relationships. This technique identifies commonalities in the data and, as more data is brought into the analysis, reacts to the presence or absence of such commonalities in each new piece of data.

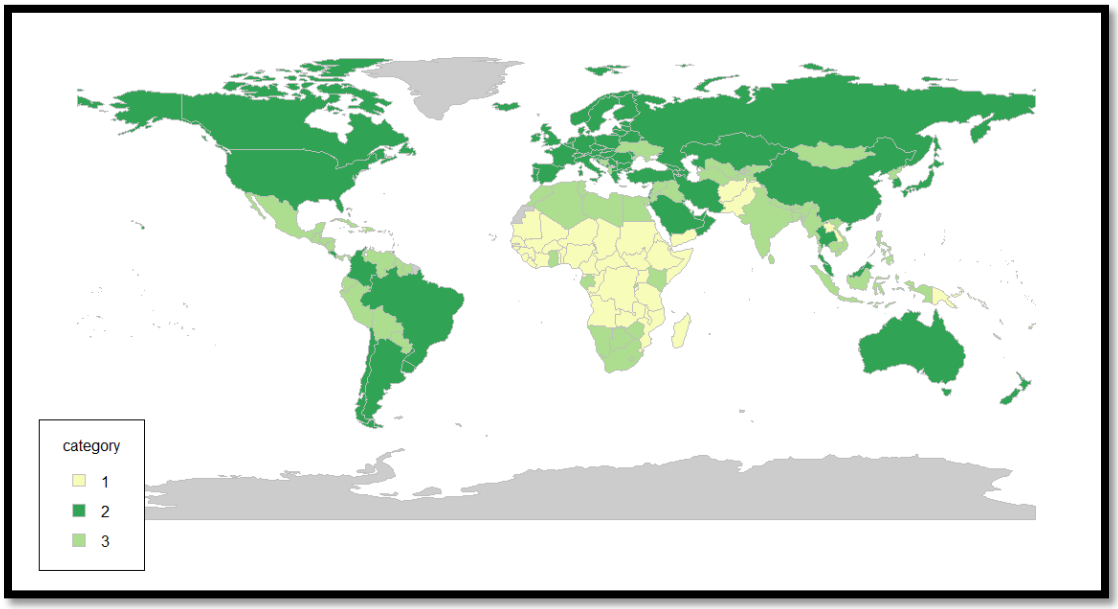
13. **We implement one of the most popular unsupervised learning clustering algorithms is K-means clustering.** It aims to divide *n* observations into *k* distinct, non-overlapping clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster (James and others, 2013). We note that the K-means algorithm finds a local rather than a global optimum, thus the results depend on the initial (random) cluster assignment of each observation. Therefore, we run the algorithm 100 times for different random initial configurations and select the best solution, i.e. that minimizes equation 1 in Appendix III. The algorithm requires the specification of the number of clusters. We use the elbow method (Thorndike, 1953), average silhouette technique (Rousseeuw, 1987) and gap statistics (Tibshirani, 2000) in deriving the proper K.[4,5]

[4] We also considered hierarchical clustering which yielded similar results. Specifically, we implemented the Divisive Hierarchical Clustering due to its superior properties in identifying relatively large clusters. We prefer the K-means algorithm since the hierarchical clustering methods are subject to arbitrary decisions of selecting both the distance metric and the linkage criteria, the time complexity of at least $O(n^2 \log(n))$, where n is the number of data points, as well as their sensitivity to noise and outliers.

[5] The elbow method is a visual approach to choose a number of clusters such that adding another one does not lead to a significant increase in the ratio of within-group to total variance explained. The average

(continued…)

Figure 1. Digital Connectivity: country groupings under K=3 clusters



Sources: ITU's ICT Indicators database, UN's E-Government Survey, UNESCO's UIS and authors' calculations.

14. **The K-means algorithm broadly groups countries into three general classes (Figure 1).**[6] The K-means unsupervised machine learning algorithm for the ICT variables for the world, taking the optimal number of clusters to be equal three, results in the grouping presented in Figure 1. The robustness checks for different cluster specifications (K equaling 2 and 4) yields similar results. The list of countries and their clusters are included in Appendix IV.

15. **PCA analysis reveals that digital connectivity can be quantified by a composite, quasi-linear index.**[7] We implement PCA on the available ICT dataset and then plot the above-mentioned groups as the function of the first two components[8] (Figure 2). The results suggest that the heterogeneity in digital connectivity can be largely explained by the differences in values of the first principal component, which in turn explains nearly half of the sample variation. Due to the linearity of the executed

silhouette technique is another visual application, which aims to maximize the within-cluster similarity, while simultaneously maximizing the across-cluster dissimilarities. Finally, the gap statistics is a statistical method that maximizes the total within intra-cluster variation for different number of clusters with their expected values under null reference distribution of the data.

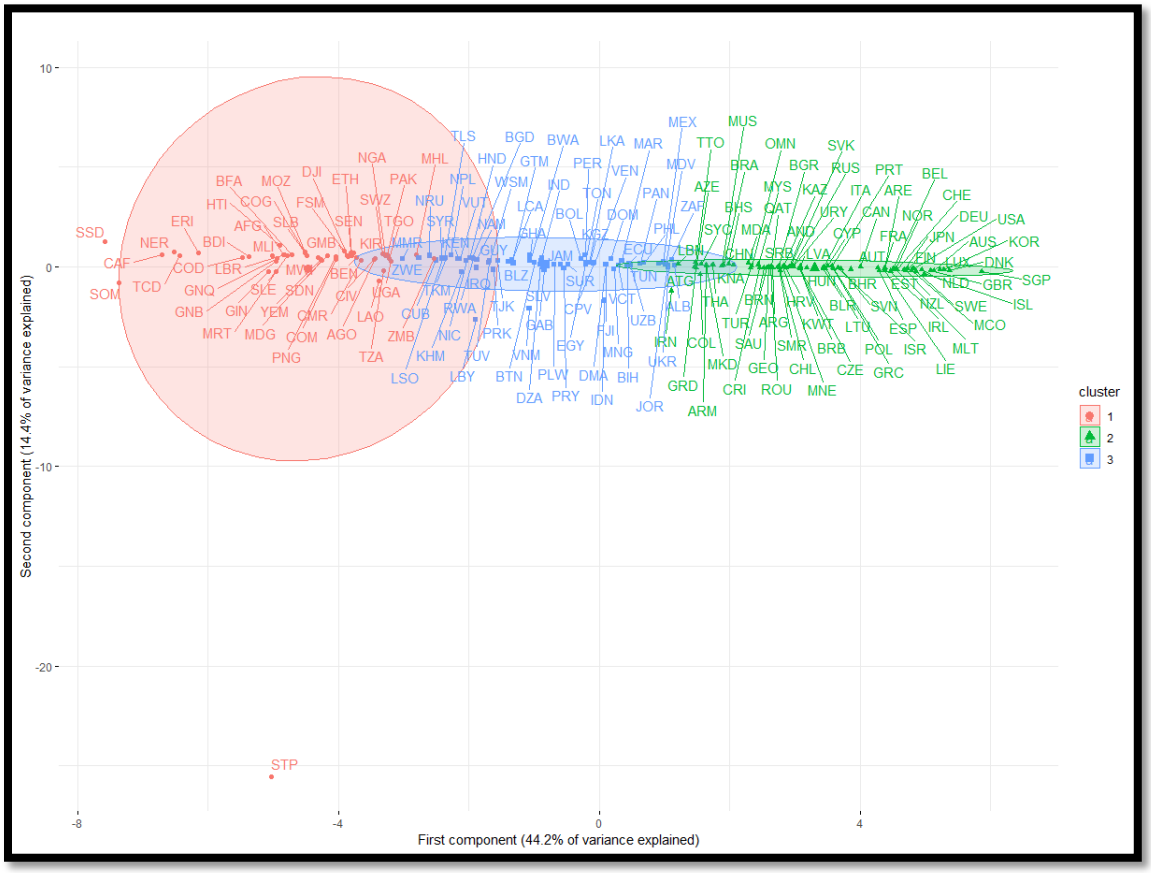
[6] The assigned colors in Figure 1, do not represent any intensity or qualitative nature of the groupings.

[7] Intuitively, PCA can be regarded as a statistical procedure to reveal the internal structure of the data in a way that best explains the variation in the data within a multivariate context. It provides a lower-dimensional picture by considering only the first few principal components to reduce the dimensionality of the transformed data.

[8] First component can be viewed as a linear transformation of indicators explaining the largest variation. The second component is still a linear transformation, orthogonal to the first one, explaining the largest portion of residual variation.

dimension reduction method, the results justify the quasi-linear composite index of digital connectivity, introduced in the next section.

Figure 2. K-means clustering under K=3, PCA



Sources: ITU's ICT Indicators database, UN's E-Government Survey, UNESCO's UIS and authors' calculations.

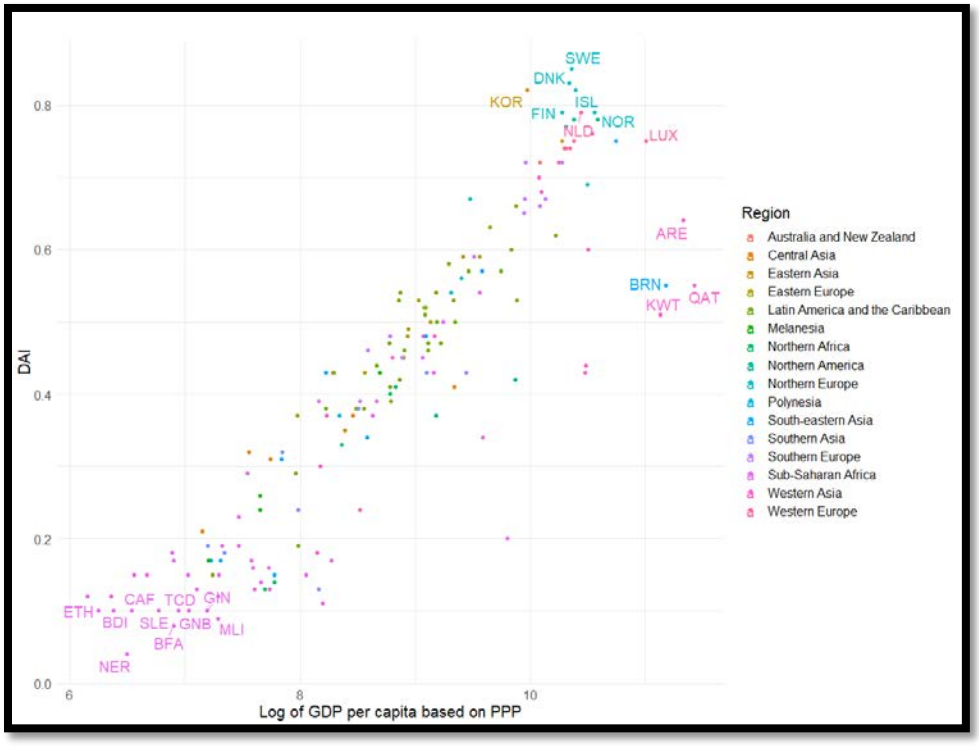
III. Constructing an Index of Digital Connectivity: EDAI

16. **In this section we describe the methodology for constructing the Enhanced Digital Access Index (EDAI) and its five sub-indices.** In addition to the results for various analytical country groupings, we check for the robustness of the outcomes by applying different weighting schemes: equal, gradual and Wroclaw. Furthermore, we present the radar graphs for several country groupings, such as oil exporters or economies in fragile situations to identify the main strengths and weaknesses in digital connectivity to inform potential areas for improved ICT adoption. The EDAI is constructed following the methodology of the Digital Access Index, introduced below which takes a holistic approach reflecting the highly complementary usage of modern technologies. We use the same data source in this section as in the unsupervised machine learning section (Appendix II).

A. Digital Access Index (DAI)

17. **DAI was launched by the International Telecommunication Union in 2003 to inform the ability of each country's population to take advantage of ICT.**[9] It was calculated for 178 economies, as a composite score of eight variables describing five categories: availability of infrastructure, affordability of access, educational level, quality of information and communication technology services, and internet usage. The variables measure access to and usage of ICT as well as education level of the population. Each variable is converted to an indicator with a value between zero and one by dividing it by the maximum value or "goalpost". Each indicator is then weighted within its category and the resulting category index values are averaged to obtain the overall DAI value. Each category is of equal importance, although some variables within categories are assigned unequal weights. The overall weighting scheme is presented in Appendix III.

Figure 3. DAI distribution



Sources: ITU's ICT Indicators database, UN's terminology database, and authors' calculations.

[9] An alternative index with the same acronym is the Digital Adoption Index of World Bank (2016). This index is based on three sectoral sub-indices covering businesses, people, and governments, with each sub-index assigned an equal weight: DAI (Economy) = DAI (Businesses) + DAI (People) + DAI (Governments). Each sub-index is the simple average of several normalized indicators measuring the adoption rate for the relevant groups.

18. **Based on the constructed DAI in 2003, SSA countries lag other regions in digital access (Figure 3).** Higher values of DAI are positively related to the Purchasing Power Parity (PPP) based Gross Domestic Product (GDP) per capita. Based on 2003 values, with the exception of Canada, ranked 10th, the top ten economies are exclusively Asian and European, while the lowest ranked economies in terms of ICT adoption were from the SSA.

B. Enhanced Digital Access Index (EDAI)

19. **The Digital Access Index (2003) has several shortcomings to assess the current global digital divide.** First, it arbitrarily imposes equal weights during aggregation. This implies a strong assumption that all indicators within each sub-index are perfectly substitutable and similarly that all sub-indices are fully substitutable. This assumption may have weak theoretical justification given the disparity among the sub-indices (for instance educational level and infrastructure). Second, availability of indicators that define the digital connectivity has expanded since the launch of the DAI in 2003 and thus the Index is currently informed by a constrained set of variables. Finally, DAI is outdated and constructed for a slightly smaller number of countries.

20. **We propose a new composite index to measure digital connectivity: EDAI.** In the same spirit of the methodology of DAI, we consider five sub-categories of digital connectivity: availability of infrastructure, affordability of access, educational level of the population, quality of information and communication technology services, and internet usage. We augment the DAI with recently available indicators included in the ICT Development Index of ITU, and the indicators from the Digitization Index (Katz and others, 2014) for a larger set of countries. We then rescale all indicators to a [0, 100] interval, following the methodology of Inclusive Internet Index, by Facebook and Economist, through the following transformation:

$$X_{New} = \frac{X_{Old} - min(X_{Old})}{max(X_{Old}) - min(X_{Old})} \cdot 100$$

21. **Using the rescaled variables and the five sub-indices, we construct EDAI values for each country by the Mazziotta-Pareto Index (MPI) methodology.** An advantage of the MPI aggregation methodology is that it avoids artificially imposing equal weighs used in constructing the DAI as well as World Bank's Digital Adoption Index, or the progressive scheme used in constructing the Inclusive Internet Index. This technique is based on a quasi-linear function that introduces a penalty for the units with unbalanced values, starting from the arithmetic mean of the normalized indicators.[10] To be more precise, the composite index is given by a similar rescaling to [0, 100] interval:

[10] This implies that for any country, each lagging value of any sub-index would act as a bottleneck and therefore reduce the EDAI.

$$\text{MPI}_i = M_i \cdot (1 - CV_i^2) = M_i - S_i \cdot CV_i$$

where M_i, CV_i, S_i denote respectively the *i*-th country mean, coefficient of variation, and its standard deviation. MPI is designed to normalize the indicators by a specific criterion that (i) normalizes values by a specific criterion and hence deletes the unit of measure and the variability effects; (ii) provides the synthesis independent from an *ideal unit*; and (iii) simplify the computations.[11] Country specific values for EDAI and its sub-indices are in Appendix VI.

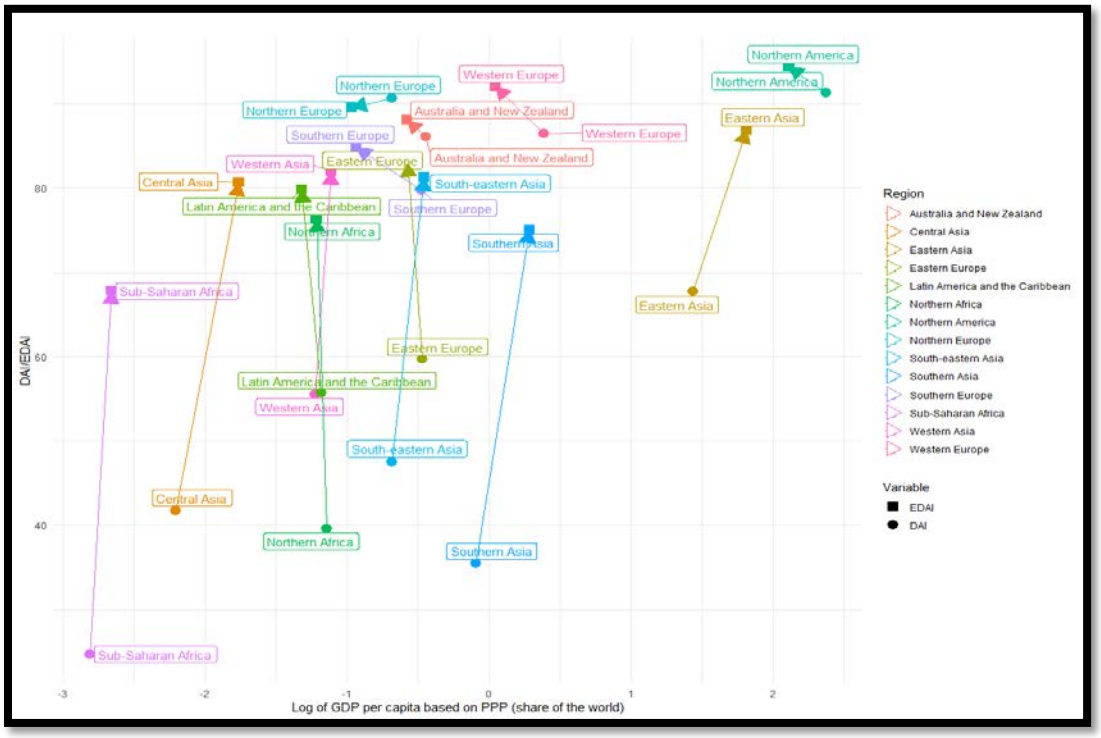
22. **The EDAI index provides further evidence of a global digital divide, but compared to 2003, the gap seems to be narrowing (Figure 4).[12]** Keeping in mind caveats regarding differences in the aggregation method and the indicator set, we note the worldwide progress in terms of digital connectivity over the last 15 years. The diffusion of technology is clearly a global phenomenon that has witnessed a steep increase in the number of people connected, rather than remaining a privilege held by a few wealthy nations. SSA countries still lag other country groupings with the high digital connectivity achieved by Western European economies. Within the SSA, countries including Cabo Verde, Mauritius, Seychelles, and South Africa remain at the top in digital connectivity, while many other nations continue to lag. We therefore view these results as providing further evidence in favor of the digital divide hypothesis, in line with the results from the unsupervised machine learning algorithm. Nevertheless, the digital divide seems to be narrowing: the SSA distance-to-frontier (DTF) in 2003 was equal to 45.1, while it reduced to 17.3 in 2017.[13] Reliance on DTF to provide evidence in favor of narrowing digitalization gap mitigates caveats due to differences in coverage, methodology, and variables in calculating the DAI and EDAI.

[11] It is important to provide the synthesis independent from the "ideal unit" because a set of "optimal values" is arbitrary, non-univocal and would potentially vary over the time. See De Muro and others (2011).

[12] This is in contrast to the key findings of the Inclusive Internet Index, commissioned by Facebook and conducted by the Economist Intelligence Unit that finds the digital divide to be widening at the bottom of the income pyramid.

[13] The values of DAI for the world, normalized through the identical transformation as in the EDAI derivation and then averaged with the GDP per capita as the share of the world weights were compared to the analogous outcomes for the EDAI: $DAI_{world} = 69.8$, $DAI_{SSA} = 24.7$, $EDAI_{World} = 85.1$, $EDAI_{SSA} = 67.8$. For more information on the distance-to-frontier consult Distance to Frontier and Ease of Doing Business Ranking (2017).

Figure 4. Evolution of Digital Connectivity: from DAI to EDAI

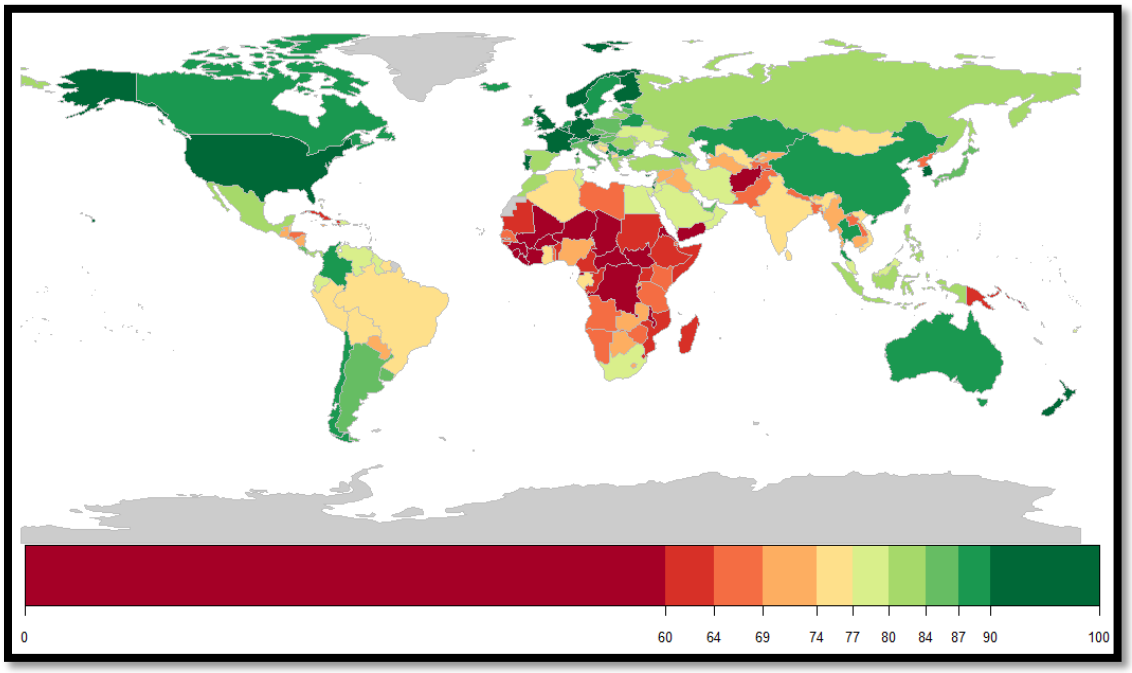


Sources: ITU's ICT Indicators database, UN's E-Government Survey, UNESCO's UIS and authors' calculations.

23. **Heterogeneity in digital connectivity is related to income and geography.** We group the EDAI values in deciles and plot them on the world map (Figure 5). North America, Europe, Western Asia, and Australia and Oceania rank highest in digital connectivity, while SSA and LIDCs in general lag in digital connectivity. This outcome is consistent with the observation that geographic location is in fact fundamentally important to issues concerning the digital divide (Grubesic and Murray, 2005) This provides evidence against the hypothesis that advances in ICT will render geographical divides as irrelevant.[14]

[14] These factors motivate the use of income and regional dummies as control variables in the fractional logit regressions (see Section IV).

Figure 5. EDAI world distribution

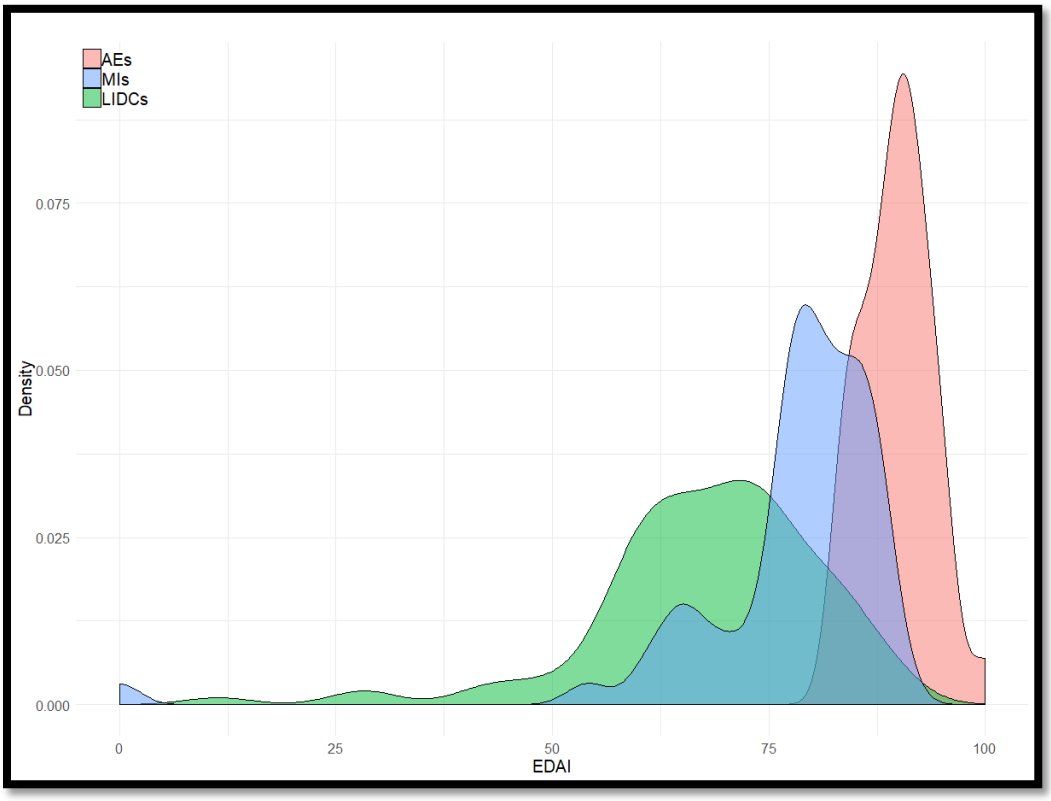


Sources: ITU's ICT Indicators database, UN's E-Government Survey, UNESCO's UIS and Authors' calculations.

24. **EDAI distribution differs between AEs, MICs, and LIDCs (figure 6).**[15] Figure 6 illustrates that the AEs perform better in terms of digital connectivity relative to the MICs, which in turn perform better relative to the LIDCs. Moreover, the distribution is much more dispersed for LIDCs.

[15] Income based country groupings are from IMF(2019c) and are in Appendix IV.

Figure 6. EDAI distribution across AEs, MICs and LIDCs



Sources: ITU's ICT Indicators database, UN's E-Government Survey, UNESCO's UIS and Authors' calculations.

25. **The digital divide is more prominent for the individual components of EDAI**. Table 1 presents digital connectivity for the world, AEs, MICs, and LIDCs, and country values within each group, weighed by their PPP based GDPs. AEs lead in every single digital connectivity sub-index, with most prominent divergences arising in Knowledge and Internet Usage categories.

Table 1. Sub-indices values

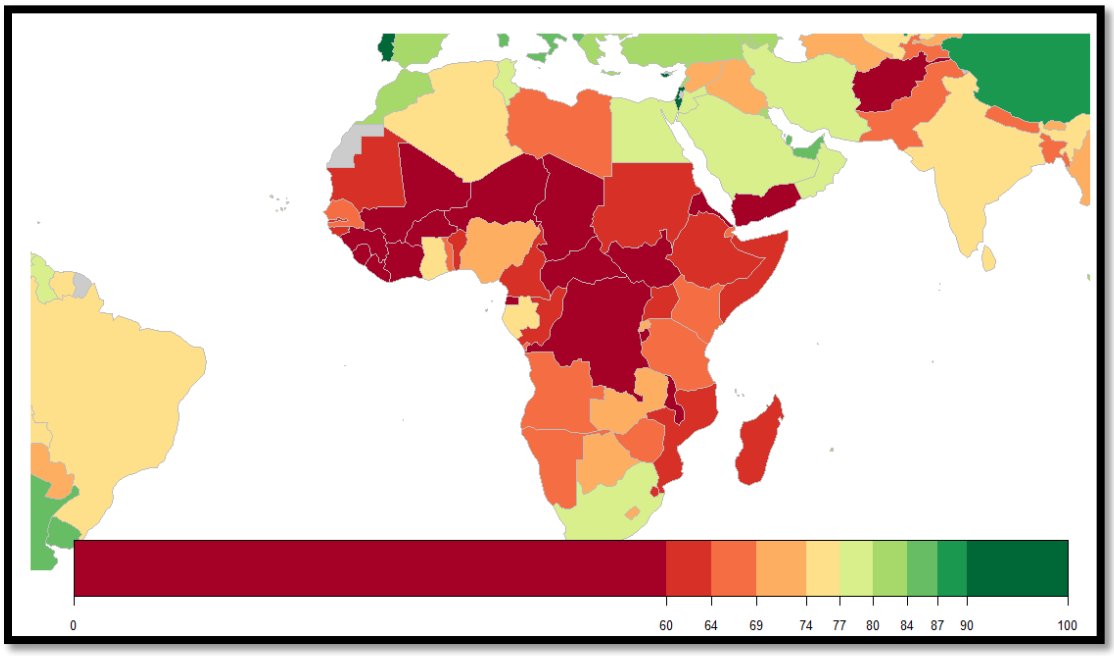
Sub-Index	World	AEs	MICs	LIDCs
Infrastructure	80.23	87.83	78.11	54.64
Quality	25.21	31.26	22.06	15.62
Affordability	20.47	22.55	19.18	18.65
Knowledge	78.99	92.06	73.54	48.28
Internet Usage	60.94	84.80	48.28	24.67

Source: Authors' calculations.

C. EDAI for SSA

26. **SSA exhibits heterogeneity in digital connectivity.** The digital divide literature tended to focus on the differences between Africa and the industrialized world, with inadequate attention to the heterogeneity in the region (Onyeiwu (2002)). Figure 7 focuses on the deciles for SSA only. It indicates the heterogeneity within SSA, with countries like Botswana, Cabo Verde, Gabon, Ghana, Lesotho, Mauritius, Rwanda, Seychelles, and South Africa ranking highest in the region in digital connectivity.[16]

Figure 7. EDAI SSA distribution



Sources: ITU's ICT Indicators database, UN's E-Government Survey, UNESCO's UIS and authors' calculations.

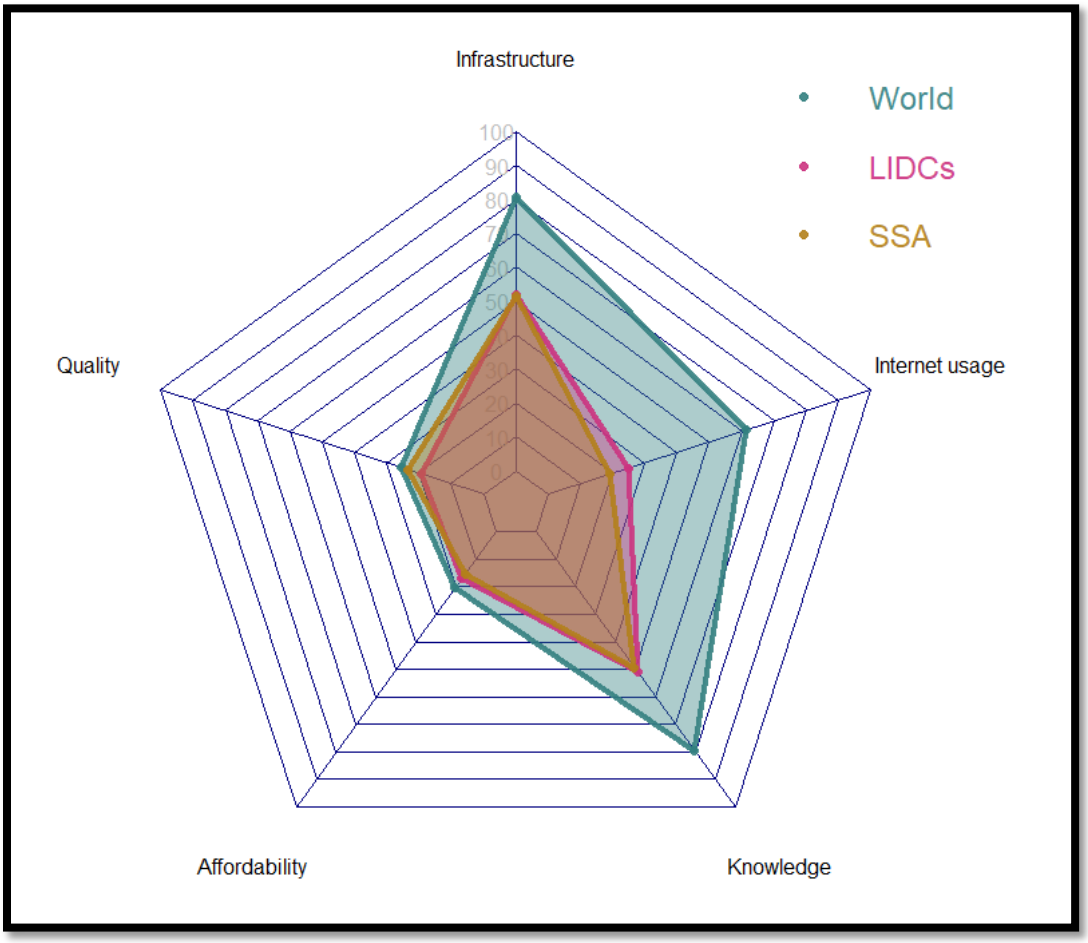
27. **In terms of sub-indices, SSA lags other regions mainly in digital infrastructure, internet usage, and human capital.** Figure 8 presents the comparison of SSA to the world and LIDCs. We observe that digital connectivity in SSA is similar to the LIDC average. In terms of the five dimensions, SSA and LIDCs are close to the rest of the world in terms of quality (speed of connection) and affordability but lag in infrastructure, internet usage, and knowledge. The comparable outcome for affordability in SSA relative to the rest of the world reflects similar SMS and internet prices in US$

[16] These are the nine countries from SSA that have EDAI values above 70. The median for the world is 78 and there are four countries from SSA above the median: Cabo Verde, Mauritius, Seychelles, and South Africa.

(continued…)

across countries[17]. Likewise, the analogous outcomes for the quality sub-index are due to the variables that measure maximum download speeds, relatively alike globally.

Figure 7. EDAI sub-indices



Sources: ITU's ICT Indicators database, UN's E-Government Survey, UNESCO's UIS, and Authors' calculations.

28. **Income differences can only partially account for the observed heterogeneity in digital connectivity across SSA countries.** Countries with similar socio-economic backgrounds continue to diverge in digital connectivity.[18] For example, Liberia and Lesotho, with comparable GDP PPP per capita and with relatively low rankings in the UN's Human Development Index, possess considerably different EDAI

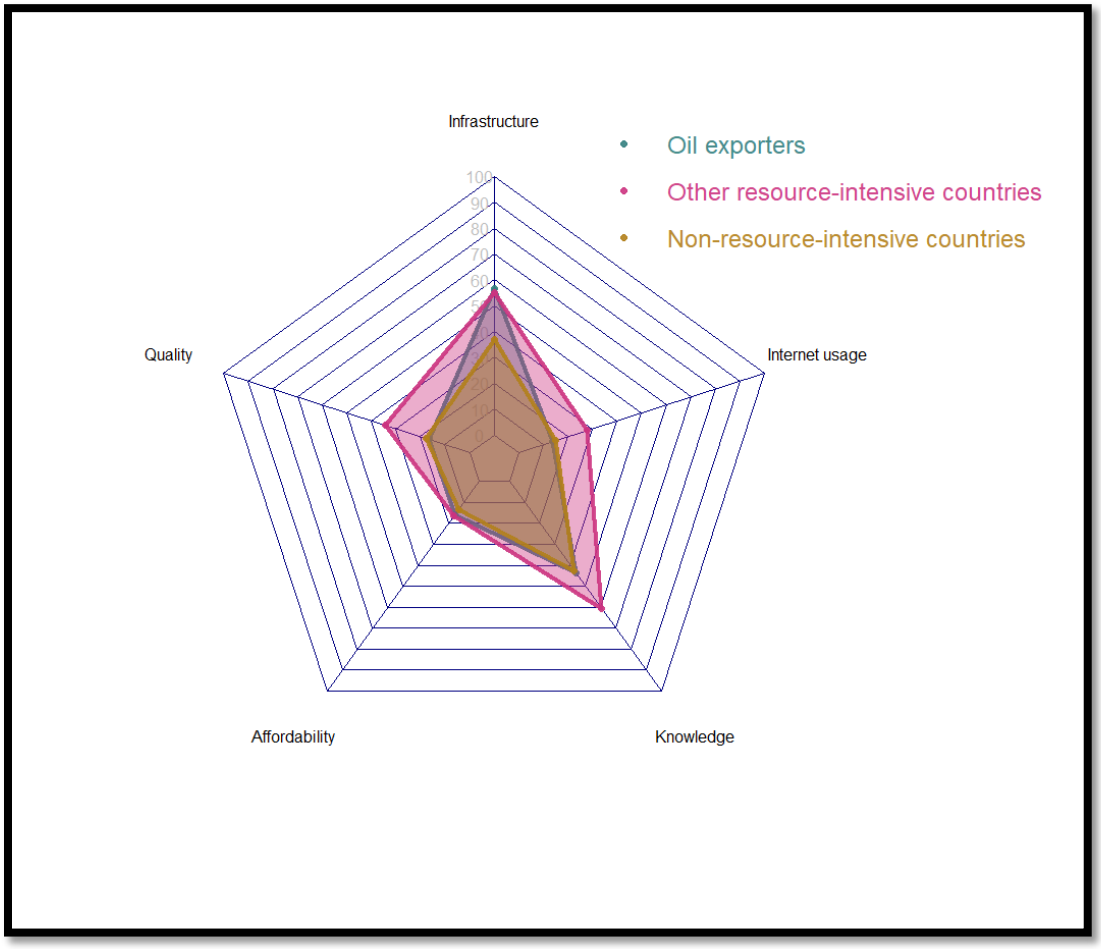
[17] Figure 7 shows that the cost of digital connectivity is similar in SSA compared to the rest of the world. However, relative to per capita income, affordability is an issue for SSA. Indeed, Abdychev and others (2018) and IMF (2019c) note the cost of a fixed broadband connection is the highest in sub-Saharan Africa compared to other regions. "Affordability" sub-index used in the aggregation of the EDAI goes beyond the broadband costs and internet penetration and also include factors such as the price of SMS and the price per minute of a peak rate call, see Appendix I.

[18] This was first noted for SSA countries by Onyeiwu (2002).

values (50 and 70 respectively). The aforementioned Enhanced Digital Access Index difference is comparable to the discrepancies between Japan (85) and North Korea (65) or among USA (95) and Ghana (75)).

29. **Economic structure can also partially explain heterogeneity in digital connectivity.** Non-resource intensive economies (mostly agricultural and commodity exporters) lag oil-exporters and other resource-intensive exporters in terms of digital connectivity, specifically quality, internet usage, and knowledge (Figure 8). This may reflect insufficient funds to invest in digital infrastructure in non-resource intensive economies. Similarly, countries classified as experiencing fragility, as defined by violence and political instability, lag in terms of knowledge.[19] This underscores the potential gains in connectivity by raising education levels in these economies.

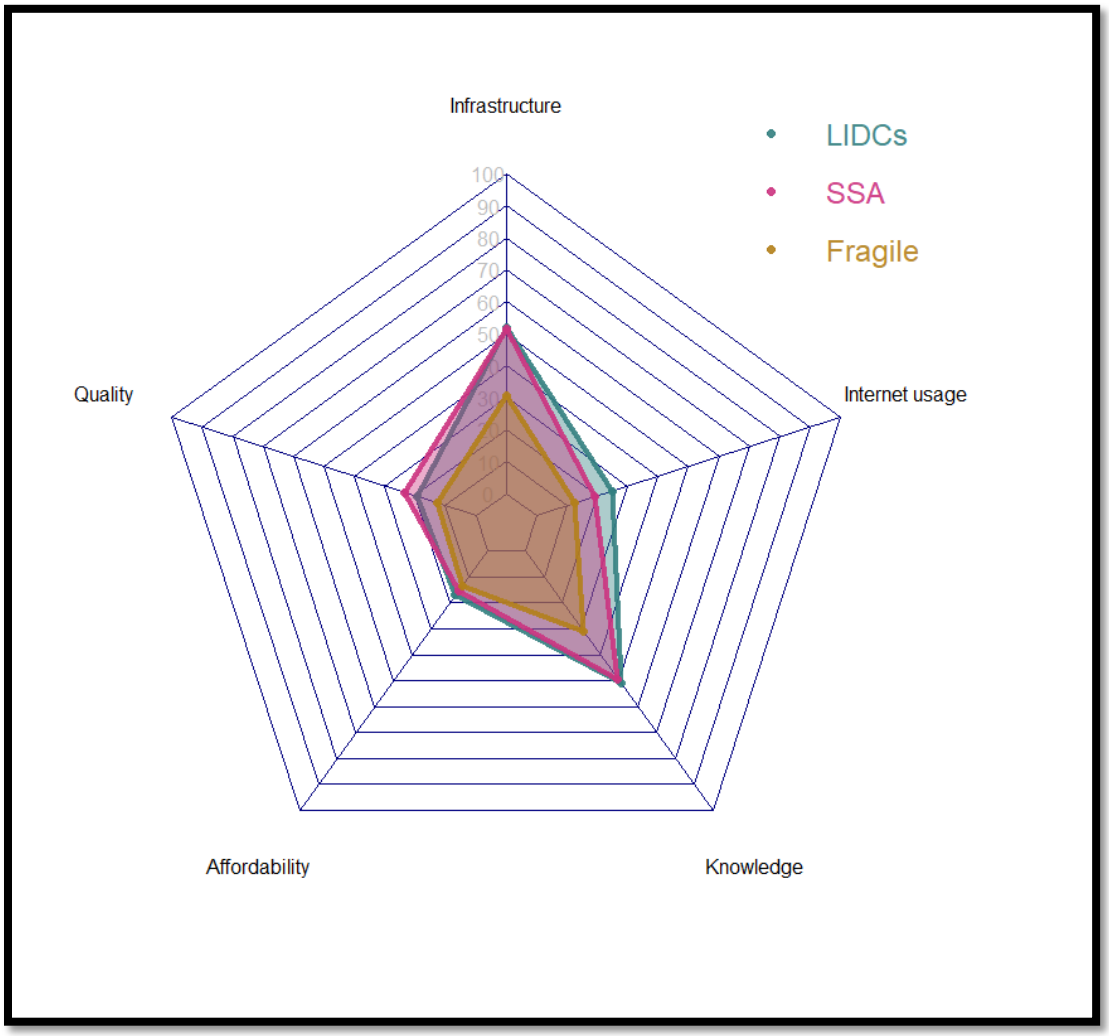
Figure 8a. EDAI for different groups of SSA economies



Source: ITU's ICT Indicators database, UN's E-Government Survey, UNESCO's UIS, and Authors' calculations.

[19] SSA country classifications are from IMF (2019c).

Figure 8b. EDAI for different groups of SSA economies



Source: ITU's ICT Indicators database, UN's E-Government Survey, UNESCO's UIS, and Authors' calculations.

IV. Estimating the Drivers: Fractional Logit Regressions

30. **We use fractional logit regressions to explore the variables associated with digital connectivity.** This method acknowledges the fractional nature of the dependent variable, can be employed for both discrete and continuous variables, and is capable of handling the extreme values of 0 and 1 without having to manipulate the data ((Papke and Wooldridge, 1996; Baum, 2008; and Mullahy, 2010). Moreover, fractional logit models allow one to capture non-linear relationships, particularly when the outcome variable is near 0 or 1 (Ramalho and others, 2011). The description of the model can be found in Appendix VIII.

31. **We implement step-wise regressions in light of the large number of potential explanatory variables.** The methodology consists of iteratively adding and removing regressors to find a subset of variables resulting in the best performing model. Performance is based on the model which minimizes the quasi-Akaike Information Criterion. Stepwise regression is useful for high-dimensional data containing multiple predictor variables. Alternative methods were also considered, such as penalized regression (ridge regression, lasso regression, elastic net) and principal components-based regression methods (Principal Component Regression, PCR, and Partial Least Squares, PLS). Nonetheless, penalized regression method can select variables correlated with each other, which may reduce interpretability (Takada and others, 2018). Similarly, the principal components options can be an effective tool for reducing dimensionality in problems where many variables are measured, particularly when there are strong linear relationships among the variables. Nonetheless, to interpret the principal components, one must filter through the coefficients (or loadings) of the linear combinations and identify patterns. This can be quite challenging in problems with many variables, which is precisely when principal components are in fact most helpful (Chipman and Gu, 2005). Thus, to facilitate interpretation, we implement the stepwise fractional logit regressions.[20]

32. **The selected explanatory variables (127 in total) can be classified under 18 broad thematical groups.** These variables are chosen with a view to provide an expansive coverage of potential factors related to digital connectivity. These include indicators related to the ease of doing business and rule of law, based on the prior literature that emphasized the role of national governments in Africa in framing ICT sector policies for investment, privatization, deregulation, and providing access in underserved areas (Sarkar and others, 2015). Furthermore, we include variables related to demographics, employment, education, and health.

Table 2. Data for Fractional Logit Regression

Variables' group	Count
Balance of Payments	9
Climate	3
Corruption, transparency and country risk	8
CPIA	16
Debt statistics	6
Demographics	2
Ease of doing business	19
Education	5
Employment	7
Financial access	5
Fiscal	6
Geography	4
Health	11
Logistics	5
Macro Indicators	5
National accounts and real sector	9
Social Development	1
Urbanization	6
Grand Total	127

Sources: Available in Appendix VII and authors' calculations.

[20] A potential extension would be to implement decipherable penalized regression method, such as Bootstrapped Ridge Regression (Lenert and Walsh, 2018).

33. **We conducted a precursor check on bivariate correlations.** We considered absolute values of pair-wise correlations. If two variables are found to be highly correlated (with absolute correlation coefficient over 0.9) then we calculated mean absolute correlation of each variable with the others and removed the one with the highest mean absolute correlation.

34. **We first estimate stepwise fractional logit model for the full sample (Table 3, column 3).**[21, 22] Through minimizing the quasi-Akaike Information Criterion, the total number of explanatory variables is reduced from 110 to 14.[23] Similar to Sarkar and others (2015), our empirical results provide evidence in favor of the role of policies for digital connectivity. In the full sample regression, better business enabling and regulatory environment with higher tax revenue yield are associated higher digital connectivity as is higher share of renewable energy share in total energy production. Moreover, higher urban access to electricity, and in general urbanization as well as lower dependency on remittances seem to matter for digital connectivity while controlling for income per capita and digital connectivity.[24] Private consumption is also positively related to digital connectivity likely capturing affordability and the availability of smart phones and personal computers to access to internet. Finally, ICT adoption seems to be related to the overall level of development, proxied by the proportion of remittances in the GDP, access to electricity and renewable energy consumption. The logit estimated coefficients can be transformed to identify as changes in odds. Assuming away the potential endogeneity issue, for the full sample, a decrease of 1 percentage point in share of rural population leads to the $exp(-0.656) - 1$, i.e., 0.48 odds increase in digital connectivity. Average share of rural population in the world is about 40 percent compared to 21 percent average in AEs. If the world average is to halve to 20 percent, i.e., a 20 percentage points decline, the odds of higher digital connectivity would rise by 0.09.

35. **The fractional logit regressions estimated for income-based country groupings reveals heterogeneity regarding the drivers of digital connectivity**. The results for AEs (Column 4) reveal that only account ownership is significantly related to digital connectivity, suggesting reforms to enhace financial access for the AEs, such

[21] In all regressions, PPP GDP per capita is used as a control variable. Additionally, for the world regression, we also impose 16 regional dummy variables as further control variables.

[22] Potential endogeneity is an issue that we are not able to address due to lack of a "good" instrument for the purposes of this study and the cross-sectional nature of the data used in estimations. In that sense, in the preceding analysis, we refrain from attributing causation and emphasis on the magnitudes of the coefficients. We rather focus on the strength of the correlations as well as the sign of the coefficients. With regular data availability across time, panel and distributed lag models could be considered as a valuable extension of further work in this area.

[23] We use the CPIA variables (16) only in the SSA regression. Hence total number of variables used in the global regression is 110.

[24] We acknowledge that the remittances as the percentage of GDP, as well as the renewable energy share as the percentage of total energy consumption may in fact capture the effect of variables not included in the analysis, being for instance a proxy for general level of development.

as further promotion of FinTech industry or full digital financial transformation through mobile banking apps, mobile money or e-wallets Conversely, estimation results for MICs (column 5) emphasizes the positive association with better regulatory and business enabling environment, better logistics, and a higher tax revenue capacity. Finally, regression results for LIDCs (column 6) underscore the importance of higher electricity access in cities as well as improved financial access and business facilitating environment for higher digital connectivity. All these interesting results require further exploration. The lower dependence on remittances and higher private consumption expenditures could reflect better affordability in form of access to smart phones and personal computers for internet usage as mentioned earlier.

36. **Results for the SSA-specific sample (column 7) further underscore the importance of a better business enabling and regulation environment, financial access, and urbanization** . This includes leveling the playing field for female entrepreneurs and investing in better government services and in people's health and providing better regulatory environment. Indeed, controlling for income per capita, higher percentage of population without postal services and lack of health regulations seem to adversely affect digital connectivity. Assuming away the potential endogeneity issue, for the SSA sample, a decrease of 1 percentage point in share of rural population leads 0.77 odds increase in digital connectivity. Average share of rural population in SSA is about 57 percent on average compared to 20 percent average in AEs. If the SSA average reduces to 21 percent, i.e., a 36 percentage points decline, the odds of higher digital connectivity would rise by 0.17. In the similar fashion, an increase of 1 percentage point in financial access as measured by account ownership leads to 3.1 odds increase in digital connectivity. Average share of account ownership is 41 percent in SSA compared to 95 percent in AEs. Hence if SSA account ownership were to improve to AE levels, odds of higher digital connectivity would rise by 1.67.

37. **The differences between regression results for LIDCs and SSA require further investigation.** Access to electricity and property registration are significant for LIDCs, but surprisingly not for SSAs. On the other hand, health regulation capacity, percentage of rural population, and population without postal services are significant SSAs, but not for LIDCs. This discrepancy could reflect for instance inclusion of MICs in the SSA sample or geographic differences among LIDCs not captured by the dummy variables.

Table 3. Fractional Logit Regression Results

Variable	Category	Enhanced Digital Access Index				
		World	**AEs**	**MICs**	**LIDCs**	**SSA**
Remittances (% of GDP)	Balance of Payments	**-1.803*****	10.166	**-5.399*****	**-1.038***	-1.485
		(0.540)	(8.500)	(1.881)	(0.609)	(1.338)
Urban access to electricity (%)	Urbanization	**1.176*****		**1.575****	**1.397*****	0.738
		(0.299)		(0.724)	(0.321)	(0.548)
Renewable energy share (% of energy consumption)	Climate	**0.399****	-0.390	**1.011****	-0.043	0.710
		(0.182)	(0.677)	(0.385)	(0.200)	(0.521)
Registering Property - Procedures (number)	Ease of doing business	**-0.033****	-0.023	-0.022	**-0.037***	-0.044
		(0.016)	(0.050)	(0.034)	(0.021)	(0.034)
Registering Property - Cost (% of property value)	Ease of doing business	**-1.504***	5.441	**-6.128*****	**-1.528***	**-3.522****
		(0.775)	(3.529)	(1.971)	(0.895)	(1.657)
Starting a Business - Time - Women (days)	Ease of doing business	**-0.005*****	0.009	**-0.006*****	**-0.003****	**-0.015*****
		(0.001)	(0.014)	(0.002)	(0.001)	(0.003)
Account ownership (% of population ages 15+)	Financial access	**0.767*****	**4.043***	0.373	**0.690****	**1.412*****
		(0.215)	(2.317)	(0.336)	(0.262)	(0.461)
Tax revenue (% of GDP)	Fiscal	**2.514*****	2.692	**2.642****	0.757	1.215
		(0.681)	(1.919)	(1.019)	(0.967)	(1.745)
Private consumption expenditure (% of GDP)	National accounts and real sector	**0.874*****	3.248	0.556	**0.758****	0.226
		(0.287)	(2.121)	(0.870)	(0.334)	(0.574)
Gross fixed capital formation (% of GDP)	National accounts and real sector	0.809	1.000	**2.230***	0.627	-0.517
		(0.498)	(3.400)	(1.259)	(0.541)	(1.044)
Services, value added (% of GDP)	National accounts and real sector	**0.856****	-1.654	1.195	0.573	1.276
		(0.390)	(2.542)	(0.918)	(0.453)	(0.834)
International Health Regulations capacity	Health	**0.240****	0.128	0.238	0.202	**0.477****
		(0.107)	(0.570)	(0.206)	(0.139)	(0.221)
Rural population (% of total)	Urbanization	**-0.656*****	0.546	-0.187	-0.461	**-1.488*****
		(0.220)	(0.944)	(0.378)	(0.308)	(0.503)
Population without postal services (% of total)	Logistics	**-0.342*****	71.496	**-1.285*****	-0.140	**-0.485****
		(0.127)	(61.660)	(0.301)	(0.149)	(0.197)
Observations		168	34	53	81	45

Note: *p<0.1; **p<0.05; ***p<0.01

The coefficients report changes in the odds ratio: value greater than 0 indicates increase in the odds ratio relative to the unconditional odds. Standard errors are reported in parentheses. All regressions control for per capita GDP in PPP terms, while the world regression is additionally controlled for the geographical regions. The variables definitions are enclosed in Appendix VII.

38. **Results from logit regression using only CPIA for SSA indicate importance of responsive governance**. Finally, we perform fractional logit regressions for the SSA sample using the Country Policy and Institutional Assessment (CPIA) indicators (See Appendix VII for the 16 indicators) while controlling for the per capita income to assess the variation in the EDAI for SSA. Among the 16 CPIA indicators, only the CPIA environmental sustainability rating indicator survive the stepwise regression estimation (Table 4).[25] With the exception of oil exporters, the environmental sustainability rating variable is robustly related to digital connectivity. This result likely reflects importance of responsive governance by the authorities and needs to be explored further.

Table 4. Fractional logit regression results for CPIA in SSA

	SSA	Oil Exporters	Other Resource Intensive Exporters	Non-resource Intensive Countries	Countries in Fragile Situations
Environmental sustainability rating	0.550***	2.305	0.278**	0.640***	0.573**
	(0.203)	(1.697)	(0.106)	(0.164)	(0.231)
Observations	45	7	16	22	17

Note: *p<0.1; **p<0.05; ***p<0.01

The coefficients report changes in the odds ratio: value greater than 0 indicates increase in the odds ratio relative to the unconditional odds. Standard errors are reported in parentheses. The regression is controlled for the per capita GDP in PPP terms.

V. CONCLUSION

39. **Digital connectivity is a key policy area to promote job creation and yield dramatic improvements in living conditions.** This is an especially relevant concerns for SSA, given that the region needs to generate 20 million jobs per year in the next two decades. In this paper, we provide an input into this debate by creating a global index of digital connectivity (EDAI) using more recent data and better fitting methodologies to assess the current stance of digital connectivity in SSA from a comparative perspective and main drivers. The EDAI can be used by policymakers to assess the level of preparedness for their countries for the Fourth Industrial Revolution.

40. **Our results indicate the existence of global digital divide and a substantial lag in connectivity in SSA.** Specifically, we find,

- Evidence in favor of global digital divide by clustering countries into three main groups;
- Significant heterogeneity in digital connectivity across different analytical country groupings based on income and geography;
- Descriptive analyses based on the EDAI suggests that the majority of SSA countries lag in digital connectivity, the exceptions include MICs such as Botswana, Cabo

[25] Environmental sustainability rating assesses the extent to which environmental policies foster the protection and sustainable use of natural resources and the management of pollution.

Verde, Gabon, Lesotho, Mauritius, Seychelles, and South Africa and LICDs such as Ghana and Rwanda;

- Among the five dimensions, SSA countries on average perform well in affordability and quality, but lag in infrastructure, internet usage and knowledge; and finally
- Fractional logit regressions underscore the importance of the business enabling regulatory environment for improved digital connectivity. Higher urbanization, financial access, share of investment and private consumption, share of renewable energy are also associated with digital connectivity
- Estimation results for SSA indicate that better business enabling and regulatory environment, financial access, urbanization, and availability of postal services are associated with higher digital connectivity. Specifically, we find that leveling the playing field for female entrepreneurs and reducing property registration costs are positively related to higher digital connectivity.

41. **Concluding, we acknowledge that the channels through which the variables affect connectivity, as well as the avenues to address endogeneity concerns need to be explored in further research**. The EDAI can allow the multi-year analysis, which should be regarded as the valuable extension in the future, with higher data availability.

REFERENCES

1. Abdychev, A., Alonso, C., Alper, E., Desruelle, D., Kothari, S., Liu, Y., Perinet, M., Rehman, S., Schimmelpfennig, A., Sharma, P., 2018. The Future of Work in Sub-Saharan Africa, International Monetary Fund African Department Paper, No. 18/18, Washington, D.C.

2. Baum, C.F., 2008. Stata tip 63: Modeling Proportions. The Stata Journal, 8(2), pp. 299–303.

3. Chinn, M.D. and Ito, H., 2006. What Matters for Financial Development? Capital Controls, Institutions, and Interactions. Journal of Development Economics, 81(1), pp.163–192.

4. Chipman, H.A. and Gu, H., 2005. Interpretable Dimension Reduction. Journal of Applied Statistics, 32(9), pp. 969–987.

5. Cullen, R., 2001. Addressing the Digital Divide. On-line Information Review, 25(5), pp. 311–320.

6. De Muro, P., Mazziotta, M. and Pareto, A., 2011. Composite Indices of Development and Poverty: An Application to MDGs. Social Indicators Research, Vol.104, No. 1, pp. 1–18.

7. Grubesic, T.H. and Murray, A.T., 2005. Geographies of Imperfection in Telecommunication Analysis. Telecommunications Policy, 29 (1), pp. 69–94.

8. Gruss, B., and Kebhaj, S., 2019. Commodity Terms of Trade: A New Database, IMF WP/19/21, Washington, D.C.

9. Hardin, J.W., and Hilbe, J.M. 2007. Generalized Linear Models and Extensions. Stata Press.

10. Hjort, J., and Poulsen, J. 2019. The Arrival of Fast Internet and Employment in Africa, American Economic review, 109(3), pp. 1032–79.

11. International Monetary Fund, 2018. Regional Economic Outlook. Sub-Saharan Africa: Capital Flows and the Future of Work. Chapter 3. The Future of Work in Sub-Saharan Africa, Washington, D.C., October.

12. _____, 2019a. World Economic Outlook: Growth Slowdown, Precarious Recovery. Statistical Appendix, Washington, D.C., April.

13. _____, 2019b. Fiscal Monitor: Curbing Corruption. Methodological and Statistical Appendix, Washington, D.C., April.

14. _____, 2019c. Regional Economic Outlook: Sub-Saharan Africa: Recovery Amid Elevated Uncertainty. Background Paper and Expanded Statistical Appendix, Washington, D.C., April.

15. International Telecommunication Union (ITU), 2003. ITU World Telecommunication Development Report: Access Indicators for the Information Society, Digital Access Index.

16. _____, 2017. ICT Development Index (IDI).

17. James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer.

18. Katz, R., Koutroumpis, P. and Martin Callorda, F., 2014. Using a Digitization Index to Measure the Economic and Social Impact of Digital Agendas. Info, 16(1), pp. 32–44.

19. Lenert, M.C. and Walsh, C.G., 2018. Balancing Performance and Interpretability: Selecting Features with Bootstrapped Ridge Regression. In AMIA Annual Symposium Proceedings (Vol. 2018, p. 1377). American Medical Informatics Association.

20. Mullahy, J., 2015. Multivariate Fractional Regression Estimation of Econometric Share Models. Journal of Econometric Methods, 4(1), pp. 71–100.

21. Norris, P., 2001. Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide. Cambridge University Press.

22. Onyeiwu, S., 2002. Inter-Country Variations in Digital Technology in Africa: Evidence, Determinants, and Policy Implications (No. 2002/72). WIDER Discussion Papers//World Institute for Development Economics (UNU-WIDER).

23. Papke, L.E. and Wooldridge, J.M., 1996. Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates. Journal of Applied Econometrics, 11(6), pp. 619–632.

24. Ramalho, E.A., Ramalho, J.J. and Murteira, J.M., 2011. Alternative Estimating and Testing Empirical Strategies for Fractional Regression Models. Journal of Economic Surveys, 25(1), pp.19–68.

25. Rousseeuw, P.J., 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, 20, pp. 53–65.

26. Sarkar, A., Pick, J.B. and Johnson, J., 2015. Africa's Digital Divide: Geography, Policy, and Implications.

27. Takada, M., Suzuki, T. and Fujisawa, H., 2017. Independently Interpretable Lasso: A New Regularizer for Sparse Regression with Uncorrelated Variables. ArXiv Preprint ArXiv:1711.01796.

28. Thorndike, R.L., 1953. Who Belongs in the Family?. Psychometrika, 18(4), pp. 267–276.

29. Tibshirani, R., Walther, G. and Hastie, T., 2001. Estimating the Number of Clusters in a Data Set via the Gap Statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), pp. 411–423.

30. United Nations, 2019. Global Indicator Framework for the Sustainable Development Goals and Targets of the 2030 Agenda for Sustainable Development.

31. World Bank Group, 2016. World Development Report 2016: Digital Dividends. World Bank Publications.

32. _____, 2017. Doing Business report 2017, Distance to Frontier and Ease of Doing Business Ranking, World Bank Publications.

33. _____, 2019. Doing Business Report 2019: Training for Reform. World Bank Publications.

APPENDIX I - Variables used in Enhanced Digital Access Index (EDAI)

Definition	Sub-Index	Source
Installation fee for residential telephone service refers to the one-off charge involved in applying for a basic residential fixed-telephone service. (US$)	Affordability	ICT
This indicator can include both; estimates and survey data corresponding to the proportion of households with computer. A computer includes: a desktop; portable or handheld computer (e.g. a personal digital assistant). It does not include equipment with some embedded computing abilities such as mobile phones or TV sets. The proportion of households with a computer is calculated by dividing the number of in-scope households with a computer by the total number of in-scope households.	Affordability	ICT
This indicator can include both; estimates and survey data corresponding to the proportion of households with Internet. The Internet is a world-wide public computer network. It provides access to a number of communication services including the World Wide Web and carries email; news; entertainment and data files. Access is not assumed to be only via a computer - it may also be by mobile phone; games machine; digital TV etc. The proportion of households with Internet access at home is calculated by dividing the number of in-scope households with Internet access by the total number of in-scope households.	Affordability	ICT
Fixed (wired)-broadband monthly subscription charge refers to the monthly subscription charge for fixed (wired)-; broadband Internet service. Fixed (wired) broadband is considered to be any dedicated connection to the Internet at; downstream speeds equal to; or greater than; 256 kbit/s. If several offers are available; preference should be given to; the 256 kbit/s connection. (US$)	Affordability	ICT
Mobile-cellular prepaid price of SMS refers to the price of sending a short-message service (SMS) message from a; mobile-cellular telephone with a prepaid subscription to a mobile-cellular number of a competing network (off-net). (US$)	Affordability	ICT
The price per minute of a peak rate call from a mobile cellular prepaid telephone to a mobile cellular subscriber of another (competing) network. Taxes should be included. If not included; it should be specified in a note including the applicable tax rate. (US$)	Affordability	ICT
Mobile-cellular prepaid connection charge is the initial; one-time charge for a new prepaid mobile-cellular subscription. Refundable deposits should not be counted. The connection fee corresponds usually to the price charged for the subscriber identity module (SIM) card; but may include other fees. It should be noted if free minutes; free SMS or other free services are included in the connection charge. (US$)	Affordability	ICT
Price of the plan; in local currency; for a mobile-broadband USB/dongle-based prepaid tariffs with 1GB volume of data. (US$)	Affordability	ICT
Fixed-telephone subscriptions 100 inhabitants	Infrastructure	ICT
Mobile-cellular subscriptions per 100 inhabitants	Infrastructure	ICT

Definition	Sub-Index	Source
Percentage of the population covered by a mobile-cellular network refers to the percentage of inhabitants within range of a mobile-cellular signal; irrespective of whether or not they are subscribers or users. This is calculated by dividing the number of inhabitants within range of a mobile-cellular signal by the total population and multiplying by 100.	Infrastructure	ICT
Percentage of the population covered by at least a 3G mobile network refers to the percentage of inhabitants that are within range of at least a 3G mobile-cellular signal; irrespective of whether or not they are subscribers. This is calculated by dividing the number of inhabitants that are covered by at least a 3G mobile-cellular signal by the total population and multiplying by 100.	Infrastructure	ICT
Percentage of the population covered by at least an LTE/WiMAX mobile network refers to the percentage of inhabitants that live within range of LTE/LTE-Advanced; mobile WiMAX/WirelessMAN or other more advanced mobile-cellular networks; irrespective of whether or not they are subscribers. This is calculated by dividing the number of inhabitants that are covered by the previously mentioned mobile-cellular technologies by the total population and multiplying by 100. It excludes people covered only by HSPA; UMTS; EV-DO and previous 3G technologies; and also excludes fixed WiMAX coverage.	Infrastructure	ICT
Active mobile-broadband subscriptions per 100 inhabitants	Internet Usage	ICT
Fixed broadband subscribers divided by population and multiplied by 100.	Internet Usage	ICT
This indicator can include both; estimates and survey data corresponding to the proportion of individuals using the Internet; based on results from national households surveys. The number should reflect the total population of the country; or at least individuals of 5 years and older. If this number is not available (i.e. target population reflects a more limited age group) an estimate for the entire population should be produced. If this is not possible at this stage; the age group reflected in the number (e.g. population aged 10+; population aged 15-74) should be indicated in a note. If no survey data are available at all; please provide an estimate specifying in detail the methodology that has been applied to calculate the estimate.	Internet Usage	ICT
Adult literacy is measured as the percentage of people aged 15 years and above who can,with understanding, both read and write a short simple statement on their everyday life	Knowledge	UNESCO (UIS)
E-Participation Index	Knowledge	EGDI
Online Service Index	Knowledge	EGDI
Expected years of schooling is the total number of years of schooling that a child of acertain age can expect to receive in the future, assuming that the probability of his or herbeing in school at any particular age is equal to the current enrolment ratio age	Knowledge	UNESCO (UIS)

Definition	Sub-Index	Source
Mean years of schooling (MYS) provides the average number of years of educationcompleted by a country's adult population (25 years and older), excluding the years spentrepeating grades	Knowledge	UNESCO (UIS)
Gross enrolment ratio is measured as the combined primary, secondary and tertiary grossenrolment ratio, of the total number of students enrolled at the primary, secondary and tertiarylevel, regardless of age, as a percentage of the population of school age for that level	Knowledge	UNESCO (UIS)
Fixed (wired)-broadband speed; in Mbit/s refers to the advertised maximum theoretical download speed; and not speeds; guaranteed to users associated with a fixed (wired)-broadband Internet monthly subscription.	Quality	ICT
International Internet bandwidth per Internet user (bit/s)	Quality	ICT
Advertised maximum theoritical download speed; and not speeds guaranteed to users associated with a 1GB USB/dongle-based postpaid plan.	Quality	ICT

Note: Selection of these variables is based on the following criterion: at least one observation for each variable should be available during 2014-17 for all countries. When a given economyhas more than one observation for a given variable, the latest data point is selected. Most of the observations are dated on 2016-17.

APPENDIX II - Country groupings

Global groupings by income (Fiscal Monitor) [26]

Advanced Economies

AUS, AUT, BEL, CAN, CHE, CYP, CZE, DNK, ESP, EST, FIN, FRA, DEU, GBR, GRC, IRL, ISL, ISR, ITA, JPN, KOR, LVA, LTU, LUX, MCO, MLT, NLD, NOR, NZL, PRT, SGP, SVK, SVN, SWE, USA

Emerging and Middle-Income Countries

AGO, ARE, ARG, AZE, BGR, BLR, BRA, BWA, CHL, CHN, CIV, COG, COL,CPV, DZA, DOM, ECU, EGY, GAB, GNQ, HRV, HUN, IDN, IND, IRN, KAZ, KWT, LBY, LKA, LSO, MAR, MEX, MUS, MYS, NAM, OMN, PAK, PER, PHL, POL, QAT, ROU, RUS, SAU, SMR, SRB, STP, SWZ, SYC, THA, TUR, UKR, URY, VEN, ZAF

Low-Income Developing Countries

AFG, ALB, AND, ARM, ATG, BEN, BDI, BFA, BGD, BHR, BHS, BIH, BLZ, BOL, BRB, BRN, BTN, CMR, CAF, COD, COM, CRI, CUB, DJI, DMA, ERI, ETH, FJI, FSM, GEO, GHA, GIN, GMB, GNB, GRD, GTM, GUY, HND, HTI, IRQ, JAM, JOR, KEN, KGZ, KHM, KIR, KNA, LAO, LBN, LBR, LCA, LIE, MDA, MDG, MDV, MHL, MKD, MOZ, MLI, MNE, MNG, MMR, MRT, MWI, NER, NGA, NIC, NPL, NRU, PAN, PLW, PNG, PRK, PRY, RWA, SEN, SLE, SLB, SLV, SOM, SDN, SSD, SUR, SYR, TCD, TGO, TJK, TKM, TLS, TON, TTO, TUN, TUV, TZA, UGA, UZB, VCT, VNM, VUT, WSM, YEM, ZMB, ZWE

Sub-Saharan Africa (SSA) groupings (Sub-Saharan African Regional Economic Outlook) [27]

Oil-exporting countries (SSA)

AGO, CMR, COG, GAB, GNQ, NGA, SSD, TCD

Other resource-intensive exporters (SSA)

BWA, BFA, GHA, NAM, NER, SLE, SOM, TZA, ZAF, ZMB

Non-resource-intensive exporters (SSA)

BEN, BDI, CIV, COM, CPV, ERI, ETH, GMB, GNB, KEN, LSO, MDG, MOZ, MUS, MWI, RWA, SEN, STP, SWZ, SYC, TGO, UGA

Countries in Fragile situations (SSA)

BDI, CAF, COM, COG, CIV, COD, ERI, GIN, GMB, GNB, LBR, MWI, MLI, SSD, STP, TCD, TGO, ZWE

[26] Classification from the IMF (2019b).

[27] Classification from the IMF (2019c).

APPENDIX III - K-means algorithm

Let $C_1, C_2, \dots, C_K$ denote the set of indices of observations in each cluster such that:

- ✓ Each observation belongs to at least one of the *K* clusters:

$$C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, n\}$$

- ✓ The clusters are not overlapping:

$$\forall_{k \neq k'} \, C_k \cap C_{k'} = \emptyset$$

K-means clustering aims to minimize the within-cluster variation:

$$min_{C_1,\dots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\} \tag{1}$$

where the within-cluster variation is defined through the squared Euclidean distance:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 \tag{2}$$

with $|C_k|$ denotying the power of set, ie. the number of observations in the *k*th cluster.

The algorithm of solving the above-mentioned problem may be described as follows:

Algorithm 1. K-means clustering

1. Randomly assign a number, from 1 to K, to each of the observations, as the initial cluster assignment.

2. Iterate until the cluster assignments stop changing:

a) For each of the K clusters, compute the cluster centroid (the vector of the p feature means for the observations in the kth cluster).

b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

APPENDIX IV – K-means clustering country groupings

Group 1

AFG, AGO, BDI, BEN, BFA, CAF, CIV, CMR, COD, COG, COM, DJI, ERI, ETH, FSM, GIN, GMB, GNB, GNQ, HTI, KIR, LAO, LBR, MDG, MHL, MLI, MOZ, MRT, MWI, NER, NGA, PAK, PNG, SDN, SEN, SLB, SLE, SOM, SSD, STP, SWZ, TCD, TGO, TZA, UGA, YEM, ZMB

Group 2

AND, ARE, ARG, ARM, ATG, AUS, AUT, AZE, BEL, BGR, BHR, BHS, BLR, BRA, BRB, BRN, CAN, CHE, CHL, CHN, COL, CRI, CYP, CZE, DEU, DNK, ESP, EST, FIN, FRA, GBR, GEO, GRC, GRD, HRV, HUN, IRL, IRN, ISL, ISR, ITA, JPN, KAZ, KNA, KOR, KWT, LBN, LIE, LTU, LUX, LVA, MCO, MDA, MKD, MLT, MNE, MUS, MYS, NLD, NOR, NZL, OMN, POL, PRT, QAT, ROU, RUS, SAU, SGP, SMR, SRB, SVK, SVN, SWE, SYC, THA, TTO, TUR, URY, USA

Group 3

ALB, BGD, BIH, BLZ, BOL, BTN, BWA, CPV, CUB, DMA, DOM, DZA, ECU, EGY, FJI, GAB, GHA, GTM, GUY, HND, IDN, IND, IRQ, JAM, JOR, KEN, KGZ, KHM, LBY, LCA, LKA, LSO, MAR, MDV, MEX, MMR, MNG, NAM, NIC, NPL, NRU, PAN, PER, PHL, PLW, PRK, PRY, RWA, SLV, SUR, SYR, TJK, TKM, TLS, TON, TUN, TUV, UKR, UZB, VCT, VEN, VNM, VUT, WSM, ZAF, ZWE

APPENDIX V – Digital Access Index (DAI) weighting scheme

Infrastructure

Fixed telephone subscribers per 100 inhabitants: **10%**

Mobile cellular subscribers per 100 inhabitants: **10%**

Affordability

Internet access price as percent of GNI x 100: **20%**

Knowledge

Adult Literacy: **13%**

Combined primary, secondary, and tertiary school enrollment: **7%**

Quality

International Internet bandwidth per capita: **10%**

Broadband subscribers per 1000 inhabitants: **10%**

Usage

Internet users per 100 inhabitants: **20%**

Appendix VI – Enhanced DAI (EDAI) and components

ISO	Infrastructure	Quality	Affordability	Knowledge	Internet usage	EDAI
AFG	25.95	0.87	6.63	34.17	7.82	58.40
AGO	52.98	20.37	18.00	47.30	9.38	65.06
ALB	85.84	39.32	19.03	77.70	43.58	86.05
AND	68.88	19.55	16.42	66.48	78.45	80.02
ARE	98.97	2.65	27.28	74.76	89.50	85.94
ARG	73.81	39.16	20.04	78.88	56.48	84.02
ARM	87.39	29.56	18.30	68.64	47.81	82.55
ATG	97.59	8.66	23.77	63.75	44.85	84.58
AUS	62.62	26.36	26.87	100.00	88.22	87.17
AUT	100.00	50.39	14.66	85.43	74.60	92.76
AZE	60.40	22.27	19.30	71.04	58.83	81.40
BDI	23.74	1.75	9.14	37.14	4.06	41.44
BEL	90.92	79.26	27.13	91.37	77.69	95.34
BEN	28.37	35.89	11.78	36.74	8.40	61.11
BFA	18.72	1.74	12.38	32.69	8.69	56.60
BGD	52.35	14.25	6.80	57.30	17.32	66.17
BGR	89.12	56.08	22.26	80.74	66.68	87.73
BHR	95.53	28.86	12.08	77.31	70.67	89.61
BHS	55.66	30.19	23.24	69.42	65.67	80.41
BIH	33.42	38.12	21.10	60.80	54.22	76.28
BLR	75.96	25.31	23.02	85.27	72.91	87.43
BLZ	68.79	1.21	24.20	50.52	23.30	73.92
BOL	70.49	48.38	22.63	63.38	34.00	77.04
BRA	50.47	5.23	21.19	77.38	50.88	77.01
BRB	98.83	25.44	23.08	77.39	62.91	87.20
BRN	68.93	8.83	24.80	72.09	64.84	83.92
BTN	73.29	4.45	0.00	47.68	28.61	68.86
BWA	77.29	2.43	28.26	42.67	31.95	71.44
CAF	9.85	9.72	13.32	23.50	2.27	48.65
CAN	94.03	7.65	32.69	88.53	81.18	87.58
CHE	79.17	10.78	27.10	85.92	91.56	89.76
CHL	83.66	16.98	33.77	83.31	64.29	87.16
CHN	86.39	25.01	17.07	77.46	61.86	87.00
CIV	34.85	2.23	17.72	25.47	29.35	53.99
CMR	12.94	36.78	19.10	46.48	13.97	61.88
COD	18.86	3.67	10.25	37.55	4.20	57.61
COG	47.28	26.12	13.10	43.29	4.27	60.40
COL	90.07	7.60	28.56	78.46	47.38	87.51

ISO	Infrastructure	Quality	Affordability	Knowledge	Internet usage	EDAI
COM	23.06	31.62	12.61	28.74	2.54	55.45
CPV	79.39	31.85	23.17	50.01	39.24	78.54
CRI	82.26	29.14	22.02	74.66	64.42	85.69
CUB	47.10	29.38	18.25	60.63	11.48	62.98
CYP	94.91	13.04	26.04	80.03	81.71	91.86
CZE	50.96	50.32	22.10	79.40	73.70	84.41
DEU	98.67	40.55	26.77	92.57	82.48	93.92
DJI	24.71	0.39	30.18	28.00	23.45	66.42
DMA	61.21	6.56	25.95	60.13	54.95	74.18
DNK	91.68	36.89	12.73	96.57	100.00	94.75
DOM	43.26	54.04	15.91	67.66	41.95	78.48
DZA	53.76	19.26	24.62	49.03	40.86	74.02
ECU	54.42	32.04	23.44	72.38	38.47	78.30
EGY	78.41	2.53	20.86	56.38	32.78	77.56
ERI	30.97	24.42	11.38	14.68	0.00	28.22
ESP	67.23	7.76	22.25	89.32	78.38	82.97
EST	95.60	15.43	19.05	88.97	88.21	88.86
ETH	35.81	0.00	10.66	39.11	9.83	61.23
FIN	87.47	54.19	26.06	96.86	90.72	91.13
FJI	84.53	10.56	21.77	60.55	31.32	79.60
FRA	75.78	46.12	19.99	89.64	88.11	90.11
FSM	11.08	10.31	29.22	48.57	13.89	59.57
GAB	64.90	2.05	30.36	50.28	37.38	76.67
GBR	95.67	16.85	22.56	94.91	88.96	89.84
GEO	93.89	21.41	17.31	76.73	54.73	88.09
GHA	33.40	70.40	16.48	59.38	31.71	74.78
GIN	28.64	24.34	13.08	25.12	7.36	59.30
GMB	54.59	20.41	14.35	30.06	8.97	62.75
GNB	57.07	5.75	8.94	29.53	2.93	62.86
GNQ	0.00	18.53	17.22	31.52	12.77	0.00
GRC	69.96	2.62	37.04	87.63	68.62	83.23
GRD	81.50	13.69	27.43	64.89	57.83	81.56
GTM	63.42	28.87	19.20	55.22	22.14	71.82
GUY	74.70	27.72	23.18	51.19	25.44	77.17
HND	61.10	20.38	12.13	56.29	18.94	68.16
HRV	60.80	4.34	18.76	78.38	67.02	79.09
HTI	35.07	22.57	10.99	40.59	8.15	61.88
HUN	73.10	36.72	22.49	79.94	60.97	85.36
IDN	86.04	16.33	24.54	62.39	32.03	81.54
IND	75.95	29.38	14.38	62.22	14.40	75.93

ISO	Infrastructure	Quality	Affordability	Knowledge	Internet usage	EDAI
IRL	62.83	15.67	20.99	94.42	78.14	83.94
IRN	81.31	7.25	34.30	68.58	49.77	79.89
IRQ	43.55	22.03	17.83	42.77	35.21	71.89
ISL	93.95	29.73	15.76	84.44	93.47	87.51
ISR	90.42	28.11	23.99	85.71	78.42	91.97
ITA	66.08	30.51	25.66	85.49	66.16	85.49
JAM	63.48	8.88	29.42	57.69	34.55	73.64
JOR	59.28	17.16	26.97	64.04	48.30	80.23
JPN	96.90	9.84	10.12	90.29	88.53	85.01
KAZ	82.98	29.01	16.51	85.28	56.56	86.99
KEN	26.45	40.73	10.57	52.79	14.44	68.20
KGZ	63.50	5.22	15.69	71.44	27.74	71.98
KHM	58.83	21.50	17.38	44.71	28.82	71.68
KIR	47.62	3.32	18.08	50.40	13.12	61.95
KNA	71.02	48.41	31.11	66.22	71.17	81.42
KOR	74.75	45.93	20.62	92.78	93.34	90.72
KWT	90.38	5.21	15.28	66.14	73.78	82.41
LAO	36.63	37.69	20.67	37.37	13.25	65.35
LBN	85.95	11.31	24.48	58.84	57.24	82.97
LBR	44.45	19.30	10.33	36.86	4.66	50.68
LBY	50.60	20.25	28.12	39.46	15.22	65.73
LCA	27.45	4.89	24.34	46.54	39.38	67.37
LIE	75.70	28.07	28.93	80.52	99.82	90.67
LKA	77.25	30.17	15.70	70.47	20.60	77.02
LSO	81.41	32.99	19.31	39.05	22.04	70.33
LTU	55.48	39.40	15.57	81.56	71.88	83.58
LUX	98.03	100.00	27.94	85.00	85.56	100.00
LVA	69.85	49.24	17.76	76.44	79.03	86.22
MAR	84.82	46.75	20.72	55.53	39.68	81.79
MCO	98.04	65.88	20.06	71.63	91.10	92.48
MDA	55.79	66.55	20.67	72.74	51.09	83.76
MDG	36.33	4.55	14.11	42.20	7.12	60.05
MDV	84.65	45.93	26.51	56.24	38.28	80.36
MEX	75.14	7.45	25.65	77.78	48.83	80.54
MHL	50.61	3.74	20.07	48.22	16.82	59.94
MKD	49.91	14.52	18.88	65.65	54.93	75.77
MLI	17.84	16.17	10.61	24.83	9.80	57.49
MLT	81.39	74.21	17.61	81.83	89.67	90.00
MMR	63.53	37.27	15.05	33.34	19.99	69.76
MNE	95.91	32.57	22.20	77.52	58.55	89.99

ISO	Infrastructure	Quality	Affordability	Knowledge	Internet usage	EDAI
MNG	36.97	9.44	19.63	73.42	31.17	73.71
MOZ	16.64	37.25	9.75	36.89	9.36	59.90
MRT	36.32	16.61	12.27	29.38	14.29	62.61
MUS	92.15	4.79	20.91	72.76	51.68	83.65
MWI	38.53	22.76	10.38	35.06	7.37	45.80
MYS	54.82	2.92	21.95	77.36	61.20	77.85
NAM	53.18	1.77	23.58	49.23	26.91	68.43
NER	17.68	10.88	17.56	10.54	3.52	55.73
NGA	64.28	13.44	14.98	43.95	14.39	69.49
NIC	71.83	4.00	15.94	51.21	20.44	71.17
NLD	66.02	33.94	28.39	94.51	89.39	88.62
NOR	89.58	36.82	15.64	93.84	87.96	91.91
NPL	50.12	46.97	13.19	53.12	18.48	68.64
NRU	80.28	20.70	24.73	37.88	16.53	77.08
NZL	92.17	12.13	29.64	95.61	82.46	94.32
OMN	50.25	21.18	20.75	73.59	53.79	78.65
PAK	65.41	17.78	15.85	43.38	9.69	66.24
PAN	68.52	21.96	25.40	70.39	42.89	80.44
PER	66.69	29.02	22.53	75.03	39.79	76.76
PHL	80.28	4.84	19.00	74.08	30.91	81.27
PLW	55.90	5.87	26.32	66.70	18.13	74.51
PNG	46.30	10.46	10.00	38.08	6.54	61.39
POL	72.18	44.40	19.34	89.44	54.47	84.53
PRK	51.47	21.04	21.97	22.12	43.92	65.65
PRT	93.16	52.49	23.43	82.81	74.53	92.73
PRY	36.72	38.55	17.87	60.14	29.81	73.54
QAT	92.93	18.17	24.69	66.31	68.05	85.02
ROU	67.54	36.01	17.45	75.42	64.71	83.54
RUS	77.56	21.29	18.73	88.96	62.67	81.89
RWA	75.79	20.44	8.73	54.27	16.07	73.00
SAU	47.20	16.91	20.41	77.37	53.92	78.29
SDN	44.73	1.70	17.58	25.01	12.98	61.02
SEN	44.34	48.12	23.79	36.65	16.66	67.96
SGP	94.95	27.96	20.54	89.87	84.76	89.05
SLB	19.86	5.41	15.68	35.25	6.34	28.17
SLE	20.78	0.85	22.52	31.43	10.23	54.90
SLV	44.78	38.58	18.09	59.88	29.55	73.06
SMR	73.67	37.98	22.66	64.14	84.01	86.41
SOM	51.54	16.30	23.92	0.00	1.19	61.14
SRB	92.15	30.25	21.42	78.19	63.11	89.49

ISO	Infrastructure	Quality	Affordability	Knowledge	Internet usage	EDAI
SSD	1.73	38.45	12.71	14.76	1.93	42.57
STP	27.03	3.19	100.00	32.50	19.22	62.29
SUR	69.29	2.77	19.45	47.63	40.88	75.45
SVK	87.21	3.74	25.56	79.02	69.37	85.52
SVN	91.55	27.24	17.91	87.00	67.08	90.89
SWE	60.36	41.30	24.10	94.54	95.12	88.53
SWZ	30.69	30.19	24.71	48.24	10.74	63.92
SYC	81.39	28.00	30.10	68.41	51.53	84.50
SYR	65.17	0.34	25.20	37.54	23.76	71.78
TCD	14.63	26.28	16.06	15.67	2.79	11.32
TGO	21.35	16.61	11.27	51.10	10.02	64.61
THA	89.61	63.44	16.73	73.45	51.48	87.84
TJK	59.25	20.56	13.07	54.12	8.90	65.21
TKM	74.88	14.70	17.74	36.29	10.04	72.99
TLS	68.81	39.27	19.88	39.92	18.39	66.65
TON	33.37	44.61	23.71	68.47	24.31	75.98
TTO	80.29	32.89	26.22	66.98	54.67	83.88
TUN	66.93	13.29	20.08	71.24	42.13	78.61
TUR	83.67	3.38	21.87	83.41	49.66	82.41
TUV	69.03	18.68	19.73	39.84	21.74	66.53
TZA	22.72	36.18	12.92	47.82	10.97	65.41
UGA	36.03	3.79	13.01	48.37	10.52	61.44
UKR	58.50	3.92	27.81	77.90	38.82	78.88
URY	89.19	20.05	23.45	81.16	73.72	86.80
USA	93.31	36.47	23.75	93.95	85.67	95.07
UZB	36.00	2.27	26.51	76.49	39.65	73.99
VCT	33.36	22.26	22.37	59.39	57.35	74.73
VEN	53.05	30.53	27.06	63.06	35.93	77.30
VNM	60.97	7.37	30.12	65.07	37.44	74.01
VUT	67.73	3.66	26.30	51.46	21.11	70.44
WSM	30.02	37.81	15.13	52.23	14.37	68.75
YEM	26.62	5.76	11.57	20.56	11.35	57.87
ZAF	75.45	39.68	17.56	75.73	38.30	79.32
ZMB	63.09	4.10	14.30	52.68	14.83	69.03
ZWE	60.70	0.70	13.87	47.30	19.43	64.90

APPENDIX VII. Variables used in fractional logit regression

Variable	Category
Current account balance (% of GDP)	Balance of Payments
Exports of goods and services (% of GDP)	Balance of Payments
Foreign direct investment, net inflows (% of GDP)	Balance of Payments
Imports of goods and services (% of GDP)	Balance of Payments
Net ODA received (% of GNI)	Balance of Payments
Total ODA (gross disbursement) for technical cooperation as % of GDP	Balance of Payments
Total ODA, by recipient countries as % of GDP	Balance of Payments
Total reserves in months of imports	Balance of Payments
Volume of remittances (in United States dollars) as a proportion of total GDP (%)	Balance of Payments
Annual mean levels of fine particulate matter in cities, urban population (micrograms per cubic meter)	Climate
Death rate due to the ambient air pollution (deaths per 100,000 population)	Climate
Renewable energy share in the total final energy consumption (%)	Climate
Bertelsmann Foundation Transformation Index	Corruption, transparency and country risk
Corruption Perceptions Index	Corruption, transparency and country risk
Countries with national statistical plans that are under implementation (1 = YES; 0 = NO)	Corruption, transparency and country risk
Global Insight Country Risk Ratings	Corruption, transparency and country risk
PRS International Country Risk Guide	Corruption, transparency and country risk
Statistical Capacity score (Overall average)	Corruption, transparency and country risk
Varieties of Democracy Project	Corruption, transparency and country risk
World Economic Forum EOS	Corruption, transparency and country risk
CPIA building human resources rating (1=low to 6=high)	CPIA
CPIA debt policy rating (1=low to 6=high)	CPIA
CPIA efficiency of revenue mobilization rating (1=low to 6=high)	CPIA
CPIA equity of public resource use rating (1=low to 6=high)	CPIA
CPIA financial sector rating (1=low to 6=high)	CPIA
CPIA fiscal policy rating (1=low to 6=high)	CPIA
CPIA gender equality rating (1=low to 6=high)	CPIA
CPIA macroeconomic management rating (1=low to 6=high)	CPIA
CPIA policy and institutions for environmental sustainability rating (1=low to 6=high)	CPIA
CPIA property rights and rule-based governance rating (1=low to 6=high)	CPIA
CPIA quality of budgetary and financial management rating (1=low to 6=high)	CPIA
CPIA quality of public administration rating (1=low to 6=high)	CPIA
CPIA social protection rating (1=low to 6=high)	CPIA
CPIA trade rating (1=low to 6=high)	CPIA
CPIA transparency, accountability, and corruption in the public sector rating (1=low to 6=high)	CPIA
IDA resource allocation index (1=low to 6=high)	CPIA
Average interest on new external debt commitments (%)	Debt statistics
Debt service as a proportion of exports of goods and services (%)	Debt statistics

Variable	Category
Multilateral debt (% of total external debt)	Debt statistics
Short-term debt (% of exports of goods, services and primary income)	Debt statistics
Total debt service (% of exports of goods, services and primary income)	Debt statistics
Total external debt divided by population.	Debt statistics
International migrant stock (% of population)	Demographics
Population ages 65 and above (% of total population)	Demographics
Burden of customs procedures - World Economic Forum	Ease of doing business
Dealing with Construction Permits - Building quality control index (0-15) (DB16-19 methodology)	Ease of doing business
Enforcing Contracts - Cost (% of claim)	Ease of doing business
Getting Credit - Strength of legal rights index (0-12) (DB15-19 methodology)	Ease of doing business
Getting Electricity - Cost (% of income per capita)	Ease of doing business
Getting Electricity - Reliability of supply and transparency of tariff index (0-8) (DB16-19 methodology)	Ease of doing business
Paying Taxes - Payments (number per year)	Ease of doing business
Paying Taxes - Time to complete a corporate income tax correction (weeks) (DB17-19 methodology)	Ease of doing business
Paying Taxes - Total tax rate (% of profit)	Ease of doing business
Percentage of firms identifying access to finance as a major constraint - Enterprise Survey	Ease of doing business
Protecting Minority Investors - Strength of minority investor protection index (0-10) (DB15-19 methodology)	Ease of doing business
Registering Property - Cost (% of property value)	Ease of doing business
Registering Property - Geographic coverage index (0-8) (DB16-19 methodology)	Ease of doing business
Registering Property - Procedures (number)	Ease of doing business
Registering Property - Quality of land administration index (0-30) (DB17-19 methodology)	Ease of doing business
Resolving Insolvency - Recovery rate (cents on the dollar)	Ease of doing business
Resolving Insolvency - Time (years)	Ease of doing business
Starting a Business - Time - Women (days)	Ease of doing business
Trading across Borders - Time to import: Border compliance (hours) (DB16-19 methodology)	Ease of doing business
Government expenditure on education, total (% of GDP)	Education
Human capital index (HCI) (scale 0-1)	Education
Proportion of seats held by women in national parliaments (% of total number of seats)	Education
Proportion of teachers who have received at least the minimum organized teacher training	Education
Pupil-teacher ratio, primary	Education
Employment in agriculture (% of total employment) (modeled ILO estimate)	Employment
Employment in agriculture, female (% of female employment) (modeled ILO estimate)	Employment
Employment in services (% of total employment) (modeled ILO estimate)	Employment
Employment in services, female (% of female employment) (modeled ILO estimate)	Employment
Ratio of female to male labor force participation rate (%) (modeled ILO estimate)	Employment
Unemployment, youth female (% of female labor force ages 15-24) (modeled ILO estimate)	Employment
Unemployment, youth male (% of male labor force ages 15-24) (modeled ILO estimate)	Employment
Account ownership (% of population ages 15+)	Financial access
Broad money (% of GDP)	Financial access
Number of automated teller machines (ATMs) per 100,000 adults	Financial access

Variable	Category
Number of commercial bank branches per 100,000 adults	Financial access
Remittance costs as a proportion of the amount remitted (%)	Financial access
Expense (% of GDP)	Fiscal
Compensation of employees (% of GDP)	Fiscal
Goods and services spending (% of GDP)	Fiscal
Subsidies and transfers (% of GDP)	Fiscal
Military expenditure (% of GDP)	Fiscal
Tax revenue (% of GDP)	Fiscal
Land area (sq. km)	Geography
Mountain area (square kilometers) as percentage of total area	Geography
Proportion of forest area within legally established protected areas (%)	Geography
Being landlocked (0/1)	Geography
Age-standardized mortality rate attributed to ambient air pollution (deaths per 100,000 population)	Health
Contraceptive prevalence, modern methods (% of women ages 15-49)	Health
Current health expenditure per capita, PPP (current international $)	Health
Health worker density, by type of occupation (per 10,000 population)	Health
Infant mortality rate (deaths per 1,000 live births)	Health
International Health Regulations (IHR) capacity, by type of IHR capacity (%)	Health
Life expectancy at birth, total (years)	Health
Number of new HIV infections per 1,000 uninfected population	Health
Prevalence of undernourishment (%)	Health
Proportion of population practicing open defecation, by urban/rural (%)	Health
Universal health coverage (UHC) service coverage index	Health
LPI international shipments score	Logistics
LPI logistics competence score	Logistics
LPI timeliness score	Logistics
LPI tracing and tracking score	Logistics
Percent of the population without postal services	Logistics
Financial Account openness (Chinn-Ito Index)[28]	Macro Indicators
IMF Export Diversification Index	Macro Indicators
Tariff rate, applied, weighted mean, manufactured products (%)	Macro Indicators
Trade openness (exports + imports)/GDP (%)	Macro Indicators
Volatility of terms of trade (standard deviation across last 10 years)[29]	Macro Indicators
GDP (current US$)	National accounts and real sector
Gross fixed capital formation (% of GDP)	National accounts and real sector
Gross fixed capital formation, private sector (% of GDP)	National accounts and real sector

[28] Chinn and Ito (2006).

[29] Gruss and Kebhaj (2019).

Variable	Category
Headline food inflation (Consumer Food Prices, 2 years average)	National accounts and real sector
Households and NPISHs final consumption expenditure (% of GDP)	National accounts and real sector
Manufacturing value added as a proportion of GDP (%)	National accounts and real sector
Nominal GDP in US$ translated using purchasing power parity (PPP) exchange rates, divided by population.	National accounts and real sector
Proportion of medium and high-tech industry value added in total value added (%)	National accounts and real sector
Services, value added (% of GDP)	National accounts and real sector
Suicide mortality rate, by sex (deaths per 100,000 population)	Social Development
Access to electricity, rural (% of rural population)	Urbanization
Access to electricity, urban (% of urban population)	Urbanization
Population density (people per sq. km of land area)	Urbanization
Population in the largest city (% of urban population)	Urbanization
Proportion of urban population living in slums (%)	Urbanization
Rural population (% of total population)	Urbanization

Note: With the exception of Balance of Payments and Macro Indicators (Chinn-Ito, IMF); Climate and Health (UN), Corruption, transparency and country risk (BTI Project, Transparency International, HIS Global Insight, PRS Group, Economist Intelligence Unit, V-Dem, World Economic Forum), all other category of variables are from World Bank's, mostly from the World Development Indicators.

APPENDIX VIII. Fractional logit regression

The model proposed by Papke and Wooldridge (1996) has the following structure:

$$E(y|X) = G(\beta X)$$

Where $G(\cdot)$ denotes the link-function satisfying $G(\cdot) \in [0,1]$, X represent a set of explanatory variables and y should be regarded as a dependent variable. The link function guaranties that the predicted values lie in the above-mentioned interval. In the following paper the authors decided to implement the logit function:

$$G(\cdot) = \frac{exp(\cdot)}{1 + exp(\cdot)}$$

Generalized linear models (GLM) are usually fitted with maximum-likelihood algorithms (Hardin and Hilbe, 2007). Papke and Wooldridge (1996) propose however a particular quasi-likelihood method, which maximizes the following Bernoulli log-likelihood function:

$$l(\beta) = y\, log\big(G(\beta X)\big) + (1 - s)\, log\big(G(\beta X)\big)$$