Knowledge Economy Classification in African Countries: A Model-Based Clustering Approach

Amavilah, Voxi Heinrich and Otero, Abraham and Andres, Antonio Rodriguez

Economics/ Division of Social and Behavioral Sciences Estrella Mountain College 3000 N. Dysart Road, Avondale, Arizona, USA, Polytechnic School, University San Pablo CEU. Urbanización de Montepríncipe s/n 28668 Boadilla del Monte Madrid, Spain, VSB-Technical University of Ostrava Faculty of Economics Department of Economics Sokolska trida 33, 702 00 Ostrava, Czech Republic

15 March 2021

Knowledge Economy Classification in African Countries: A Model-Based Clustering Approach

Antonio Rodríguez Andrés (Corresponding Author)

VSB-Technical University of Ostrava

Faculty of Economics

Department of Economics

Sokolska trida 33, 702 00 Ostrava

Czech Republic

E-mail address: antonio.rodriguez.andres@vsb.cz

Voxi Heinrich Amavilah

Economics/ Division of Social and Behavioral Sciences

Estrella Mountain College

3000 N. Dysart Road, Avondale, Arizona, USA

E-mail: vhsamavilah@gmail.com (preferred)

voxi.amavilah@estrellamountain.edu

Abraham Otero

Polytechnic School, University San Pablo CEU.

Urbanización de Montepríncipe s/n

28668 Boadilla del Monte

Madrid, Spain

E-mail: abraham.otero@gmail.com

ABSTRACT

Knowledge economy (KE) has been a central issue in the political economic literature of advanced economies, but little research has focused on the transition towards a KE in Africa. Using a latent profile analysis, six clusters of the KE were found in the region. The clusters range from very prepared with good performance in all KE dimensions (institutional, education, and innovation output) to very unprepared with low performance inf each KE dimension. Lastly, we offer policy recommendations that shed some light on the national and international economic policies towards a more knowledge-oriented environment. One such recommendation is that effective policies should consider both the similarities and dissimilarities of African knowledge economies. How precise that can be done is one direction future research can take.

KEYWORDS: *Africa, model-based clustering; knowledge economy; cross-country comparisons; exploratory analysis*

JEL Classification: O10, O34, O38, P00, P48, C45

1 Introduction

Knowledge-based economies (henceforth, KBE) have received central attention in key policy reports from different international organizations such as the Organization for Economic Co-operation and Development (OECD, 1996) and the World Bank (2007)[1]. Popularized by Drucker (1969), the concept of knowledge economy (henceforth, KE), which forms the basis of KBE, was primarily introduced by Machlup (1962) who classified knowledge depending on its application to areas of economic activities. Before that Stigler (1961) viewed knowledge as an economic category with an emphasis on information searching costs. Other scholars consider KE as an economic system where knowledge is a key factor (or resource) of production and economic growth (see for instance, Kochetkov and Vlasov, 2016). The fundamental determinants of the KE include significant dependence of the economy on intellectual abilities than on physical inputs or natural resources in the context of the integration of new knowledge at each level of the production process (see, Powell and Snellman, 2004). In addition to this, in KBE the share of intangible capital (for example IPRs is part of it) is greater than that of tangible capital in the overall stock of capital.

Research continues to investigate the precise channels and mechanisms by which information technology affects economic growth and development as we illustrate later. Despite such efforts focused attention on KE in African countries remains limited. The formal literature there continues to use placeholders like the *Africa dummy* to capture the effects of innovation and technology on growth, assuming that Africa is distinctly different from other regions of the world, homogeneous within itself in all possible manners (Conway and Greene, 1993). The assumption is mistaken, because Gyimah-Brempong and Wilson (2004), for example, used an expanded Solow dynamic panel model to examine the effects of *health human capital* on the economic growth of Sub-Saharan African and OECD countries. They found that the difference between the two groups of countries is only of the order of magnitude, not of arithmetic signs; health human capital influenced economic growth positively in both groups of countries, suggesting that Africa is not different. Amavilah (2006) found that the technical capability of 14 African countries is diverse and concluded that blanket policies may not be effective. Asongu et al. (2018) stylized framework which divides African KBEs into leaders and laggards relative to the knowledge frontier without an endogenous categorization of the national economies reveal that countries perform differently on different dimensions of the KE, with South Africa leading in innovations, Botswana and Mauritius in institutional elements, North African countries in education, and so on. Not only is the assumption of technological knowledge homogeneity unreasonable, it also appears that the technical capabilities of African countries are ever-changing, sometimes in convergent, and other times in divergent, ways. This understanding holds out possibilities for expanding research into the direction of leader-follower models of innovation of the type proposed by Stiglitz (2015) for Nordic countries.

Moreover, the measurement of technological capabilities is still controversial from a methodological as well as conceptual viewpoints (see Rizk et al., 2017). To deal with the controversy, numerous composite indicators have been attempted to measure how different prepared countries are on the path towards a KE. These indicators can be relevant as initial assessment tools for policy makers and can be used as benchmarks to compare each country with its competitors or partners. Even so, at the macro level the indices differ in the variables included as well as the statistical methodology used to generate them. Additionally, many of the indices focus on high income countries with little representativeness of African countries, although the

region consistently ranks the lowest, and stands to gain the highest marginal returns. By looking at the KE over time, some scholars noted a decline in Sub-Saharan African countries' development towards a KE during the 2000-2012 period (see Anyanwu, 2012; Asongu et al., 2018). How precise these findings are is not easy to ascertain because composite indicators rely on the quality and accuracy of the national statistical institutions, which might be particularly problematic for African countries that suffer from a statistical tragedy where weak capacity, inadequate funding and lack of coordination have yielded unreliable estimates (see, for instance, Devarajan, 2013; Jerven, 2013).

Another related branch of the literature worth pointing out here is that of the technology clubs, in which earlier attempts to explore the clustering of technological knowledge of countries do exist, but they too are mainly focused on high income countries, or world-wide datasets. For example, Castellacci (2011), and Castellacci and Archibugi (2008) employ a hierarchical clustering technique. The former paper investigates the relationship between technological factors and their long run economic performance, and how this hypothesized relationship differs across countries. The results indicate three specific clusters according to the threshold values of the variables of country's technological capabilities: ***advanced*** *(*mostly advanced economies*),* ***followers*** (nearly all Latin America, the Middle East, Southern Europe, and East Asia), and ***marginalized*** (mostly South Asian and African economies). This method tells us that each country-year observation belongs to only one cluster, but when Castellacci and Archibugi (2008) approached the same problem by clustering 120 countries in a dataset with indicators of technological knowledge, African countries were excluded from the sample. Lastly, there is only one study that applies the k-means clustering approach to categorize countries according to their level of KE based on World Bank's KE procedure (Paz Marín et al., 2018). Besides the fact that African countries as well as emerging countries are not included in the sample, this study has two inherent problems. The first is that the selection of clusters is not based on any statistical criterion, and the second is that the study does not allow countries to change from one cluster to another over time. Thus, as Rao and McNaughton (2019) demonstrate, the value of knowledge is diminished when the dynamic dimension as well as the nature of panel design are neglected. In addition to this, while focusing on a sample of OECD countries, the study largely ignores that the conditions for, and transition towards, a KE are rather different for different countries.

Our main research question is: *Are African countries technologically diverse or homogenous in their transition towards a KE through different dimensions?* This question can be unbundled as the following sub-questions:

- What is the African countries' potential to enable a KE society through the development of ICT?
- What is the institutional framework needed to facilitate the transformation of national economies to a capable KE?
- What are the human capital requirements needed to absorb the full potential of innovations that support KE?
- How have innovation outputs become a key driver of further advances toward KE?

We argue that the question (or its sub-questions) justifies this investigation, because the current lack and deficiency of classification models in the literature on the KE may undermine the usefulness to policy and future research of composite indicators for quantifying the progress towards KE at the national level,

especially in developing countries. We also argue that given the importance that policy makers place on African countries' progress towards a KE a richer and deeper approach is needed as an alternative strategy to using simple composite indices to capture the level of the KE across countries. This study demonstrates the use of a refined model-based clustering approach to distinguish underlying homogeneous groups, or latent classes, of countries belonging to similar levels of KE. The main idea is to create a taxonomy of countries that allows us to make comparisons and generate a new classification according to their respective KE levels. For that purpose, we employ a certain number of KE dimensions: education, institutional factors, innovation, and ICT infrastructure. Instead of picking KE dimensions arbitrarily ourselves, we take the ***Knowledge Assessment Methodology*** (henceforth, KAM) as a benchmark (see Chen and Dahlman, 2006) and we apply a General (Gaussian) Mixture Models (henceforth, GMMs) clustering approach due to its advantages in comparison to traditional clustering techniques (such as k–means) as we describe later in that they provide an objective mathematical criterion for determining the number of clusters present (Fraley and Raftery, 2002).

The relevance of a GMM, and our novel application of it, is that it permits us to determine endogenously how the different components of the KE play an important role in the development context and how similar African countries are with respect to the levels of KE, which in turn allows us to generate a ranking or classification of countries that is dynamic and non-arbitrary (see for instance for a similar application, Abad-González and Martínez, 2017). We also contribute to the modeling strand of literature in that our approach has both exploratory as well as confirmatory elements to it. However, since the existing literature is too thin to permit for a categorical choice of between exploratory analysis and confirmatory analysis, and we are, therefore, unable to construct testable hypotheses in a conventional way, we favor a GMM approach over alternatives like k-means. The approach serves as a complement for the judgement of stakeholders to know the real stages of countries are in, and how they are evolving or can move forward This is a significant contribution because even though composite indices like KEI exists, it is not clear how they classify African countries. If countries are miss-classified, then the measured effects of ICT or IT on development would be incorrect. To stress the point, in regression models often ICT, IT, or collectively KEI is often represented by the Africa dummy variable which assumes that African countries are technologically homogenous. Our classification model and analysis bring clarity to how correct these matters are and so by extension to the nexus between development and ICT or IT.

Moreover, previous research also reveals the strengths and weaknesses of each African country over the different KE dimensions. Hence, the study enhances understanding of how African countries may improve their technological competitiveness and consequently the quality of life for their citizens. This is the overarching goal and to accomplish -- it we borrow from the KAM initiative published by the World Bank (2012) and apply a GMM to classify 50 African countries over the period 1996–2017 according to their progress towards a KE. Our results show **four representative clusters of KE levels**: ***Very* prepared, *Prepared,* Unprepared, *and Very* unprepared.** Thus, African countries are technologically diverse and statistical tests confirm the presence in that data of different clusters. One key conclusion we draw is that even if they do exist technology clubs are not static but dynamic and they can coexist with technology gaps. Once the data is labelled, a classifier is applied. In this way, we provide some ranking for assessing the degree of competitiveness of the African countries in the context of KBE. Another conclusion is that

different countries can belong to different clusters at different times for different reasons, which questions the relevance of static classifications. Thus, our study makes three significant contributions to the KE literature: First, most studies involve comparisons of KE across a relatively small number of countries, and mainly focusing on high income countries. Our study focuses on a sample of 50 African countries. This is significant because the Africa region has been largely neglected in formal studies on KE. Indeed, the choice of the sample might also lead to different outcomes in terms of KE convergence. Most of the empirical literature has focused on the macro determinants of the KE in African countries (Andrés et al., 2015; Asongu et al., 2018; Asongu et al., 2020). They have employed ad hoc statistical techniques to identify causality by using longitudinal data. Particularly, they employ generalized method of moments (GMM) that deals with endogeneity issues, although many of the empirical outcomes might not be valid as GMM is quite sensitive to the choice of valid external or internal instruments. Our approach is to endogenously determine how similar African countries are with respect to the levels of KE in a way that allows us to generate a dynamic ranking or classification of countries that is endogenous and non-arbitrary, which adds to the formal literature on KE.

The second contribution of the paper is that it also evaluates the state of the KE in African countries and explores the differences across the different KE dimensions. Hence, it fuels the debate over the ways to measure the KE through composite indicators as elaborated by several international organizations (World Bank, European Commission, and World Economic Forum, among others). Finally, the study contributes to existing literature by seeking to formalize the diverse typologies of KE in African countries within the economics literature using a model-based clustering approach rather than other cluster algorithms (for instance, k–means), again, thereby enabling us to create classificatory topologies of African countries' KE levels based on strong statistical criteria.

The remainder of the paper is as follows: In Section 2, we review the current literature in two subsections. The first subsection overviews the KE assessment methodologies, whereas the second describes clustering to justify why we selected GMM to study KE. Section 3 describes the variables and data we utilize, and the proposed methodology: the GMM technique. The results are presented and discussed in Section 4, while Section 5 concludes the paper with the implications for policy and further research.

2 Literature review

2.1 KE Assessment Methodologies

Conceptually, knowledge is a source of competitive advantage in the 21st Century. This is not a new idea; economic theory has long appreciated the importance of knowledge in economic performance and human welfare (Dodgson and Gann, 2018, pp. 12-32). Schumpeter (2005), for example, saw technology as a key driver of economic growth and productivity that could both create and destroy jobs, and often does both simultaneously (Becker, 2005; Dodgson and Gann, 2018, pp.12-32). This is Schumpeter's well-known "gale of creative destruction," and it means that the endogenous growth approach emphasizes knowledge as both an output- input (Romer, 1986; 1990; Lucas, 1988; 1993; Rebelo, 1990; Grossman and Helpman, 1991). As input, knowledge allows technological advancement and associated innovations to drive long-

run growth. Nevertheless, from the empirical side, it is hard to net out the contribution of knowledge from the total factor productivity, because of the measurement problem (either at micro or macro level). Krugman (2013), for example, cast doubt on the empirical verification of the theory by pointing out that there are plenty of assumptions about how unmeasurable things affect other unmeasurable things. The doubt is understandable because knowledge also involves combination of factors that interact in intangible ways. According to Kaplinsky (2005), there are several types of knowledge rent: technological, human resources, organization and marketing and design.

Leaving aside the conceptual issues associated with the multidimensional concept of KE, the literature on measuring knowledge in developing countries has also been limited, despite the efforts such as those by Samoilenka and Osei-Bryson (2018) who argue for "context-specific micro-economic" assessment of national ICT capabilities in the DEA analytical framework (see, also, Samoilenka and Osei-Bryson, 2008). As Carter (1996) has pointed out it is not only difficult to measure knowledge at firm level, but also at country level. The difficulty in quantifying knowledge and innovations has obstructed research and policy, leading to an assumption of homogeneous regional dummies as representations of innovations and technology like the so-called *Africa dummy* now common to many growth regressions in which countries are arbitrarily group according to their income levels (Barro, 1991; Collier and Gunning, 1999a; 1999b; 1996; Collier, 2007; Barro and Lee, 1993; Mauro, 1995; Easterly and Levine, 1997; Burnside and Dollar, 1997; Knedlik and Reinowski, 2008; Englebert, 2000; Jerven, 2011; Easterly, 2001; Azam et al., 2002). The results of these studies have clearly extended the Solow-Swan tradition, but just as clearly, they continue to leave the growth effects of technology and technological change unexplained. Despite existing gaps there has been little research on these issues in African countries.

We put emphasis on the measurement of KE at macro level in this paper. In this respect, a variety of composite indicators have been proposed that acknowledge the multidimensional aspect of the KE concept. Nevertheless, there is no clear consensus yet about indicators employed to measure a KBE. One of the main criticisms is that existing composite indicators tend to be data driven, meaning that they use only the information available across countries, which essentially means countries for which data does not exist, or the available data has missing values, these countries are left alone in the dark (Shapira et al., 2006). For example, Archibugi et al. (2009) make a comparison of aggregate indicators of technological capabilities,[2] and conclude that the rankings at country level have consistently significant discrepancies for some nations. Moreover, these indicators are less suitable to capture changes in technological knowledge over time. In many situations, the choice of indicators is restricted due to data availability and the mutual interdependencies between a combination of inputs and outputs indicators. Lastly, these composite indicators rely on the quality and accuracy of the national statistical institutions -- a "statistical tragedy" in Africa's case (Devarajan, 2013).

In this paper, we take a more holistic approach in examining the determinants of KE, assuming that their relative performance can be assessed through a benchmarking methodology.[3] This is an old concept in the context of organizational comparisons but one that has, nevertheless, been commonly also used in the context of country comparisons to identify and compare the degree of competitiveness of countries (see for instance, Dolowitz and Marsh, 2000). Some examples of the benchmarking methodology are: first, the OECD's going for Growth Exercises which identify five productivity related policy priorities for each

OECD member (OECD, 2005), second, the European Commission's Internal market Scoreboard (European Commission, 2020) which ranks member countries' performance in the implementation of the required legislation for internal market convergence, and lastly, the KAM that measures the countries' capacity to compete in what the World Bank has named the Knowledge economy (World Bank, 2007).

The formal literature leads us to employing the KAM (1996) framework for selecting the input and output indicators to measure countries' capacities to compete within the same KE (see Parcero and Ryan, 2017; Širá, 2020). KAM is a reasonable starting point because World Bank researchers constructed the knowledge economy index (KEI) as an element of the general knowledge index (KI) in any national economy in response to the need we outlined above. In addition, using data for the last available year for each indicator, Archibugi et al. (2009) show that there are positive and high correlations, ranging from 0.47 to 0.92, of this index with other composite measures of technological knowledge. Nonetheless, various composite indicators might measure different things, for if the correlations are relatively high, it simply means support for the choice of the indicators. The KAM is reasonable also because it is the most inclusive methodology in comparing and assessing the level of KE across countries, and it recognizes that the conditions leading to a KBE should include an institutional regime offering the right incentives, an educated and skilled labor force, a modern information infrastructure, and an effective innovation system. Shortly, the methodology involves four pillars or dimensions of KE and 148 indicators for 146 countries in the world, which are described briefly below (see, Chen and Dahlman, 2006; World Bank, 2012):

- **Pillar 1: Economic and institutional regime**: It provides incentives for creation, dissemination, and use of the existing knowledge. It covers a diversity of f issues and policy areas ranging from aspects of the business environment, finance and banking, macroeconomic framework, regulations, governance, and institutional quality. The importance of institutions and their impact on economic growth has widely been recognized in the formal literature (North, 1990; Landes, 1998). Although inadequate drivers, institutions of governance are shown to promote economic growth (Blackburn and Forgues-Puccio, 2010) and so too are the incentives that economic and financial institutions offer (Ryan and Shinnick, 2011; Tchamyou, 2016; Andrés et al., 2015; Kauffman et al., 2010; World Bank, 1996; Chen and Dahlman, 2005). The selected proxy variables for this pillar are corruption control index, regulatory quality, and the rule of law.

- **Pillar 2: Education**: One of the most relevant pillars of the KE is human capital, education is a critical factor in the creation and dissemination of knowledge, and for the use of knowledge effectively. In addition, most new ideas and inventions are generated in knowledge clusters where scientific skills are required (Buesa et al., 2010, Marrocu et al., 2013, among others). The selected indicators for the education pillar are gross primary, secondary, and tertiary enrollment rates.

Pillar 3: Information and communication infrastructure (ICT): It facilitates the effective communication, processing, and dissemination of information. ICT can be defined as a combination of hardware, software and communication networks that enable electronic information capture, storing, processing, and transfer. ICT and supporting technologies work in synergy in sustaining business activities and socioeconomic development (Borgmann, 2006). The economics literature has paid attention to the effects of ICT on innovation and socioeconomic development. The ICT term is largely used as an extension of or interchangeable with information technologies (IT). Research suggests that total telephones and

mobile phones positively influence innovation (see, Carayannis et al., 2013). Moreover, empirical research has also documented the impact of ICT on economic growth (Driouchi et al., 2006, Thompson and Walsham, 2010; Tripathi and Kumar Innani (2020), and Datta and Agarwal, 2004). Furthermore, Chavula (2013), and Qureshi (2013), among others, find that telecommunications infrastructure plays an important role in promoting economic growth, while for Grant and Yeo (2018), ICT affects the economy through technology investment and financing in the manufacturing and service industries. ICT indicators such as total telephones per person, internet use per person, and fixed broadband internet are used as proxies for the ICT pillar.

- **Pillar 4: Innovation system:** This dimension is more concerned with innovation outputs. A good innovation system consists of an interconnected array of universities, research centers, firms, consultants, and other organizations that generate, assimilate, and adapt knowledge. Previous research has paid attention to the characteristics of the national innovation systems and their relevance for economic growth and competitiveness (Lundvall et al., 2009). In terms of intellectual protection, it has been argued that stronger IPRs protection leads to more innovations (Arrow, 1962). It is also clear that the IPRs systems are not well developed in many countries, and this a clear evidence for the low levels of intellectual property creation measured by the patents per capita. The same applies to the universities as a source of new knowledge. African universities are focus on teaching and are behind in terms of research. This is mainly due to inadequate computers and network systems, unstable power supplies, and limited capacity to pay for subscription content (see, Marfo et al., 2015: Mitchell et al., 2020). Proxies such as scientific journal articles and patent applications in per capita terms are used to capture innovation.

Notice that countries should keep a proper balance among the four pillars to create, disseminate and use of knowledge efficiently. Clearly, all these pillars are interrelated and connected (see Figure 1). From theory, we know that knowledge and technology can contribute to a country's wealth, because the generation of wealth at both the country and firm level can be represented by a conventional production function of the following log-log framework:

$$\ln(Q) = a_0 + b_1 \ln(Conventional\ factors) + b_2 \ln(KEI)$$

where at the country-level Q is a measure of development, *Conventional factors* include capital and labor, and KEI include ICT, and so logically IT. In that sense, ICT is a relevant factor in the generation of wealth at country level as Torero and von Braun's (2006) have illustrated at both the national and firm level.

In addition, this framework allows us to fully understand a country's strengths and weaknesses relative to the other countries, therefore being useful from the policy dimension as it can reveal country's problems and opportunities where policy makers can implement national strategies for achieving a KE. Lastly, we explicitly agree that there are inputs and outputs in our conceptual framework. Thus, we use patent data that is more an innovation output and education is more a necessary input to acquire technology or new inventions. Nevertheless, we do not explore the efficiency of economies on their way to KE (see for instance, Samoilenka and Osei-Bryson, 2008) but rather how the dynamics towards a KE has changed through these four dimensions at national level.

Insert Figure 1 near here

Each of the pillars outlined above has several indicators as proxy variables for each dimension that ranges from zero to ten, suggesting that a higher index means a higher KE level. Based on data from 2012 (the latest year), Sweden tops the list with the KEI score of 9.43, followed by Finland with a 9.33 score. For comparison, the United States ranks 12th with a score of 8.77. Among African countries included in our sample, Mauritius is at 61st with a score of 5.52, followed by South Africa with a score of 5.21, Botswana with a score of 4.31, and Namibia with a score of 4.10. African countries with the lowest KEI scores in our sample are Angola and Sierra Leone with scores of 1.08 and 0.87, respectively. By comparing African KEIs to the rest of the world, most Sub-Saharan African countries (SSA) are still in the KE infancy with only a few countries such as Mauritius, and South Africa close to being on the transitional path in the journey towards a viable KE (see Agyapong and Oseifuah, 2015). As pointed out earlier in a related area, some scholars have in fact noted a decline over time in Sub-Saharan African countries' development towards a knowledge economy between the period 2000-2012, not only in the total KEI score but also in terms of the three pillars of the KE (education, ICT infrastructure, and institutional quality), see e.g., Anyanwu (2012), and Asongu et al. (2018). Indeed, the education and ICT pillars are the weak pillars in comparison with the innovation dimension, which is a matter of profound concern and deserve research attention, because overall human capital and ICT infrastructure are among the main facilitators of KE.

2.2 Clustering Overview: GMM vs. k-means

Cluster involves the partioning of a set of objects into a useful set of mutually exclusive such that the similarity between the observations within each cluster (i.e., subset) is high, while the similarity between the observations from the different clusters is low (see Mardia et al., 1979). Before proceeding with an overview of the clustering techniques and further justification of our choice in our analysis, we next discuss the reasons for doing clustering.

One of the reasons is to find a set of natural groups and the corresponding description of each group (see for instance, Samoilenka and Osei-Bryson, 2008). Moreover, this approach allows researchers to generate a classification of groups that is endogenous and non-arbitrary (based on cut-off points, and ad-hoc weights). Hence, the use of cluster analysis assumes that there are natural groupings in the data. Secondly, cluster analysis is a powerful tool because it allows researchers to explore the socio-economic phenomenon through interaction with organization, technology, and people (Balijepally et al., 2011). Xiong et al. (2014) point out that cluster analysis should be used in combination with other research methods, such as in determining the number of clusters, validating clusters, and multicollinearity among variables.

Cluster algorithms can be categorized in various ways such as: Partitioning (e.g, k-means, k-median, hierarchical, fuzzy, density-based, and model-based clustering (Hair, Black, Anderson, and Tatham, 2006). This paper utilizes the latter methods as its title indicates. We will now provide a general overview of the last approach and highlight its differences and advantages with other classical approaches (e.g. k-means).

Model-based clustering advances the earlier clustering methods like hierarchical and k-means clustering that are heuristic and less formal so that it is possible for different runs of one k-means algorithm to generate different results even when the user specifies the optimal number of clusters. Model-based clustering

overcomes these weaknesses by considering the data as coming from a distribution that is mixture of two or more clusters unlike k-means which assumes a specific probability for each cluster (Fraley and Raftery 2002, Fraley et al., 2012). Theoretically, both k-means and GMM are partitioning clustering techniques, which try to divide the feature space into different regions and represent each of these actions by means of a prototype or centroid. The objective of this prototype is to be as representative as possible of the instances that fall in the region of space associated with it.

In the case of k-means, the prototype is the mean vector of the feature vectors of the instances that have fallen in the region of the feature space associated with the prototype. These regions are bounded by linear edges and they are called Voronoid regions. By contrast, in GMM techniques, Gaussian probability distributions ("Models") are used as prototypes that represent each region of space. Multi-dimensional Gaussians are characterized by a mean vector (note that this is the only representation that k-means uses for its prototypes) and a covariance matrix. In a way, the main advantage of using GMM over k-means is similar to the advantage of characterizing a population using a probability distribution (GMM) or only the mean value of the population (k-means). There is no doubt that using a probability distribution better characterizes the population. Some of the consequences of this difference are:

1. k-means generates clusters with a spherical shape, while GMM provides more flexible representations by allowing the Gaussians to have different variation in different directions, as well as different orientations in space. Of course, in the case of using a Gaussian oriented with the axes of the feature space, and with the same variation in all dimensions, we have a sphere. In other words: everything that can be represented by a k-means prototype can be represented by a GMM prototype. The opposite is not true.
2. In k-means the membership of a cluster is total; an instance either belongs or does not belong to a cluster. The real world is often more complex than this situation; for example, there may be countries that are in a transition between two states, without fully presenting the characteristics of either of them. In GMM each instance has associated a mathematical probability of belonging to a cluster. This probability, of course, could be 1 or 0, these being the cases in which the membership is clearer, and being equivalent to the representation capability of K-means. But it can also be any continuous value between (0,1), and an instance can belong, for example, to two different clusters representing two different states. This endows GMM with a greater ability to characterize transitions between states. On the other hand, the value of the probability of belonging to a cluster is a measure of the confidence that the sample actually belongs to the cluster. In k-means, all the samples that belong to a cluster, belong to it with probability 1. In GMM we have information about the certainty that we have about said membership as a continuous probability between [0,1]; i.e., we have more detail about the certainty of membership.
3. The evaluation of the k-means results is usually carried with ad hoc geometric criteria such as the Davies-Bouldin index, Dunn's index (Xiao, Lu, and Li, 2017; Havens, Bezdek, Keller, and Ppescu, 2008), Levine-Domany index (Levine and Domany, 2001), the silhouette coefficient (Zhou and Gao, 2014), and other similar criteria (Bezdek and Pal, 1998). In these criteria, the quality measure of a clustering configuration is usually based on two basic concepts: (1) the instances that belong to a cluster should be as similar as possible to each other, and (2) the instances should be as different as possible from the instances of the rest of the clusters. These indexes try to quantitatively formalize both criteria using ad hoc strategies that seem reasonable to human intuition, but for which there are no mathematical proofs that lead to an optimum configuration, beyond satisfying the ad hoc criterion itself. Consequently, there is no guarantee that these different indexes will select the same clustering configuration when applying different criteria to the same

clustering configurations. Furthermore, there is no mathematical proof that if one of the tested clustering configurations corresponds to the true underlying data structure, the index in question would prefer that configuration over the others. It is possible to use all these indices to evaluate a clustering configuration obtained with GMM. However, this is typically not done, and it is preferred to use the Bayesian Information Criterion (Chen and Chen, 2008). The Bayesian Information Criteria (henceforth, BIC) penalizes complexity and rewards for parsimony when comparing different models that differ in the extracted number of clusters, and at the same time it tries to maximize the likelihood of observing the data set. BIC is not an ad hoc criterion but is based on a powerful mathematical formalism: the Bayesian theory. It can be mathematically proven that if BIC is used to find the optimum model (clustering configuration in the case that concerns us) that best fits a set of observed data from a set of candidate models, and the real model that has generated the data ("true model") it is present in the set of candidate models, BIC will always choose the true model as long as the data set is large enough to allow its adequate estimation (Neath and Cavanaugh, 2012). Note that BIC depends on the existence of mathematical models (probability distributions) for its computation, and therefore it cannot be applied to k-means. The authors are unaware of any reason to prefer ad hoc geometric criteria (such as the silhouette coefficient) over BIC to evaluate a GMM cluster configuration and definitely, in the machine learning literature, BIC is by far the most common measure used when evaluating GMMs.

4. k-means is more sensitive to the initialization of centroids and has a greater tendency to get stuck at local minima compared to GMM. In this sense, we must highlight that the GMM implementation used in the paper uses a deterministic initialization for the Gaussians based on hierarchical clustering that guarantees that the same results will always be obtained when executing the algorithm on the same data.

In short: k-means is computationally more efficient than GMM: i.e., it requires less execution time. From the authors' point of view, this is the only advantage k-means has over GMM. The rest of the aspects, including greater flexibility in the shape of the clusters (this approach does not bias the structure of the clusters to have a specific structure as k-means does), a more powerful description of these (probability distributions instead of mean value; the possibility that a sample belongs to several clusters with different probabilities instead belonging to a single cluster), a robust mechanism for selecting the number of clusters (BIC), and less tendency to fall local minimums, GMM is superior to k-means. This situation is quite expected when the algorithm is currently more than 60 years old (Jain, 2010). Obviously, there has been some additional progress in clustering research in those 60 years. That said, for small to medium data sets (a few hundreds, or a few thousand data points), each represented by one or two dozen variables), GMM typically runs in less than a second, so this should not be a problem in practice. For large data sets (millions, tens of millions of data), the GMM run time can be considerably longer and k-means may be preferred as a less powerful but faster alternative (see for instance, Marquez, Felix, Garcia, Tejedor, and Otero, 2019).

3. The Proposed Methodology

3.1. Overview of the Dataset

We gathered the data from the World Development Indicators see https://data.worldbank.org/indicator). Observations on the key variables were selected based on data availability, because institutional variables from the World Bank, for instance, are available only since 1996. For that reason, our study covers the

1996-2017 years. Table 1 displays the World Bank's classification of the 50 selected economies according to 2019 Gross National Income (GNI) per capita[4]. There is only one African country in the group of high-income countries: Seychelles. In the middle there are countries classified as "upper middle economies" (14%) while most of the African countries are concentrated in the lower middle income (40%) and low-income group (44%).

Insert Table 1 near here

Variable definitions, and data sources are displayed in Table 1A. Table 2A presents descriptive statistics, while Figure 1A shows the correlation matrix of all variables employed in the empirical analysis. The data employed in the empirical examination were analyzed using the R statistical package (R Core Team, 2019). This means that before implementing the cluster technique, we carry out a thorough descriptive analysis of the variables selected for our cluster analysis. We end up with 23% missing values. Table 2 shows the distribution of missing values across variables in our sample. Most of the missing values correspond to the fixed broadband variable for internet subscribers. Value imputation strategies are detailed further in Table 3A in the Appendix.

Insert Table 2 near here

3.2 Description of the methodology

We shall now describe the main elements of our cluster approach. It is crucial to stress the GMM since is not widely used as the traditional methods in Economics.[5] Here, the data generation process (DGP) is assumed to be given by some finite mixture of probability distributions $f\,(X|\theta)$, where $X = (x_1, x_2, \dots, x_i, \dots, x_n)$ is an *nxm* matrix of *n* instances, each of them comprised of *m* features; i.e., $x_i = (x_i^1, x_i^2, x_i^3, \dots, x_i^m)$; x_i represents one of the African countries, and each of the m features is one of the variables in Table 1A; and $\theta = \left(\theta_1, \theta_2, \dots, \theta_g, \dots, \theta_K\right)$ are the parameters of the K Gaussian probability distributions that form the mixture (Verbeek et al., 2003), i.e., $\theta_i = \{\mu_i\ , \Sigma_i\}$, μ_i being the mean of the i Gaussian and Σ_i its covariance. Then the likelihood of an instance x_i having been generated by the mixture of Gaussians will be the sum of the likelihood that that instance has been generated by any of the Gaussians. This implies that the density of $\boldsymbol{x}$ will be given by a finite mixture of the form:

$$f\ (\theta) = \sum_{g=1}^{K} w_g\, f\left(\theta_g\right) \qquad (1)$$

where *K* is the number of Gaussians; and w_g acts as a weight that permits modeling the fact that different groups (clusters) may have a different number of instances within them (Ahlquist and Breunig, 2012).

Since there are a total of *n* instances, the likelihood that all of these instances *X* having been generated by the mixture of Gaussians will be the multiplication of the likelihood of each of the instances x_i having been generated by the mixture, i.e.,

$$f\ (\theta) = \prod_{i=1}^{n} \quad f\ (\theta)\ = \prod_{i=1}^{n} \quad \left(\sum_{g=1}^{K} \quad w_g\, f\left(\theta_g\right) \right) \qquad (2)$$

In practice, the data *X* is known, but the parameters of each of the mixtures $\theta = \left(\theta_1, \theta_2, \ldots, \theta_g, \ldots, \theta_K\right)$ as well as their weights $w = \left(w_1, w_2, \ldots, w_g, \ldots, w_K\right)$ are not known. In model-based clustering, a process of searching for the parameters that maximize the likelihood of observing the complete set is carried out by means of a two-step Expectation Maximization algorithm (EM) (Jung et al., 2014). In a first step the likelihood that each instance belongs to each of the mixtures is calculated. In a second step, the parameters of the mixtures and their weights are updated trying to maximize the overall likelihood (the so-called E-Step). These two steps are repeated multiple times, until the likelihood does not change, or until the changes in likelihood are negligible (the M-Step). Figure 2 below lays out the entire process implied by Equations 1-2.

Insert Figure 2 near here

In the GMM algorithms, there is a robust statistical criterion that assists the analyst in the selection of the optimal value of *K*: the classical BIC. The BIC is defined as

$$BIC \equiv ln\ ln\ (n)\ \cdot p - 2 \cdot ln\ ln\left(f\ (\theta)\right) \qquad (3)$$

where $f\ (\theta)$ is given by Equation (2), *n* is the number of observations, and *p* is the number of parameters of the model (the number of parameters of all Gaussians in the case that concerns us). The smaller the value of the BIC, the stronger the evidence in favor of the corresponding model; i.e. BIC prefers simple models (a smaller number of parameters *p* implies a lower value of $ln\ ln\ (n) \cdot p$ (given that *n* is constant) that have high likelihood (the term $-2 \cdot ln\ ln\left(f\ (\theta)\right)$ decreases when the likelihood increases). The number of clusters is not considered an independent parameter for the purposes of computing the *BIC*. By calculating the *BIC* for different values of K and looking for the one that minimizes Equation (3), we can find the optimal number of clusters according to this criterion. Although Equations (1–3) may appear complicated, the steps involved in our modelling approach can be easily summarized as Figure 3 below clearly illustrates.

Insert Figure 3 near here

For the problem under study, we shall use the ***MCLUST*** package developed by Fraley and Raftery (1998), for which Scrucca et al. (2016) designed the R language for the application of clustering based on GMMs. We perform clustering with *K = 1* up to *K = 10*, and we will evaluate the quality of each clustering configuration by using the *BIC* criterion. Our focus here is to describe the KE according to four dimensions

outlined previously. Here the unit of observation is the pair country (*i*)-year(*t*). In this context, countries may stay in the same cluster over the entire period or move up or down to other clusters. We present and discuss the results of the analysis next.

4. Findings and Discussion of Results

After imputation, we ran the GMM-based clustering over the data trying from 1 to 10 different models. The best solution for the clustering, based on the BIC criterion (see Equation 3), was obtained for 6 models. A summary of the selected mixture model is displayed in Table 3.

Insert Table 3 near here

The classification of each cluster seems to be balanced. The first cluster contains 273 observations, the second 131, the third 143, the fourth 326, the fifth 164, and the sixth 113. For instance, 28 % of the data points belong to cluster 4, and only 10 % belong to cluster 6. Based on this statistical criterion, we select six clusters. Recall that we should look for the model that maximizes BIC (see Table 3, the best model is the VEV = ellipsoidal, equal shape) with 6 clusters. The parameterization of the covariance matrix, *VEV* means variable volume, equal shape, and variable orientation. In our application, we get a BIC value equal to −14513.05. The Integrated Completed Likelihood criterion[6] (ICL= -14624.42) is nearly identical to the BIC, implying that the E-Step and M-Step generated stable probabilities. Figure 4 displays the spherical plots for the initial classification for six clusters. The figure is a result of a principal components analysis (PCA) that projects our data on the first two principal components (linear combination of our original variables). The two dimensions (Dim2 on the y-axis and Dim1 on the x-axis) capture the most variation in our data (around 58%). The first component explains 44% of the variation and the second one accounts for 14%. This may sound too technical for our non-expert readers, but the main point is simple: There are clusters in the data.

Insert Figure 4 near here

By examining the entropy values for the *K* clusters, merging from 6 to 4 clusters is necessary since the decrease in entropy is large. Note that the entropy is only an exploratory tool that can help us to separate the clusters rather than a formal inference tool (see, for instance Baudry et al., 2010).[7] The lower entropy coefficient means better clustering. Moreover, from a simple inspection of the radar plots (see Figures 5a and 5b), we can see that Cluster 3 and Cluster 4 are quite similar. Further analysis of these two groups reveals that Cluster 4 performs badly in the education dimension (in particular, secondary, and tertiary school enrolment). Both clusters are similar in the ICT dimension and perform equally in terms of scientific output. If we look at the institutional variables, Group 3 is slightly better than Group 4, but Group 4 is better in terms of trade openness than Group 3. Lastly, we also merge cluster 1 with the previous merged cluster 3 and 4 resulting in four clusters. Table 4 provides the means of clusters.

***Insert Figures 5a and 5b near here side-by-side ***

Insert Table 4 near here

Having grouped our six initial clusters into four, we shall now perform some statistical tests to check if the clusters are statistically different. We apply a Kruskal-Wallis test (henceforth, KKW, see Kruskal and Wallis, 1952)[8] for each of numerical variable in our dataset. The results of the KKW test are statistically significant (p value < 0.01). Finally, for pairwise comparison across clusters, we carry out a Wilcoxon test (see Bauer, 1972). The results also display that there are statistically significant differences between paired clusters (1-2, 2-3, 3-4, 1-3, 1-4, and 2-4) with p-values < 0.01.

Since two of the six initial clusters are fuzzy (nearly similar), we can categorize countries into four groups according to their transition towards a KE. However, we do not neglect to assign certain labels to these clusters instead of relying solely on descriptions to explain each profile. Let us describe each of the clusters:

- Cluster 1 *(Very prepared*): This cluster contains the pair country and year observations that perform the best in each of the KE dimensions. They are national economies with high quality institutional frameworks. They perform slightly better in terms of innovation output (patents and scientific articles). They also perform better in terms of educational variables and are also strong in terms of ICT indicators.

- Cluster 2 (*Prepared*): This cluster group follows the country-year observations in Cluster 1. There is a clear hierarchy according to the mean of the features of each KE dimension. For instance, there is significant difference in the education indicators of Cluster 2 compared to those of Cluster 1. The innovation variables are similar in terms of patents per capita, but there is a difference in performance related to scientific articles.

- Cluster 3 (*Unprepare*d): This cluster shows up low performance in education (low primary, secondary, and tertiary school enrolment), and in innovation (low number of patents and scientific articles per capita). In relation to institutional variables, they show quite similar values on average to Cluster 2. The countries classified in this cluster show low values of ICT (internet users and broadband internet penetration).

- Cluster 4 (*Very unprepared*): This cluster groups has low performance in each KE dimension.

Whereas Table 4A in the Appendix presents a complete country-year picture of the composition of the four clusters, Table 5 compares Algeria and Botswana as an example. The table tells us that different economies are differently prepared for different reasons and at different times, i.e., clusters are not static. Algeria started off unprepared for transition to the KE, held back by the weakness of all the dimensions of KE. Improvements in the education and innovation dimensions allowed the transition to Cluster 2, but it was not until the quality of institutions made possible by the end of civil strife and national reconciliation (Hamdy, 2007). Botswana on the other hand took a different path; the country's stable and good quality of institutions permitted it to jump from being unprepared (Cluster 3) to being very prepared (Cluster 1). However, because, the education, innovations, and ICT dimensions of KE in Botswana remained weak, the quality of institutions has been the primary driving force. It makes good sense then that it is sensitive to shocks (real or perceived) to the general economy which in turn affected KE so that in 2009, 2011, and

2015 Botswana's KE slipped back to Cluster 2. These findings are consistent with what we discern from UNESCO's reports and from Hamdy (2007), Isaacs (2007) and Ouedraogo et al. (2021).

Insert Table 5 near here

Lastly, Figure 6 displays the classification tree representation for African countries. Classification trees are easy to interpret and can give us an indication on which variable is more relevant within each cluster (see, Samoilenko and Osei Bryson, 2008). This is, by no means, a way to validate our cluster formation; it is a way of providing plausible explanations that are easily interpretable by human beings of what variables can explain to which cluster each country belongs. The subsets created by the splits are nodes and the subsets which are not split are called terminal nodes. Each terminal node is assigned to one of our two labels (prepared and unprepared). At the top is the root node with the TELEP3 variable. If FIXBI is lower than 6.915, then it goes down to another node. There we ask if the level of patents per capita is lower than 0.004725, then we classify the observations in the cluster unprepared and prepared. For the sake of interpretation, this final node tells us that 897 observations fall in the unprepared class, and 14 are incorrectly classified (Table 6). *According to our decision trees, 180 observations in the cluster prepared can be explained with ICT variables, and only in 22 cases with the innovation indicators (patents and scientific publications per capita).*

Insert Figure 6 near here

We also computed the overall accuracy of our model, and the accuracy is of 97%.[9] Table 5 displays the confusion matrix for our classifier once we merged all clusters into two categories.

Insert Table 6 near here

Table 6 displays that 899 observations were predicted in the class unprepared, and it turns out to be true. Similarly, 225 observations were predicted in the class prepared, and it turns out to be true. Seven observations can be classified as False Positive, and 19 observations as False Negative. *This classification tree has also remarkably interesting policy implications; it gives us relevant information on how countries end up in a similar stage of KE using different paths*. These results demonstrate clearly that one-size-fits-all policies are mistaken; African KE's are similar, but not identical and hence require specific policies targeted to dimensions in which weakness lies. The findings also show that observed differences are irrespective of the level of development shown in Table 1, political institutions, or region.

5. Conclusions

The aim of this paper is to apply a GMM methodology to provide an alternative strategy to using simple composite indices to capture the level of the KE across countries over time. As a manageable objective for accomplish the goal we perform GMM clustering. Countries were grouped by the classical dimensions that characterize the KE at the country level according to the World Bank's KAM that consists of four dimensions: education, economic and institutional regimes, ICT, and the innovation. The method we employ is a promising technique that better aligns an empirical approach to understanding the KE with recent theoretical frameworks. Subsequent clustering analysis obtained: ***very prepared (Cluster 1),***

prepared (Cluster 2), unprepared (Cluster 3), and very unprepared (Cluster 4). Further analysis identified the evolution of any country over time as a country can belong to a different cluster in different years for different reasons. The results are consistent with the literature on both technology gaps and technology clubs. A simple interpretation of the latter is that technology clubs are dynamic, not static. The former suggests that technology gaps can exist within countries belonging to the same clubs.

The results clearly indicate that nations, irrespective of their different levels of economic development, face different issues in their KE pillars. Blanket policies may be necessary, but they are inadequate promotors of KE. In the light of the findings, we conclude that not all African countries are KEI-challenged; countries in Cluster 1 did well at least in some years. However, the evolution of KEI even in this cluster is not static; it depends on changes in the features of the pillars. *If this result holds that KEI is generally dynamic, that countries move in and out of clusters, then the World Bank's ranking of countries by their KE levels is severely misleading as it might well be that some developing countries have stronger KEs than developed countries in some years*. This is an area that needs further inquiry because our findings do not agree with previous studies which conclude that technology clubs are dynamic only for high-income countries and static only for low-income countries. An important implication of these findings is that pooling one class overlooks the heterogeneity in the KE process and leading to incorrect conclusions about the KE in African countries. Moreover, the findings are not optimistic. Most African countries are not prepared for progress towards KE.

Our research methodology helps us clarify further what we know of IT for development in that both the use of an advanced clustering technique, and the robust imputation of the missing values, have allowed us to find meaningful clusters of countries that helped us to understand better the KE phenomenon in Africa. Particularly, we now understand better and more clearly the dimensions or forces behind the transition of African countries towards a KE, and dimensions (institutional, education, ICT, and innovation) that each African country should reinforce to facilitate its transition towards a better KE, and hence sustainable social and economic development.

Two extensions from a methodological point of view can be considered. First, the choice of the variables in each dimension can be a bit arbitrary and result in selection bias in the clustering method because they are only proxy variables for each dimension. Even so, we see two directions for further examinations. One, future research can explore additional or alternative ways of measuring each dimension of the KE at the national level. In this respect, the methodology presented by Fop and Murphy (2018) can be employed for variable selection within model-based clustering. They compare models with different variables based on the BIC criterion. They employ a similar approach to stepwise regression. For that purpose, they employ a forward/backward and backward/forward feedback. While we are aware of potential benefits from alternative methodologies, we took them for granted in favor of the obvious advantage offered by the KAM dimensions. Even with this weakness, our approach is novel, the problem we addressed is relevant, and the results we obtained are robust and informative to both policy and future research.

Secondly, another extension of the current work is the use of a latent profile analysis in a panel data context. This approach possesses superior complexity, accounting for both time and cross-sectional dimension, and capturing the variability in each KE indicator over time. These models have been largely investigated in

other fields such as: applied statistics, and biostatistics. Nevertheless, little attention has been paid on socio-economic applications. Some exceptions are Fruhwirth-Schnatter and Kaufmann (2008), and Fruhwirth-Schnatter (2011). As more data concerning the variables under the study becomes available for African countries additional studies should be conducted to improve the robustness of these findings.

The potential implications of our paper in terms of ICT for development, as we discussed earlier include the fact that ICT is a key driver for economic growth and wealth generation. Our results show that only a small group of African countries perform well in the ICT dimension. These group of countries are experiencing rapid growth in the adoption of ICT. We also show that broadband is not relevant variable for this group at this stage, but total number of mobile subscribers would be more important dimensions for both policy and further research to stress. It is also important to notice that prepaid mobile cards subscribers are a quite importance due to weak ICT infrastructure and lack of optical cables in these countries. Future research should take this into consideration because of their implications for development in those countries.

Finally, this methodology can be valuable from a managerial point of view as it can be used as an additional tool in the decision-making process; it allows managers to identify the current country's stage, its evolution over time, and what needs done to facilitate its progress. This is a benchmarking methodology that allows us to compare results with other reference countries that belongs to the same group and to learn from their best practices. This analytical approach can also help policy makers in a similar vein of the composite indices of KE constructed by international organizations, but with an added benefit of a robust classification even in the presence of missing data as often is the case in developing, especially African, countries. Indeed, we can also simulate different scenarios that could help to anticipate the group of the countries according to the values of the variables employed to empirically measure the KE.

6. Appendix [at end]

Notes

1. The OECD defines KBEs at a very general level as economies that are directly based on the production, distribution, and use of knowledge and information (OECD, 1996, p.7).
2. They employ the Technological Readiness Index and the technological innovation index from the World Economic Forum, the Technological Advanced index edited by UNIDO (United Nations Industrial Development, the Global summary index from the European Commission, the Technological Activity Index (TAI) FROM UNCTAD, and ArCO (Archibugi and Coco, 2004).
3. Benchmarking can be defined as a sequence of activities that involves process and assessment (see Watson, 1993).
4. Recent applications of mixture models in different contexts can be found in Csereklyei et al. (2017), Sulkowski and White (2016), Alfo et al. (2008), Seo and Thorson (2016), Kumar (2019), and Clements (2020).
5. Available at https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups
6. See for further details, Biernacki et al. (2000).
7. Details available upon request.
8. This is a non-parametric statistical test that assesses the differences among three or more independently sampled groups on a single, non-normally distributed continuous variable,
9. R code is available upon request. The classification tree has been generated with rpart package (Therneau and Atkinson, 2019). rpart stands for recursive portioning and regression trees. In our context given that our variable is a factor then we deal with a classification tree. By default, rpart () function uses the **Gini** impurity measure to split the node. The higher the Gini coefficient, the more different instances within the node.

Acknowledgements

We would like to thank four anonymous referees and the two editors for their constructive comments. We would also like to thank participants at the *First International Conference on Regional Integration "Taking action to approaching regional integration: Africa, Mediterranean and European Union* in a Global Age", *WEAI's Virtual International Conference*, and the *Future of Growth Conference* for valuable comments on previous versions of this paper. Data collection assistance was provided by Ms. Nilufar Safarova. This paper is dedicated to the memory of both my parents for always loving and supporting me during my academic career: Francisco Rodríguez Gude and Esperanza Andrés González who died from COVID-19 in 2020. The usual caveat applies.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Financial support from the Czech Science Foundation (GA19-25280S), from the Ministry of Science, Innovation and Universities of Spain, and the European Regional Development Fund of the European Commission (No. RTI2018-095324-B-I00) is gratefully acknowledged.

Code and data availability

The R source code and the original dataset are available at:
https://github.com/antonio1970/Clustering-Algorithms/tree/master/code.

The dataset was assembled from the public sources listed in Table 1A.

References

Abad-González, J. and Martínez, R. (2017). Endogenous categorization of the human development. *Applied Economics Letters*, 24:4, 243–246.

Acuña, E. and Rodríguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In: Banks D., McMorris F.R., Arabie P., Gaul W. (eds). *Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation*. Springer, Berlin, Heidelberg.

Agyapong B. G., and Oseifuah, E.K. (2015). Knowledge and economic growth: A comparative analysis of three regional blocks in Sub-Saharan Africa. *Environmental Economics*, 6(4-1): 196–208.

Ahlquist, J., and Breunig, C. (2012). Model-based clustering and typologies in the social sciences. *Political Analysis*, 20: 92–112.

Alfo, M., Trovato, G., and Waldmann, R. (2008). Testing for country heterogeneity in growth models using a finite mixture approach. *Journal of Applied Econometrics*, (4): 487–514.

Amavilah, V. (2006). Non-parametric diversity indices of technical capability of African countries. *African Development Review*, 2(18): 205–220.

Amavilah, V., Asongu, A., and Andrés, A.R (2017). Effects of globalization on peace and stability: Implications for governance and the knowledge economy of African countries. *Technological Forecasting and Social Change*, 122: 91–103.

Andrés, A.R, Amavilah, V, and Asongu, A. (2017). Linkages between Formal Institutions, ICT Adoption, and Inclusive Human Development in Sub-Saharan Africa. ***Catalyzing development through ICT adoption: The Developing World Experience***. Chapter 10. pp. 175-203. Editors. Dr. Ewa Lechman, Dr. Adam Marszk, and Dr. Harleen Kaur. Springer Verlag.

Andrés, A.R., Asongu, S.A., and Amavilah, V.H. (2015). The impact of formal institutions on the knowledge economy. *Journal of Knowledge Economy*, 6(4): 1034–1062.

Anyanwu, J. (2012). Developing knowledge for economic advancement in Africa. *International Journal of Academic Research in Economics and Management Sciences*, Vol. 1: 73–111.

Archibugi, D., and Coco, A. (2004). A new indicator of technological capabilities for developed and developing countries (ArCo). *World Development*, 32(4): 629–654.

Archibugi, D., and Coco, A. (2005). Measuring technological capabilities at the country level: A survey and a menu for choice. *Research Policy*, 34(2): 175–194.

Archibugi, D., Denni, M., and Filippetti, A. (2009). The technological capabilities of nations: The state of the art of synthetic indicators. *Technological Forecasting and Social Change*, 76(7): 917–931.

Arrow, K. (1962). Economic Welfare and the Allocation of Resources to Invention. In The Rate and Direction of Inventive Activity: Economic and Social Factors, edited by the Universities- National Bureau Committee for Economic Research and the Committee on Economic Growth of the Social Science Research Councils, 609–26. Princeton, NJ: Princeton University Press.

Asongu, S., Andrés, A.R., and Amavilah, V. (2019). Business dynamics, knowledge economy, and the socio-economic development of African countries. *Information Development,* 36(1): 128–152

Asongu, S., Tchamyou, V., and AchaAnyi, P. (2018). Who is who in knowledge economy in Africa? *Journal of the Knowledge Economy*: 1–33.

Azam, J.P., Fosu, A., and Ndung'u, N.S. (2002). Explaining slow growth in Africa. *African Development Review*,14(2): 177–220.

Balijepally, V., Mangalaraj, G., and Iyengar, K. (2011). Are we wielding this hammer correctly? A reflective review of the application of cluster analysis in information systems research. *Journal of the Association for Information Systems*, 12(5), 375–413

Barro, R. (1991). Economic growth in a cross section of countries. *The Quarterly Journal of Economics*, 106(2): 407–443.

Barro, R.J., and Lee, J.W. (1993). International comparisons of educational attainment. *Journal of Monetary Economics*, 32(3): 363–394.

Baudry, J.P., Raftery, A., Celeux, G., Lo, K., Gottardo, R. (2010). Combining mixture components for clustering models. *Journal of Computational and Graphical Statistics*, 19 (2): 332–353.

Bauer, D.F. (1972). Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, 67: 687–690.

Becker, M. (2005). Development: Josep A. Schumpeter. *Journal of Economic Literature*, 1(XLII):108–111.

Beretta, L., and Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16(3. Suppl 74):198–208.

Bezdek, J. C., and Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(3), 301–315.

Biernacki, C., Celeux, G., Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Blackburn, K., Forgues-Puccio, G.F. (2010). Financial liberalization, bureaucratic corruption and economic development. *Journal of International Money and Finance*, 29(7), 1321–1339.

Borgmann, A. (2006). Technology as a cultural force: For Alena and Griffin. *The Canadian Journal of Sociology*, 31(3), 351– 360.

Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Kindle Direct Publishing, 1st Edition.

Burnside, C., and Dollar, D. (1997). Aid policies, and growth. *World Bank Policy* Research Working Paper (1777).

Carayannis, E.G., Clark, S.C., Valvi, D.E. (2013). Smartphone Affordance: Achieving Better Business Through Innovation. *Journal of the Knowledge Economy*, 4, 444–472.

Carter, A.P. (1996). Measuring the performance of a knowledge-based economy. In D. Foray, & B.A. Lundvall (Eds). Employment and growth in the knowledge-based economy (pp. 61-68). Paris. Organisation for Economic Cooperation and Development.

Castellacci, F. (2011). Closing the technology gap. *Review of Development Economics*, 15(1): 180–197.

Castellacci, F., and Archibugi, D. (2008). The technology clubs: The distribution of knowledge across nations. *Research Policy*, 37: 1659–1673.

Chavula, H.K. (2013). Telecommunications development and economic growth in Africa. *Information Technology for Development*, 19(1): 5–23

Chen, J., and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95 (3): 759–771.

Chen, H., and Dahlman, C. (2006). The Knowledge Economy, the KAM Methodology, and World Bank Operations. World Bank Institute Working Paper 37256. The World Bank, Washington D.C.

Clement, J. (2020). Social protection clusters in sub-Saharan Africa. *International Journal of Social Welfare*, 29: 20–28. https://doi.org/10.1111/ijsw.12378

Collier, P. (2007). Africa's economic growth: Opportunities and constraints. *African Development Review*, 19(1): 6–25

Collier, P., and Gunning, J.W. (1996). Policy towards commodity shocks in developing countries. IMF Working Paper 84.

Collier, P., and Gunning, J.W. (1999a). Explaining African performance. *Journal of Economic Literature*, 37(1): 64–111.

Collier, P., Gunning, J.W. (1999b). Why has Africa grown slowly? *Journal of Economic Perspectives,* 13(3): 3–22.

Conway, P., and Greene, J. (1993). Is Africa different? *World Development*, 21 (12): 2017– 2028.

Csereklyei, Z., Thurner, P., Langer, J., and Kuchenhoff, H. (2017). Energy paths in the European Union: A model-based clustering approach. *Energy Economics*, 65: 442–457.

Datta, A., and Agarwal, S. (2004). Telecommunications and economic growth: A panel data approach. *Applied Economics*, 36(15): 1649–1654.

Devarajan, S. (2013). Africa's statistical tragedy. *Review of Income and Wealth*, 59(1): S9–S15.

Dodgson, M., and Gann, D. (2018). *Innovation: A Very Short Introduction*. 2nd edition. Oxford: Oxford University Press.

Dolowitz, D.P., and Marsh, D. (2000). Learning from abroad: the role of policy transfer in contemporary policy making. *Governance*, 13(1), 5–23

Driouchi, A., Azelmad, E.M., and Anders, G.C. (2006). An econometric analysis of the role of knowledge in economic performance. *Journal of Technology Transfer*, 31: 241–255.

Drucker, P. (1969). *The Age of Discontinuity*. New York: Harper Business.

Easterly, W. (2001). Can institutions resolve ethnic conflict? *Economic Development and Cultural Change*, 49: 687–706.

Easterly, W., and Levine, R. (1997). Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics*, 112(4): 1203–1250.

Englebert, P. (2000). Solving the mystery of the African dummy. *World Development*, 28(10): 1821–1835.

European Commission (2020). *Develop an Upgraded Single Market Scoreboard as a Governance Tool for the Single Market*. Available at https://ec.europa.eu/internal_market/scoreboard/

Fop, M., and Murphy, T. (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, 12: 18–65.

Fraley, C., and Raftery, A. (1998). How many clusters? which clustering method? Answer via model-based cluster analysis. *The Computer Journal*, 8(41): 578–588.

Fraley, C., and Raftery, A. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97: 611–631.

Fruhwirth-Schnatter, S. (2011). Panel data analysis: A survey on model-based clustering of time series. *Advances in Data Analysis and Classification. Theory, Methods, and Applications in Data Science*, 5: 251–280. Springer.

Fruhwirth-Schnatter, S., and Kaufmann, S. (2008). Model based-clustering of multiple time series. *Journal of Business & Economic Statistics*, 26: 78–89.

Grant, D., and Yeo, B. (2018). A global perspective on tech investment, financing, and ICT on manufacturing and service industry performance. *International Journal of Information Management*, 43, 130–145.

Grossman, G., and Helpman E. (1991). *Innovation and Growth in the Global Economy* Cambridge, Mass, MIT Press.

Gyimah-Brempong, K., and Wilson, M. (2004). Health human capital and economic growth in Sub-Saharan African and OECD countries. *Quarterly Review of Economics and Finance*, 2(44): 296–320.

Hair, Jr., J.F., Black, W.C., Babin, B.J., Anderson, R.E., and Tatham, R.L. (2006). *Multivariate Data Analysis*, 6th Edition. Upper Saddle River (NJ): Pearson Prentice Hall. Chapter 8.

Hamdy, A. (2007). Survey of ICT and Education in Africa: Algeria Country Report. InfoDev ICT and Education Series. World Bank, Washington, DC. World Bank. https://openknowledge.worldbank.org/handle/10986/10683 License: CC BY 3.0 IGO.

Havens, T. C., Bezdek, J. C., Keller, J. M., and Popescu, M. (2008, December). Dunn's cluster validity index as a contrast measure of VAT images. In *2008 19th International Conference on Pattern Recognition* (pp. 1-4). IEEE.

Isaacs, S. (2007). Survey of ICT and Education in Africa: Botswana Country Report. InfoDev ICT and Education Series. World Bank, Washington, DC. World Bank. https://openknowledge.worldbank.org/handle/10986/10713 License: CC BY 3.0 IGO.

Jain, A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.

Jerven, M. (2011). The quest for the African dummy: Explaining African post-colonial economic performance revisited. *Journal of International Development*, 23(2): 288–307.

Jerven, M. (2013). *Poor Numbers: How we are misled by African development statistics and what to do about it*. Ithaca. Cornell University Press

Jung, Y.G., Kang, M.S, and Heo, J. (2014). Clustering performance comparison using K-means and expectation maximization algorithms. *Biotechnology & Biotechnological Equipment*, 28, sup1: S44–S48.

Kaplinsky, R. (2005). *Globalization, Inequality, and Poverty: Between a Rock and a Hard Place*. Cambridge, UK: Polity.

Knedlik, T., and Reinowski, E. (2008). The African growth gap, development policy, and the realization of the MDGs. New In book: *African Development Perspectives Yearbook*, No. 13 Edition: Publisher: LIT Verlag Editors: Wohlmuth et al

Kochetkov, D. M., and Vlasov, M. V. (2016). Teoretiko-metodologicheskie podkhody k analizu ekonomiki znanii na regional'nom urovne [theoretical and methodological approaches to the analysis of the knowledge economy at the regional level]. *Zhurnal ekonomicheskoi teorii [Russian Journal of Economic Theory], 4*, 242–247.

Krugman, P. (2013). The New Growth Fizzle. New York Times.

Kruskal, W.H., and Wallis, W.A (1952). Use of ranks in one criterion variance analysis. *Journal of American Statistical Association*, 47: 583–621.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software. Articles*, 28(5):1–26.

Kumar, N. (2019). A model-based clustering approach for analyzing energy related financial literacy and its determinants. CER-ETH Economics working paper series 19/312, CER-ETH - Center of Economic Research (CERETH) at ETH Zurich.

Landes, D. (1998) *The Wealth and Poverty of Nations: Why Some Are So Rich and Some So Poor*. New York: **W.W. Norton.**

Levine, E., and Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11): 2573–2593.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Second Edition. *John Wiley & Sons*.

Lucas R. (1988). On the mechanics of economic development. *Journal of Monetary Economics,* 22: 3– 42.

Lucas, R. (1993). Making a miracle. *Econometrica*, 61(2): 251–272.

Lundvall, B. Å., Joseph, K., Chaminade, C. and Vang, J. (2009). *Handbook on Innovation Systems and Developing Countries: Building Domestic Capabilities in a Global Setting*, Edward Elgar.

Machlup, F. (1962). *The Production and Distribution of Knowledge in the United States*. Princeton. Princeton University Press

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Academic Press London. New York.

Marfo, K., Pence, A., LeVine R.A., and LeVine, S. (2011). Strengthening Africa's Contributions to Child Development Research: Introduction. *Child Development Perspectives*, 5, 104–111.

Márquez, D.G., Félix, P., García, C. A., Tejedor, J., Fred, A. L., and Otero, A. (2019). Positive and negative evidence accumulation clustering for sensor fusion: An application to heartbeat clustering. *Sensors*, *19*(21), 4635.

Mauro, P. (1995). Corruption and economic growth. *The Quarterly Journal of Economics*, 110(3): 681–672.

Mitchell, R., Rose, P., and Asare, S. (2020). Education research in Sub-Saharan Africa: Quality, visibility, and agendas. *Comparative Education Review*, 64(3), 363–383.

Neath, A. A., and Cavanaugh, J.E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 199–203.

North, D.C. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

OECD (2005). *Micro-policies for Growth and Productivity: Synthesis and Benchmarking user guid*e. Paris. OECD

OECD (1996). *The Knowledge Based Economy*. Paris: OECD.

Ouedraogo. R., Sourouema W.S., and Sawadogo, H. (2021). Aid, growth. and institutions in Sub-Saharan Africa: New insights using a multiple growth regime approach. *World Economy*, 44:107–142.

Parcero, O.J., and Ryan, J.C. (2017). Becoming a knowledge economy: The case of Qatar, UAE, and 17 benchmark countries. *Journal of the Knowledge Economy*, 8, 1146–1173

Paz-Marín, M., Gutiérrez, P.A., and Martínez, C.H. (2018). Classification of countries' progress toward a Knowledge economy based on machine learning classification. *Expert Systems with Applications*, 42: 562–672.

Powell, W.W. and Snellman, K. (2004). The Knowledge Economy. *Annual Review of Sociology*, 30:1, 199–220

Qureshi, S. (2013). What is the role of mobile phones in bringing about growth? *Information Technology for Development*, 19(1): 1–4.

Rao, L., and McNaughton, M. (2019). A knowledge broker for collaboration and sharing for SIDS: the case of comprehensive disaster management in the Caribbean. *Information Technology for Development*, 25(1), 26–48.

Rebelo, S.T. (1990). Long run policy analysis and long run growth (No. w3325). National Bureau of Economic Research.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rizk, N, El Said, A., Weheba, N., De Beer, J. (2017). Towards an Alternative Assessment for Innovation in Africa. Working Paper, Open AIR.

Romer, P. (1986). Increasing returns and long run growth. *Journal of Political Economy,* 94(2): 1002–1037.

Romer, P.M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5): S71–S102.

Samoilenko, S., and Osei-Bryson, K.M. (2008). Increasing the discriminatory power of DEA in the presence of the sample heterogeneity with cluster analysis and decision trees. *Expert Systems with Applications*, 34(2), 1568-1581.

Schumpeter, J.A. (2005). Development. *Journal of Economic Literature*, 43(1): 108–120.

Scrucca, L., Fop, M., Murphy, T., and Raftery, A. (2016). mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R Journal*, 1(8): 289–317.

Seo, H., and Thorson, S. (2016). A mixture model of global internet capacity distributions. *Journal of the Association for Information Science and Technology*, 67(8): 2032–2044.

Shapira, P., Youtie, J., Yogeesvaran, K., and Jaafar, Z. (2006). Knowledge economy measurement: Methods, results, and insights from the Malaysian knowledge content study. *Research Policy*, 35(10): 1522–1537.

Širá, E., Vavrek, R, Kravčáková V. Ivana, Kotulič, R. (2020). Knowledge economy indicators and their impact on the sustainable competitiveness of the EU countries. *Sustainability*, 12(10).

Stigler, G. (1961). The economics of information. *Journal of Political Economy, 69*(3), 213–225.

Stiglitz, J. (2015). Leaders and followers: Perspectives on the Nordic model and the economics of innovation. *Journal of Public Economic*, 127: 3–16.

Sulkowski, A., and White, D. (2016). A happiness Kuznets curve? Using model-based cluster analysis to group countries based on happiness, development, income, and carbon emissions. *Environment, Development, and Sustainability*, 18(4):1095–1111.

Therneau, T., and Atkinson, B. (2019). rpart: Recursive
Partitioning and Regression Trees.
R package version 4.1–15. https://CRAN.R-project.org/package=rpart

Thompson, M., and Walsham, G. (2010). ICT Research in Africa: Need for a Strategic Developmental Focus. *Information Technology for Development*, 16(2): 112–127.

Torero, M., and von Braun, J.(ed.) (2006). Information and communication technologies for development and poverty reduction: The potential of telecommunications. IFPRI books, International Food Policy Research Institute (IFPRI), number 0-8018-8041-6, December.

Thurow, L. (2000). Globalization: The product of a kbowledge-based economy. *The Annals of the American Academy of Political and Social Science, 570*, 19–31.

Tripathi, M., and Inani, S.K. (2020). Does information and communications technology affect economic growth? Empirical evidence from SAARC countries. *Information Technology for Development*, 26(4): 773–787.

UNESCO Institute of Statistics (Online) Botswana: Education and literacy. http://uis.unesco.org/en/country/bw. Accessed 3/6/2021.

UNESCO Institute of Statistics (Online), Botswana: Science, Technology, and Innovation. http://uis.unesco.org/en/country/bw?theme=science-technology-and-innovation. Accessed 3/6/2021.

UNESCO Institute of Statistics (Online) Algeria: Education and literacy. http://uis.unesco.org/country/DZ/. Accessed 3/6/2021.

UNESCO Institute of Statistics (Online) Algeria: Science, Technology, and Innovation. Accessed 3/6/2021.

van Biljon, J., Osei-Bryson, K.W (2020). The communicative power of knowledge visualizations in mobilizing Information and Communication Technology Research. *Information Technology for Development* 26(4): 637–652.

Verbeek, J., Vlassis, N., and Kröse. B. (2003). Efficient greedy learning of Gaussian mixture models. *Neural Computation*, 15 (2): 469–485.

Watson, G.H. (1993). *Strategic Benchmarking: How to Rate Your Company's Performance against the World's Best*. Wiley, New York, NY.

Weber, A.S. (2011). The role of education in knowledge economies in developing countries. *Procedia Social and Behavioral Sciences*, 15: 2589–2594.

World Bank (2007). *Building Knowledge Economies. Advanced strategies for Development*. WBI Development Studies. The World Bank. Washington D.C.

World Bank (2008). *Measuring World Knowledge in the World's economies. Knowledge Assessment Methodology and Knowledge Economy Index*. Knowledge for Development Program. World Bank Institute

World Bank (2012). *Knowledge for Development (k4d), Knowledge Assessment Methodology 2012*. Technical report.

World Bank (2016). *World Development Report: Digital Dividends*. World Bank, Washington.

World Bank (2019). World Bank country and lending groups. World Bank. Available at https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups. Accessed on 10/11/2019.

Xiao, J., Lu, J., and Li, X. (2017). Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis*, *21*(6), 1327–1338.

Xiong, J., Qureshi, S., and Najjar, L. (2014). A Cluster Analysis of research in Information Technology for global development: Where to from here? AIS Electronic Library.

Zhou, H.B., and Gao, J.T. (2014). Automatic method for determining cluster number based on silhouette coefficient. In *Advanced Materials Research* (Vol. 951, pp. 227–230). Trans Tech Publications Ltd.

Figure 1. The four pillars of knowledge economy. Source: Own elaboration.

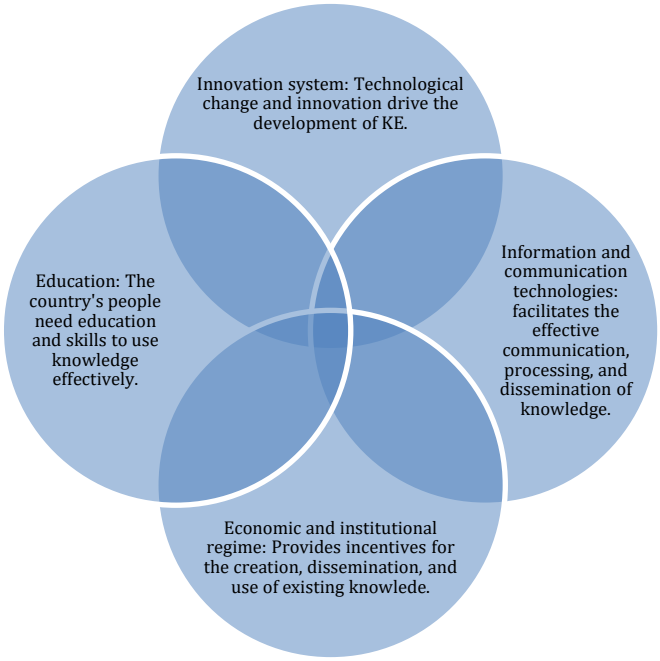

Innovation system: Technological change and innovation drive the development of KE.
Education: The country's people need education and skills to use knowledge effectively.
Information and communication technologies: facilitates the effective communication, processing, and dissemination of knowledge.
Economic and institutional regime: Provides incentives for the creation, dissemination, and use of existing knowlede.

Figure 2. Analytic process in mixture model clustering. Source. Own elaboration

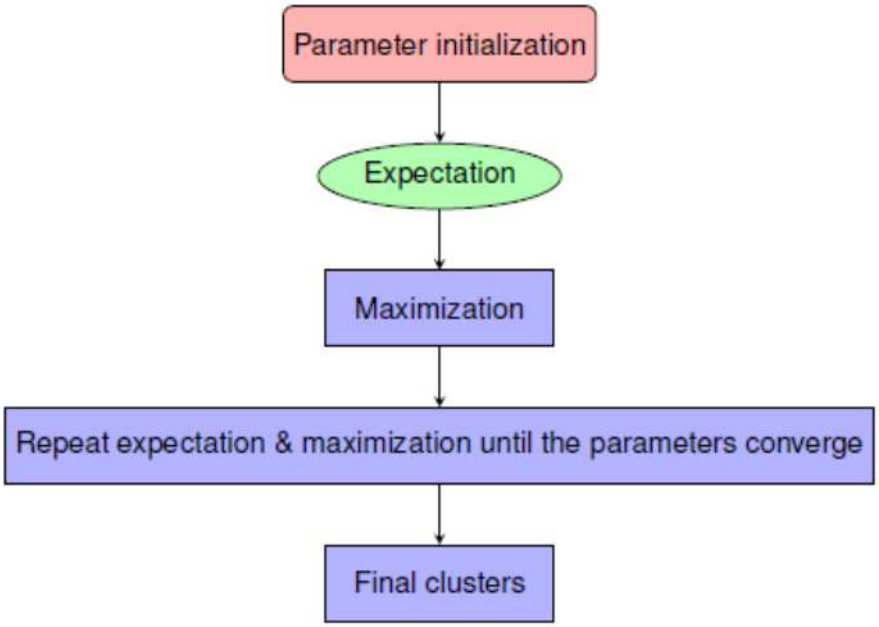
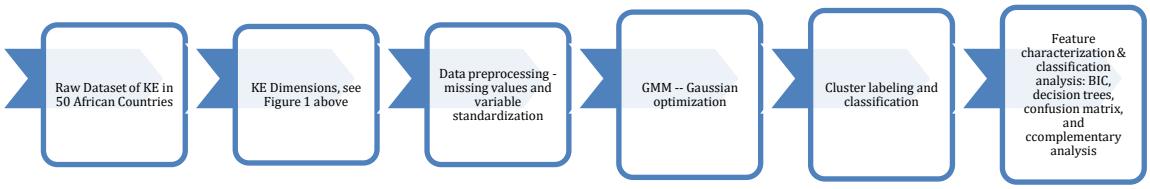
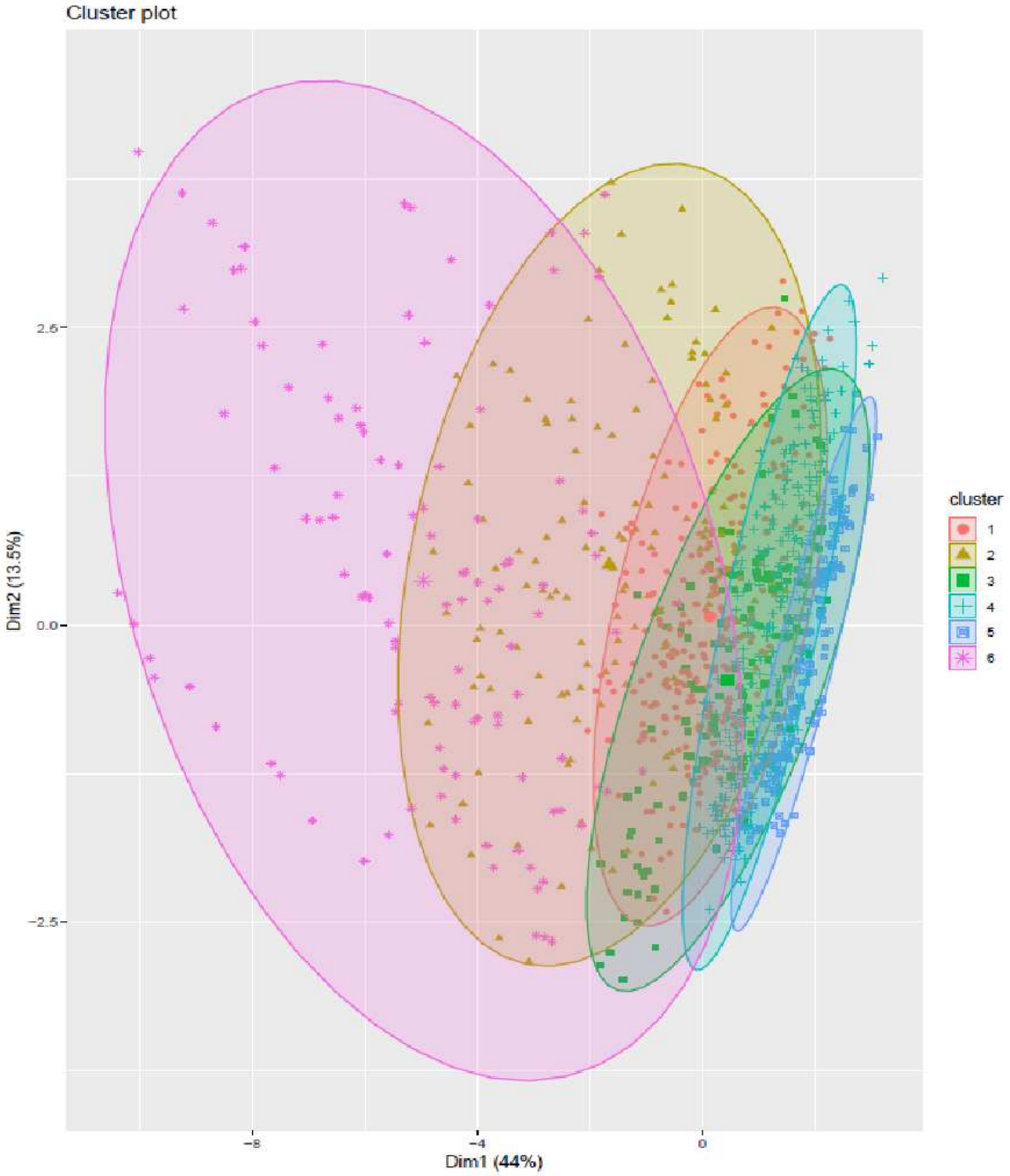


Figure 3. GMM analytical steps.



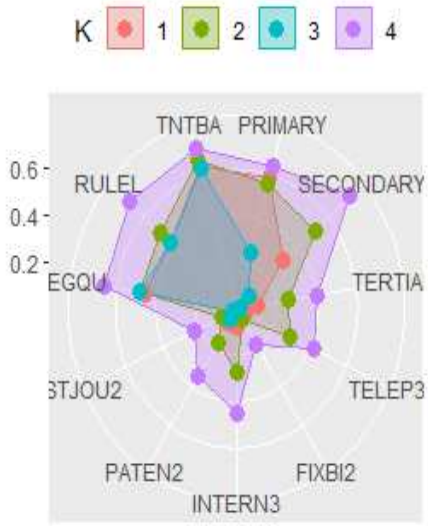
Source: own elaboration

Figure 4. Country clusters plot. Own elaboration



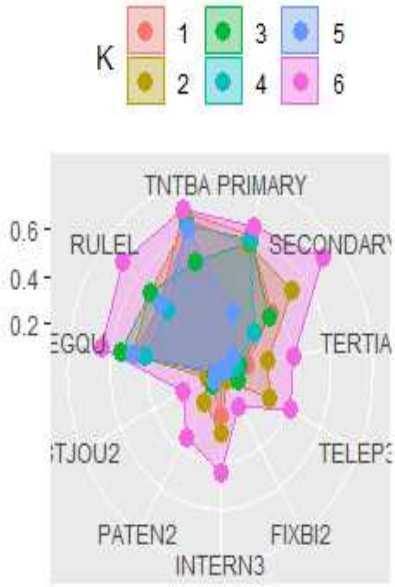
Source: own elaboration

Figure 5a. K-4 radar plot of African KE.

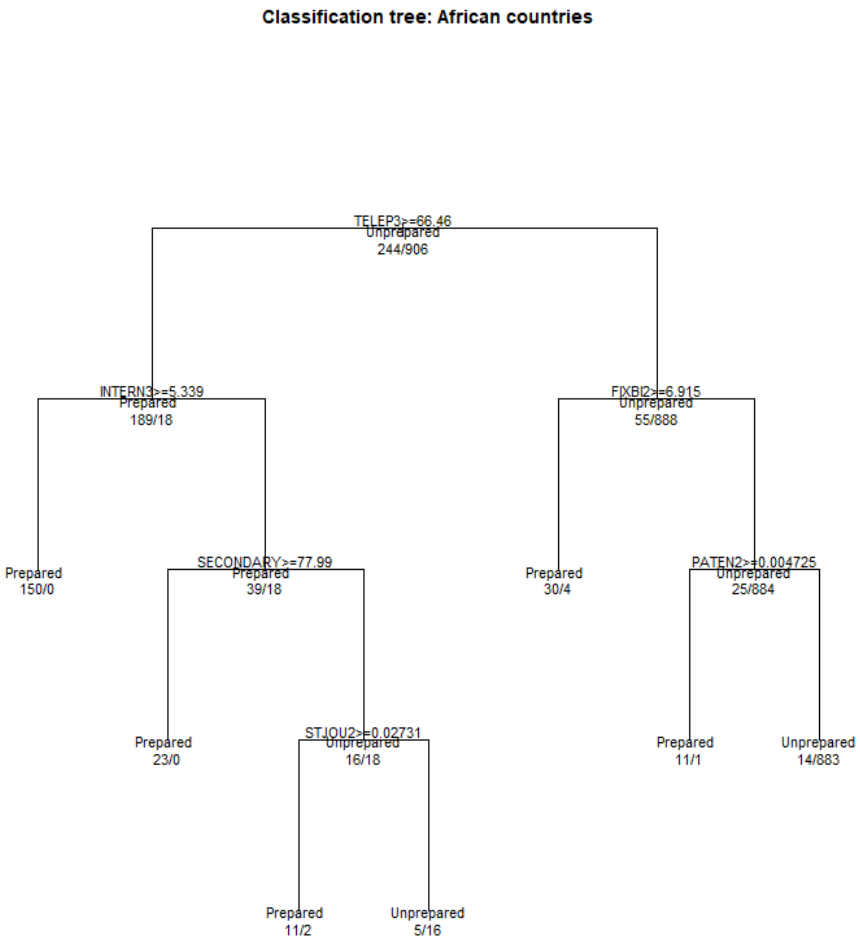


Source: own elaboration

Figure 5b. K-6 radar plot of African KE.



Source: own elaboration



APPENDIX tables and other illustrations

Table 1A *Variable Definitions, and Data Sources*		
Variable	***Definition***	***Source***
Regulatory quality (REGQU)	This indicator measures the incidence of market-unfriendly policies such as price controls or inadequate bank supervision, as well as perceptions of the burdens imposed by excessive regulation in areas such as foreign trade and business development	Worldwide Governance Indicators (WGI). The World Bank. Available at https://info.worldbank.org/governance/wgi/#home
Rule of law (RULEL)	This indicator includes several indicators that measure the extent to which agents have confidence in and abide by the rules of the society	Worldwide Governance Indicators (WGI). The World Bank. Available at https://info.worldbank.org/governance/wgi/#home
Tariff and non-tariff barriers (TNTBA)	This is a score assigned to each country based on the analysis of its tariff and non-tariff barriers to trade, such as import bans and quotas as well as strict labeling and licensing requirements	The Heritage Foundation's Trade Freedom score. The Heritage Foundation. Available at https://www.heritage.org/index/trade-freedom
Patent applications (PATEN)	Patent grants by country of origin and patent office, per 1000 people	World Development Indicators. Data Bank. The World Bank. Available at https://data.worldbank.org/
Scientific and technical journal articles (STJOU)	Scientific and engineering articles published by country per 1000 people	World Development Indicators. Data Bank. The World Bank. . Available at https://data.worldbank.org/
Internet users (INTERN)	Internet users are individuals who have used the Internet (from any location) in the last 3 months	World Development Indicators. Data Bank. The World Bank. . Available at https://data.worldbank.org/
Fixed telephone subscriptions (TELEP)	The number of subscriptions per 1000 people. It includes Integrated services digital network channels and fixed wireless subscribers	World Development Indicators. Data Bank. The World Bank. Available at https://data.worldbank.org/
Fixed broadband internet subscribers (FIXBI)	Fixed broadband internet subscribers per 1000 people	World Development Indicators. Data Bank. The World Bank. Available at https://data.worldbank.org/

Primary enrolment (% gross) (PRIMARY)	Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the primary level of education	World Development Indicators. Data Bank. The World Bank. Available at https://data.worldbank.org/
Secondary enrolment (% gross) (SECONDARY)	The ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the secondary level of education	World Development Indicators. Data Bank. The World Bank. Available at https://data.worldbank.org/
Tertiary enrolment (% gross) (TERTIARY)	The ratio of total enrolment, regardless of age, to the population of the age group that officially corresponds to the tertiary level of education	World Development Indicators. Data Bank. The World Bank. Available at https://data.worldbank.org/

Source: own elaboration

Table 2A. Summary Statistics				
Variable	***Mean***	***Standard deviation***	***Min***	***Max***
Regulatory quality	-0.643	0.58	-2.30	1.13
Rule of law	-0.649	0.595	-2.13	1.07
Tariff and non-tariff barriers	58.5	14.1	0	90
Patent applications	0.002	0.003	0.000013	0.00081
Scientific and technical journal articles	0.024	0.043	0	0.0052
Internet users	7.50	11.75	0	0.21
Fixed telephone subscriptions	34.73	57.41	0	326.53
Fixed broadband internet subscribers	5.226	18.064	0	194.52
Primary enrolment (% gross)	94.78	22.88	27.8	152.2
Secondary enrolment (% gross)	42.79	23.12	5.21	114.4
Tertiary enrolment (% gross)	8.857	8.580	0.22	60.51

Table 3A. Data preprocessing and missing value imputation strategies

Data preprocessing

So, before we run the clustering process, we need to select a missing value estimation method. In this respect our approach differs from the mainstream approaches in economics journals where the missing values are often replaced by the mean value or just simply ignored via listwise deletion. While the latter is reasonable, if the instances with missing values differ systematically from the observed instances, this could bias the complete-case analysis. There are several imputation methods. For a review of several imputation methods, see Little and Rubin (2002). In this paper missing values are imputed using the K–nearest neighbors imputation (henceforth, kNN). The kNN method relies on metric measures and the main idea is to find the K-closest neighbors to the observations with missing data and imputing them based on the non–missing values in the neighbors. The missing value is then replaced by a weighted mean of the k nearest neighbors, where the weights are proportional to the inverse of the Euclidean distances (the closer an instance is to the one that has a missing value, the more weight it has in computing the average to impute the missing value). In that sense, it seems to be more appropriate to replace missing values with plausible values from the observed dataset via imputation procedures. Indeed, Acuña and Rodríguez (2004) find that kNN imputation method becomes more robust when the number of missing values increases.

In this application, we used the R package ***CARET*** (Kuhn, 2008) to carry out the imputation of missing values. One of the main criticisms of this technique is the critical choice of *k*, the number of neighbors. We tried with several *k*, and we finally decided to use $k = 6$ based on the accuracy of the method (for a critical review on imputation methods (see Beretta and Santaniello, 2016).

Missing value imputation

As for the imputation strategy, and for some of the variables, it is possible to assume that missing values from the dataset have a certain value. In our dataset, for instance, the variable fixed broadband internet subscribers (FIXBI) can be assumed to be zero for some missing values. In the initial years of this study, broadband internet penetration was almost non-existent throughout the world. For the first five years of the period under study no country reported FIXBI data. Thus, for the first years there was no other temporally close value of FIXBI to be used for imputation with k- nearest neighbors. Thus, it does not seem a wise idea to calculate the mean of broadband penetration of several countries in 2005 to impute a value of 1996. Therefore, and under the hypothesis that typically when the countries do not report this data is because they still do not have broadband (once a country starts reporting FIXBI, it is missing only in 2.1% of the cases) to impute this parameter we proceeded as follows. A value of zero was set for all countries for all years starting in 1995 until the first year it reports a value for FIXBI. We verified that the first reported value is extremely low (almost always below 0.1), which is consistent

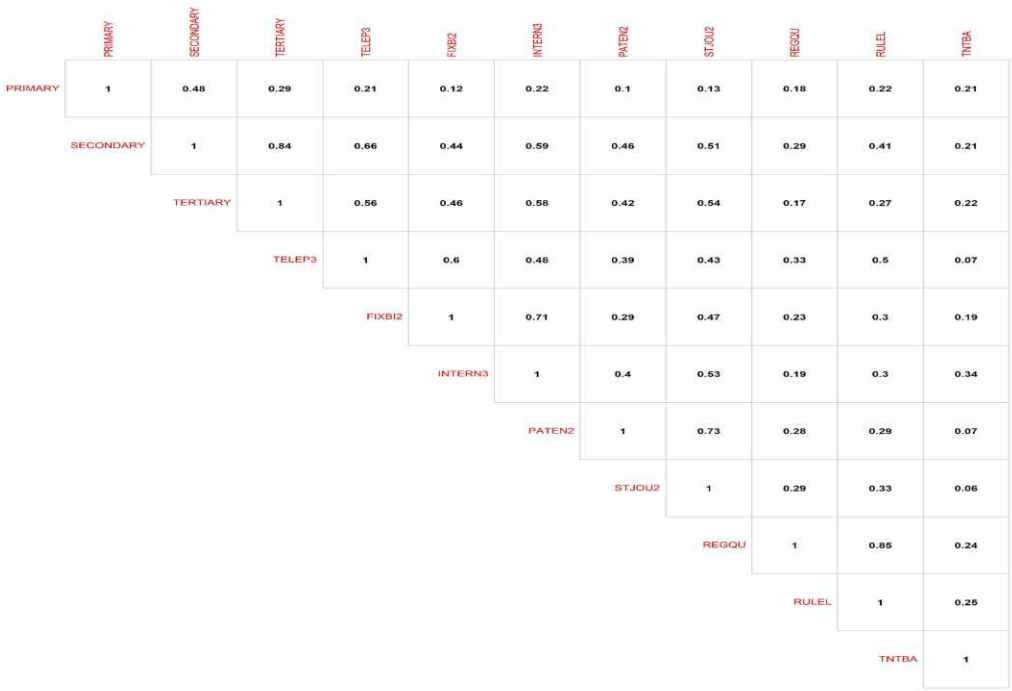
with the previous values being 0 or close to 0. The remaining 2.1% of FIXBI missing values were imputed through k nearest neighbors, in a similar way as the missing values for the rest of the features in our dataset.

Table 4A. Composition of four clusters

- ***Cluster 1: Very prepared:*** Algeria 2008; Algeria 2014-2017; Botswana 03-08; Botswana 10, 12-14, 16-17; Cabo Verde 04-13, Egypt 07-10, Egypt 15-17; Libya 99-03; Mauritius 20062017; Morocco 2013-14, Morocco 2016-2017, Namibia 2013, 2015-2016, Sao Tome and Principe 2016-2017, Seychelles 2003, 2006-2017, and South Africa 1998-2016.
- ***Cluster 2: Prepared***: Algeria 2005-2007, Algeria 2009-2013, Botswana 2009, Botswana 2011, 2015, Cabo Verde 2014-2017, Cameroon 2012-2017, CAF-2017, Djibouti 2008-2017, Egypt 1996, 1998-2006, 20112014 South Africa 1995, Tunisia 1999, Tunisia 2003-2017, Equatorial Guinea 2017, Eritrea 2013-2017, Lesotho 1995, Liberia 2017, Libya 2004-2017, Mali 2017, Mauritius 1995, 1997, 1999-2005, Morocco 1998, Morocco 2000, Morocco 2005-2012, Morocco 2015, Namibia 2009-2012, 2014, 2017, Congo Republic 2017, Senegal 2010-2013, Seychelles 1995-1997, 1999-2002, 2004-2005, South Africa 1996-2017, Sudan 2008-2011, Sudan 2013, 2015-2016, Tanzania 2016-2017, Gambia 2017, Tunisia 1996-1998, Tunisia 2000-2001, Zimbabwe 1995, 2013-2017.
- ***Cluster 3: Unprepared***: Algeria 1995-2004, Angola 1995-2000, 2003, 2006-2017; Benin 1996– 1998, Benin 2001–2014, 2016-2017, Botswana 1995–2002, Burkina Faso 2010-2017, Burundi 2005–2017, Cabo Verde 1995–2003, Cameroon 1995–2011, CAF-1995, 1997–1998, 2000, 2005, 2007-2016; Chad 2008–2017, Comoros 1995–2017, Cote D'Ivoire 1995–2000, 2004–2017, Congo. Dem. Rep. 1995-1998, 2000–2001, 2003–2017, Egypt 1995, 1997, Equatorial Guinea 1995–2016, Eritrea 2002–2012, 2014–2016, Ethiopia 2006–2017, Gabon 1995–2017, Ghana 1995–2017, Guinea 1997, 2002-2003, 2005– 2017, Kenya 1995, 1997–2017, Lesotho 1996–2017, Liberia 1995–2016, Libya 1995–1998, Madagascar 1995–1996, 1998–2017, Malawi 1995– 2017, Mali 2006–2016, Mauritania 1997–2002, 2007–2017, Mauritius 1996, Morocco 1995–1997, 1999, 2001–2003, Mozambique 1997–1998, 2003–2017, Namibia 1995–2008, Niger 2015–2017, Nigeria 1995–2017, Congo Rep. 1995–2016, Rwanda 1995–2017, Sao Tome and Principe 1995–2015, Senegal 1997, 2003–2009, 2014–2017, Sierra Leone 1995– 2000, 2002–2017, Sudan 1995–2007, 2014–2017, Tanzania 2000, 2002–2015, Gambia 1996–2016, Togo 1995–2014, Togo 2016–2017, Tunisia 1995, Uganda 1997–2017, Zambia 1995–1997, 2003–2009, 2011–2017, Zimbabwe 1996–2013.
- ***Cluster 4: Very unprepared*:** Zambia 00-01, Uganda 95-96, Gambia 95, Tanzania 95-99, 01; Sierra Leone 00-01; Senegal 98-02, Senegal 95-96, Niger 95-14; Mozambique

99-02, Mozambique 95-96; Tanzania 03-06, Tanzania 95-96; Mali 95-05; Madagascar 97; Kenya 96; Guinea 04; Guinea 98-01, Guinea 95-96; Ethiopia 95-06; Eritrea 95-01; Djibouti 95-07, Cote D'Ivoire 01-03; Chad 95-07; Central African Republic 96 , Central African Republic 99, 01-04, 07; Burundi 96-04; Burkina Faso 95-09; Benin 99-00, Benin 95- 97 and Angola 01-05.

Figure 1A: Correlation matrix



	PRIMARY	SECONDARY	TERTIARY	TELEP3	FIXBI2	INTERN3	PATEN2	STJOU2	REGQU	RULEL	TNTBA
PRIMARY	1	0.48	0.29	0.21	0.12	0.22	0.1	0.13	0.18	0.22	0.21
SECONDARY		1	0.84	0.66	0.44	0.59	0.46	0.51	0.29	0.41	0.21
TERTIARY			1	0.56	0.46	0.58	0.42	0.54	0.17	0.27	0.22
TELEP3				1	0.6	0.48	0.39	0.43	0.33	0.5	0.07
FIXBI2					1	0.71	0.29	0.47	0.23	0.3	0.19
INTERN3						1	0.4	0.53	0.19	0.3	0.34
PATEN2							1	0.73	0.28	0.29	0.07
STJOU2								1	0.29	0.33	0.06
REGQU									1	0.85	0.24
RULEL										1	0.25
TNTBA											1

Source: own elaboration

Code availability statement

The R source code and the original dataset are available at: https://github.com/antonio1970/Clustering-Algorithms/tree/master/code