Comparative clustering and visualization of socioeconomic and health indicators: A case of 47 counties in Kenya

EVANS KIPTOO KORIR (evanskorir6@gmail.com)
University of Szeged

Research Article

Keywords: cluster analysis, socioeconomic and health indicators, counties, Dimensionality reduction, Principal component analysis, Hierarchical and K-means clustering.

Posted Date: December 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3771097/v1

Additional Declarations: No competing interests reported.

Comparative clustering and visualization of socioeconomic and health indicators: A case of 47 counties in Kenya.

Evans Kiptoo Korir, (evanskorir6@gmail.com)

[a] *University of Szeged, Hungary.*

Abstract

Kenya as a country is still faced with uneven regional development seen across various economic sectors. The regions experienced inequalities in economic diversity, economic development infrastructure, human development levels, social structure and living conditions, as well as political representation and participation in decision-making.

The aim of this article is to examine the dynamics of marginalization and regional inequalities in Kenya by grouping counties based on socioeconomic and health indicators. We also analyzed aggregated data from the World Bank country website, online publications, and county government websites. The data includes 24 variables and metrics describing counties; Fertility, mortality, social, education, population and development and access to social services. Principal component analysis was used to visualize socioeconomic similarities between counties and identify the indicators of maximum variation. Agglomerative hierarchical and K-means clustering were then used to identify county groups with similar socioeconomic and health performance. The grouped counties were then visualized on a geographic map. Five clusters were identified that may be useful to county and state governments in future plans to promote inclusive and sustainable economic development.

Keywords: cluster analysis, socioeconomic and health indicators, counties, Dimensionality reduction, Principal component analysis, Hierarchical and K-means clustering.

1. Introduction

Until the promulgation of the Kenyan constitution in 2010, the country was divided into 8 provinces [1]. The provincial system was replaced by a decentralized government under the Ministry of Decentralization, which established 47 counties as units. The size, boundaries, and geographic location of the counties are based on the 47 statutory regions under the 2010 Constitution [2, 3]. The list of 47 counties is shown along the horizontal axis in Table. 3. With the creation of the counties, several key functions of the national government, such as healthcare, education, agriculture, and transportation, among others, were transferred to the counties [1, 2, 3]. Each county is headed by a governor, elected by the residents of the county and supported by the executive committee and members of the county assemblies.

Over the last decade, counties have implemented significant economic reforms that have contributed to sustained economic growth, improved health care, and higher school enrollment rates. However, development has been hampered by key challenges such as corruption, poverty, inequality, transparency and accountability, national government interference, climate change and low investment attractiveness [4, 5]. This has led to socioeconomic and health disparities between counties and triggered inter-county migration. One of the biggest differences is the county's population. The country's population was 47.56 million in 2019, with a population growth rate of 2.3 percent [3]. In 2009, the country's population was counted at 38.6 million. Since the country's independence in 1963, the population has increased fivefold. The population is dominated by young people aged 15 and younger as they make up 39 percent of the population. This proportion has decreased from 43 percent in 2009 [6]. Only 57 percent of the country's population is employed (aged 15 to 64), with 29 percent being youth. Older people (60 years and older) make up 6 percent of the total population [3]. Nairobi County has the highest population percentage (9.24 percent), while Isiolo has the lowest at just 0.56 percent [6, 7, 8, 9, 10].

According to the Kenya Economic Report 2020, only 15.0 percent of the 47 counties have significant manufacturing activities, with most counties heavily dependent on agriculture. Agriculture is the most important economic sector in most regions, followed by the service sector [11]. Counties with relatively well-established manufacturing and agricultural sectors have larger populations [2]. Differences between counties also exist in poverty rates, with Nairobi County reporting the lowest rates at 16.7 percent, while

Turkana County recorded the highest rate at 79 percent. The report shows that 22 counties are still below the national level, which is 36.1 percent, while the poverty rate in 16 counties is below 8.6 percent, the national core poverty rate [3, 10, 12].

The government has increased its resource allocation efforts to combat poverty and inequality in counties through the budget [13, 14]. Turkana and Mandera, the least developed counties, receive a relatively larger share of revenue allocation. The allocation is based on the poverty factor in the Commission's Revenue Allocation Formula (CRA), which accounts for 18 percent of the revenue allocation. However, there is no clear understanding of the other indicators that need to be included in this CRA formula to better understand the counties that require special attention [13, 14].

Recent research has shown a direct connection between socioeconomic and health indicators, and cluster analyzes have been shown to reveal patterns of hidden regional economic development [15, 16, 17, 18]. By comparing and characterizing counties, we can test the theory that socioeconomic and health heterogeneity may account for the spatial diversity of regions. The introduction of cluster strategies as part of the overall regional development strategy offers the opportunity to increase the county's competitiveness.

In this study, we used dimensionality reduction techniques such as principal component analysis and an agglomerative hierarchical algorithm to represent the clusters through a dendrogram with different levels of granularity based on the 24 socioeconomic and health indicators that could explain the variability of the 47 counties. In order to provide a more accurate clustering result, data scaling is performed and K-means clustering is implemented to validate the obtained clusters based on the hierarchical clustering algorithm.

2. Theoretical framework

Several studies have been conducted to group regions based on socioeconomic and health indicators. [16] analyzed aggregated survey data and demographic data to classify 29 SSA countries into three clusters based on 48 socioeconomic and HIV-related indicators. The study used PCA to identify the variables that explain the maximum variation. The study found that designing interventions that incorporate social and behavioral factors can help reduce HIV transmission in SSA countries.

The socioeconomic and demographic factors were also used to group 146 WHO member countries to identify disease-specific deaths and suggest inter-

ventions to reduce mortality rates in the regions [15]. The authors used PCA to identify the underlying data patterns before applying the hierarchical clustering algorithm using the Ward linkage method to group the countries that have similar causes of death. The study found that income, spending, education and causes of death are related in a country. The cluster differences proved to be statistically significant.

Another study [19] ranks 180 countries by COVID-19 cases and deaths to examine countries' preparedness and control for the pandemic. Based on the countries' daily COVID-19 cases, the study used a hierarchical clustering technique to group the countries into five clusters. The clusters showed that more developed countries belonged to the same cluster, while the least developed countries, such as African countries, were placed in a separate cluster. The results suggest that countries' economies are closely linked to the health system and therefore deaths.

For good clustering and a better shaped dendrogram, multiple linkage methods are used; average, complete, station and single. [20] compared these methods when examining the causes of poverty in the North Sulawesi region. The author used agglomerative hierarchical clustering with the linkage methods and compared the results with the RMSSTD value. The study found that the Ward linkage method resulted in the smallest RMSSTD values compared to other methods and therefore can be a good criterion for multiple analysis.

3. Materials and Methods

3.1. Data

To classify counties with similar characteristics in Kenya into a small number of clusters, socioeconomic and health data for all counties were extracted from the reports of the National Council for Population and Development, Kenya Economic Growth 2020, Ministry of Decentralization, Individual County Website, Kenya, collected census report 2019, Knoema data website and Kenya Property Developers Association (KPDA) website. The selected data capture residential structure, demographic, health, social and economic information and characterize various dimensions of the districts. A total of 24 raw and calculated indicators were selected for cluster analysis, as listed in Table. 1.

We transformed the raw data to standardize some of its characteristics because some of the variables such as county GDP, population and density, household size, infant mortality rates, and crime index vary widely and the

Socioeconomic and Health Indicators		
Fertility measures	**Mortality measures**	**Social measures**
contraceptive prevalence	Infant Mortality rates	Employment rate
Fertility rates	Under-five mortality rates	Crime index
Birth rate	Death rates	Poverty rates
Household size	Healthcare facility delivery	Unemployment rate
Education level measures	**Population and Development**	**Access to social services**
HIV prevalence rates	Population size	Urbanization rates
Education level	Population density	Electricity access
Literacy rates	GDP	land size
Child marriage rates	growth rates	Healthcare facility Density

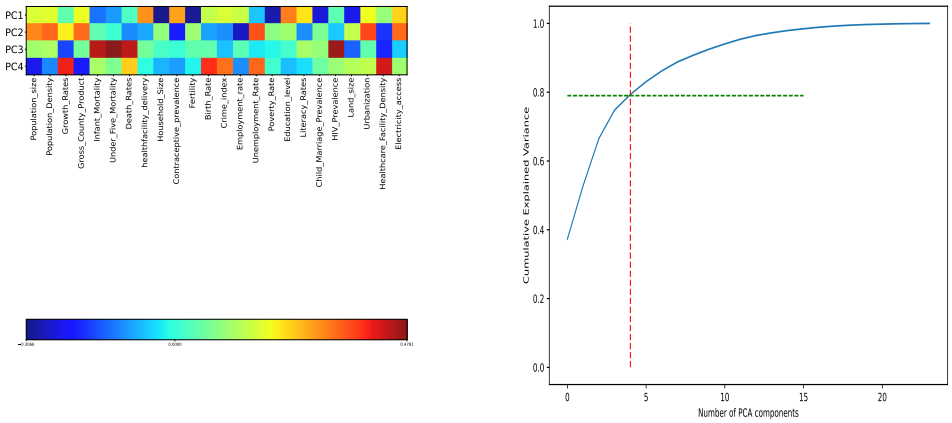
Table 1: The 24 socioeconomic and health indicators categorized into 6 measures that forms the basis for clustering counties.

model learns from the most important characteristics. The variables were scaled using the standard scalar method in scikit-learn [17].

3.2. Principal Component Analysis

Modern applications of statistical theory involve massive amounts of data with a larger number of features compared to data points. To avoid dimension curse that affects learning models and training performance due to high-dimensional data, feature selection and dimensionality reduction techniques such as principal component analysis (PCA) are often used [21, 22]. The PCA technique is used for many reasons; Identify relationships between variables, detect outliers, identify patterns, and reduce data dimensionality [23]. In dimensionality reduction, PCA projects the data ($X = x_1, x_2, ..., x_n$) from a high-dimensional space $\mathbb{R}^m$ onto a low-dimensional space $\mathbb{R}^d$ towards maximum variation, where n represents the total number of observations [24, 25, 23, 20, 16, 26, 17, 22]. The PCA space has d principal components that are orthonormal and uncorrelated and represent the direction of maximum variance. The principal components can be easily determined using the covariance method or the singular value decomposition (SVD)[21, 23, 22].

Using the SVD approach, PCA was applied to the 24 socioeconomic and health datasets and the data were projected and transformed from $\mathbb{R}^{24}$ into four feature vectors, $\mathbb{R}^4$ as shown in Fig.1a. The resulting principal compo-



(a) (b)

Figure 1: Figure (a) corresponds to the heat map of factor loading for the first four principal components. The colors represent the correlation coefficient between the original feature and the principal component. Their values are provided by the horizontal bar on the lower side. The cumulative variance plot (b) shows the total percentage of variance explained versus the number of principal components.

nents and the amount of variance explained by each component are shown in Fig. 1b.

We determined the number of principal components that optimizes our algorithm by looking at the graph 1b. Thus, the optimal number of components is 4 and this explains 74 percent of the variance in the data set 1b. The first principal component explains about 36 percent of the variation, the second component about 16 percent, the third about 13 percent, and the fourth component only about 8 percent. Remarkably, the latter components explain only a small part of the total variability [23].

3.3. Clustering

Grouping large data sets has gained importance in various fields such as health, medicine, social, and spatial [20]. Grouping such data requires cluster analysis such as clustering, an unsupervised algorithm for identifying clusters within a data set such that the within-cluster similarity is high and low in-between clusters. In particular, there are various methods for clustering,

such as: Feature selection, distance-based, partition-based methods, density-based, probabilistic techniques, grid-based methods, among others [17, 19, 21, 27, 28].

In this study, distance-based algorithms are used due to its wide range of applications, simplicity and ease of implementation compared to other clustering methods [21]. Firstly, agglomerative clustering is considered, in which in the first step each observation is viewed as a separate cluster. The closest sets of clusters are merged at each level and then the dissimilarity matrix is updated accordingly [17, 19, 20, 21, 27, 28]. This process of agglomerative merger continues until the final maximum cluster is reached. Clusters are aggregated based on decreasing level of similarity until we arrive at a single cluster that includes all data points. This would represent the culmination of our dendrogram and mark the completion of the merging process [27]. In this study, Ward's criterion was used to bring the individual clusters together using the distance measure because it is based on minimizing the inner sum of square error (SSE). For any two clusters, C_1 and C_2, this link is calculated by measuring the increase in the clustering SSE value caused by merging them, $C_1 \cup C_2$. The Ward's criterion is defined as follows

$$W(C_{1\cup 2}, c_{1\cup 2}) - W(C, c) = \frac{N_1 N_2}{N_1 + N_2} \sum_{v=1}^{M} (C_{1v} - C_{2v})^2$$

$$= \frac{N_1 N_2}{N_1 + N_2} d(C_1 - C_2)$$

where d is the chosen distance measure. In this study, the Euclidean distance is taken into account. After the dendrogram is created, the number of clusters can be easily determined. As a rule of thumb, we consider the vertical distance with the largest gap between the two clusters [27, 28].

Secondly, k-Means is considered to validate the results obtained using hierarchical clustering. In this algorithm, the partitioning representatives correspond to the mean of each cluster. The euclidean distance is also used in order to compute distances. Other measures which can be used are Manhattan distance and Cosine similarity.

Given a dataset $D = \{x_1, x_2, ..., x_N\}$ which consists of N points, and denote the clustering obtained after applying K-means clustering by $C = \{C_1, C_2, ..., C_k..., C_K\}$. The objective is to find a clustering that minimizes the SSE score [27].

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|^2$$

$$c_k = \frac{\sum_{x_i \in C_k} x_i}{\mid c_k \mid}$$

where c_k is the centroid of cluster C_k.

3.4. Research design

Several research steps were carried out in this study. The research design of the study is summarized in the figure 2

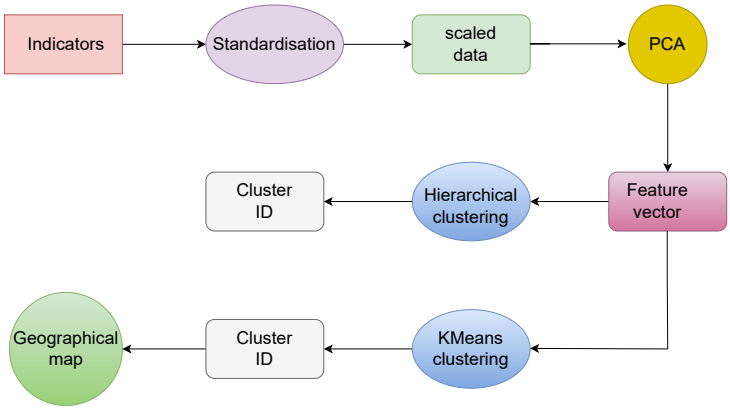


Figure 2: Flowchart summarizing the research design. Rectangles show the output of a module, represented by ellipsoids that contain function calls that execute a specific part of the method. As can be seen, the socioeconomic and health indicators are the input of the study. We apply standardization to the indicator data to obtain the scaled data. In the next step, we apply PCA to the scaled data. The output feature vectors of PCA are used in the hierarchical and K-means clustering algorithm to obtain the clusters. The geographical map visualizes the clusters from the K-means in a geographical context.

All statistical analysis, PCA, hierarchical and K-means clustering are performed using the Python object-oriented programming framework using built-in libraries and open source packages. Cluster map visualization runs Folium in a notebook environment. Scikit Learn is used to develop and run PCA and clustering algorithms. The code is publicly available on GitHub.

4. Results

We analyzed aggregated data from National Council for Population and Development reports, Kenya Economic Growth 2020, Ministry of Decentralization, Individual County Website, Kenya, 2019 Collected Census Report and other websites to sort counties with similar characteristics in Kenya into a small number of clusters.

4.1. Visualization of socioeconomic and health characteristics

Socioeconomic and health variables were grouped into six broader measures, as shown in the Table.1. The values of these variables are then plotted using a heatmap as shown in Fig.3 for easy visualization. The variables are arranged in the order in Table.1 and the values of each variable in each measure are presented in small triangles for each county. Looking at access to social services in Fig3, urbanization is represented by (top), access to electricity (right), land size (bottom), and healthcare facility density (left). The size of the values increases with the intensity of the color. For example, Nairobi and Mombasa have higher rates of urbanization and access to electricity (green color). Narok and Nyandarua have low urbanization rates, while Mandera and Tana River have low electricity access (lightest shade). Likewise, Lamu has the highest density of health facilities (green color), while Marsabit and Turkana have the largest land area (green color, lower triangle)

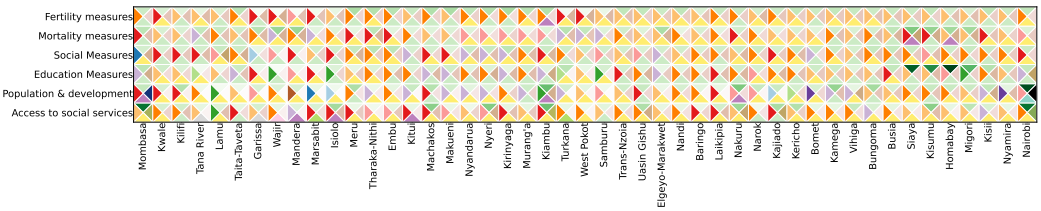


Figure 3: Stacked socioeconomic and health variables categorized into six measures presented on the vertical axis for all 47 counties. The variables are arranged in the order in Table.1. For example, access to social services, urbanization (top), access to electricity (right), land size (bottom) and health facility density (left). Their values are indicated by the strength of the color, with smaller values having a lighter color. The dark color indicates counties with higher readings.

Population and development measures in Fig.3 were divided into population size (top), population density (right), GDP (bottom), and growth

rates (left) in Table.1. The two largest cities; Nairobi and Mombasa have the highest population density, while Nairobi and Kiambu have the largest population size, with Nairobi recording a high GDP. Marginalized counties such as Lamu, Marsabit and Isiolo have high population growth rates.

We grouped HIV prevalence rates, educational attainment, literacy rates, and child marriage rates in Table.1 into measures of educational attainment (3). Based on the figure, Siaya, Kisumu, Homabay and Migori have the highest HIV prevalence rates. Marriage rates are particularly high in Wajir, Isiolo and Samburu counties.

In this document, social measures combine the employment rate, the crime index, the poverty rate and the unemployment rate as indicated in Table.1. Unemployment rates are high in Kilifi, Tana River, Isiolo, Machakos, Makueni, Kajiado and Nairobi counties. Lamu, Mombasa and Taita Taveta counties have the highest crime rates.

Infant mortality rates, under-five mortality rates, death rates and health facility provision are the measures of mortality in this article. Partly due to high HIV prevalence rates, Siaya, Kisumu, Homabay and Migori have reported the highest mortality rates. This also applies to other underdeveloped counties such as Garissa, Wajir, Mandera and Marsabit. The provision of health facilities is better (red color) in Mombasa, Meru, Tharaka-Nithi, Embu, Siaya, Kisumu and Kisii counties (3).

Finally, contraceptive prevalence, fertility rate, birth rate and household size are the variables that determine fertility rates (1). Counties along the Lake Victoria region (Siaya, Kisumu, Homabay and Migori) have the highest birth rates. Kwale, Garissa, Wajir, Marsabit, Turkana and West Pokot counties have higher household size (red color).

4.2. Clustering of counties and analysis of associated socioeconomic and health indicators

The hierarchical grouping of the 47 countries resulted in a dendrogram 4. The advantage of this clustering technique is that one can manually cut the hierarchy at any level and retrieve the clusters accordingly without having to run the algorithm again. Using the rule of thumb, we can separate the clusters by defining a threshold. The choice of threshold is completely user dependent. By dividing the dendrogram at a cluster distance of 10, five separate clusters were formed, as shown in figure 4 and in Table 2.

The counties in the first cluster (in black) have less productive economic activities and therefore contribute little to the country's gross GDP. This

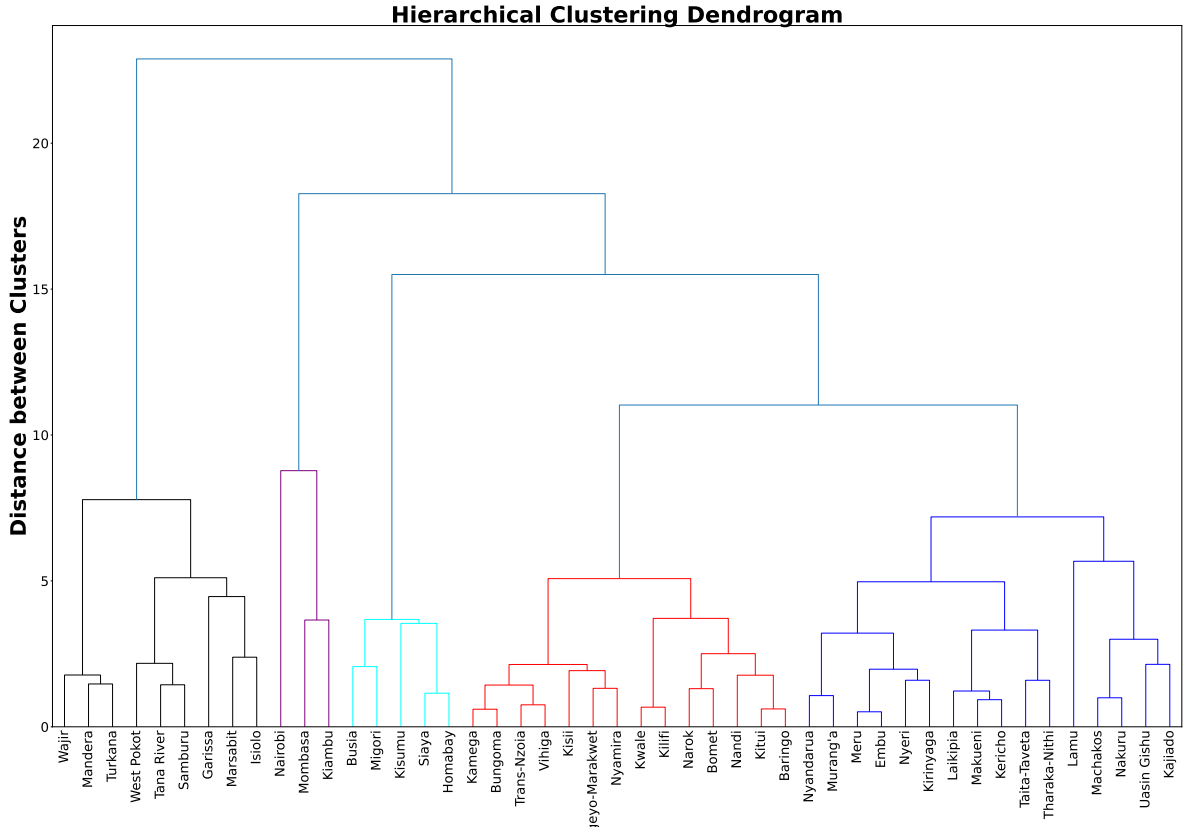


Figure 4: Dendrogram related to the grouping of counties based on the socioeconomic and health indicators. The vertical axis indicates the distance between the two clusters being joined. The agglomerative nature of the algorithm can be seen by following the tree from the bottom up. Counties with the same hue belong to the same cluster.

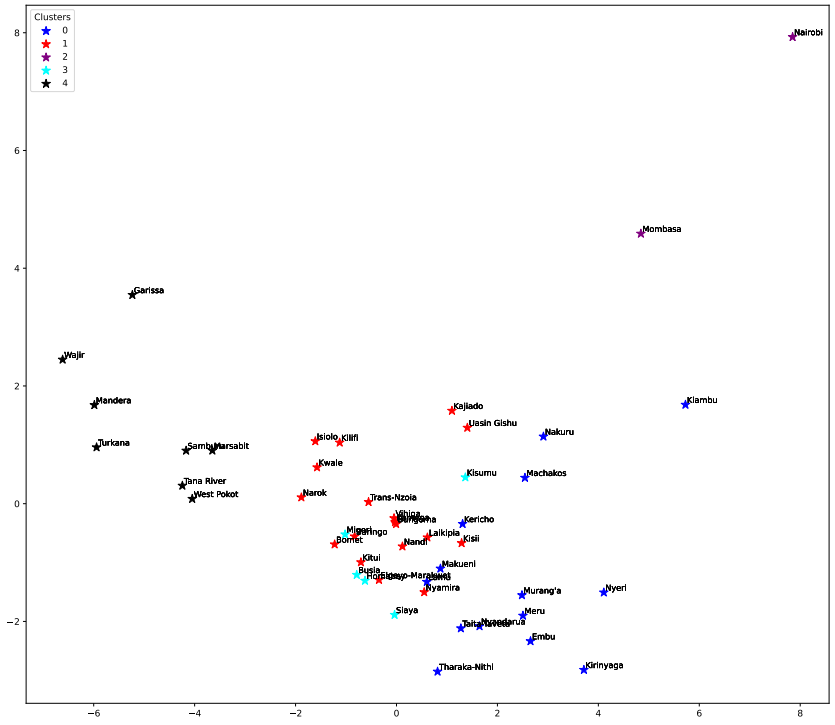


Figure 5: Grouping of counties based on the socioeconomic and health indicators using K-means algorithm. The output graph is clearly showing the five different clusters with different colors.

cluster included counties in the northern part of Kenya; an arid area with average temperatures of up to 35 degrees Celsius in some areas that do not support significant economic activities such as agriculture. In these counties the people are mainly nomadic pastoralists. For the k-means algorithm, Isiolo was grouped into cluster 4 (see Fig. 5 and Table 3).

Cluster 2, in purple, includes three large developed counties in Kenya; Nairobi, Mombasa and Kiambu. Nairobi is the capital of the country. It has the highest population, the highest contribution to GDP and is more industrialized 3. Mombasa is the second largest capital after Nairobi and is

particularly suitable for international trade. Kiambu borders Nairobi and is growing rapidly due to its large population and active trade. These counties are more developed compared to other counties and, notably, Nairobi appears to be an outlier. The K-means method omitted Kiambu from this cluster, as shown in Fig.5 and Table.3.

The third cluster counties, colored cyan, include counties from Nyanza Province. Fishing is the main economic activity in these counties. Notably, these counties have high HIV prevalence and mortality rates (Fig. 3). For this cluster, the two clustering algorithms produced the same clusters.

The fourth cluster in red included counties from all regions except the northern part. These counties have average socioeconomic and health indicators that are no better than counties in the second and fifth clusters (Fig. 3). Most of the counties in this cluster are from the western region. The K-means method included Laikipia, Uasin Gishu and Kajiado counties in this cluster as shown in Fig.5 and Table.3.

Cluster 5, in blue, includes counties that are well distributed across the country. These counties are not well developed, but perform better economically than counties in clusters 1, 3 and 4 3. The counties have larger cities, are moderately developed, and their socio-economic indicators suggest average performance, since agriculture, fishing and trade are mainly carried out in these counties. Table 2 and Table 3 provide the final list of clusters using hierarchical and K-means clustering, respectively.

5. Discussion.

This paper presented the socioeconomic indicators that define the county cluster, as well as the necessary factors that a county should utilize to be economically successful. Child mortality, infant mortality and HIV prevalence rates are highly correlated as they are indicators of mortality rates and life expectancy. Counties with high mortality rates form a cluster. It is also obvious that in these counties there is little use of contraceptives, the number of early marriages is moderate and the poverty rate is low. The county's GDP is an important indicator of the county's growth. This depends on the county's population size and density, access to electricity for transformation industrial development, education and urbanization to provide residents with access to essential and affordable amenities. GDP growth is influenced by high birth and mortality rates, which is evident in this study. Counties with

CLUSTER 1	**CLUSTER 2**	**CLUSTER 3**	**CLUSTER 4**	**CLUSTER 5**
Wajir, Mandera, Turkana, West Pokot, Tana River, Samburu, Garissa, Marsabit, Isiolo.	Nairobi, Mombasa, Kiambu.	Kisumu, Siaya, Homabay, Busia, Migori.	Vihiga, Kakamega, Bungoma, Trans-Nzoia, Elgeiyo-marakwet, Kwale, Kilifi, Narok, Bomet, Nandi, Kitui, Baringo.	Meru, Embu, Nyeri, Kirinyaga, Laikipia, Kericho, Taita-Taveta, Makueni, Nyandarua, Muranga, Tharaka-Nithi, Kisii, Nyamira, Lamu, Machakos, Nakuru, Uasin Gishu, Kajiado.

Table 2: Clusters Produced based on Socioeconomic and health indicators for the 47 counties in Kenya using hierarchical clustering method.

CLUSTER 1	**CLUSTER 2**	**CLUSTER 3**	**CLUSTER 4**	**CLUSTER 5**
Wajir, Mandera, Turkana, West Pokot, Tana River, Samburu, Garissa, Marsabit	Nairobi, Mombasa.	Kisumu, Siaya, Homabay, Busia, Migori.	Isiolo, Vihiga, Kakamega, Bungoma, Trans-Nzoia, Elgeiyo-marakwet, Kwale, Kilifi, Narok, Bomet, Nandi, Kitui, Baringo.	Kiambu, Meru, Embu, Nyeri, Kirinyaga, Laikipia, Kericho, Taita-Taveta, Makueni, Nyandarua, Muranga, Tharaka-Nithi, Kisii, Nyamira, Lamu, Machakos, Nakuru, Uasin Gishu, Kajiado.

Table 3: Clusters Produced based on Socioeconomic and health indicators for the 47 counties in Kenya using K-means algorithm.

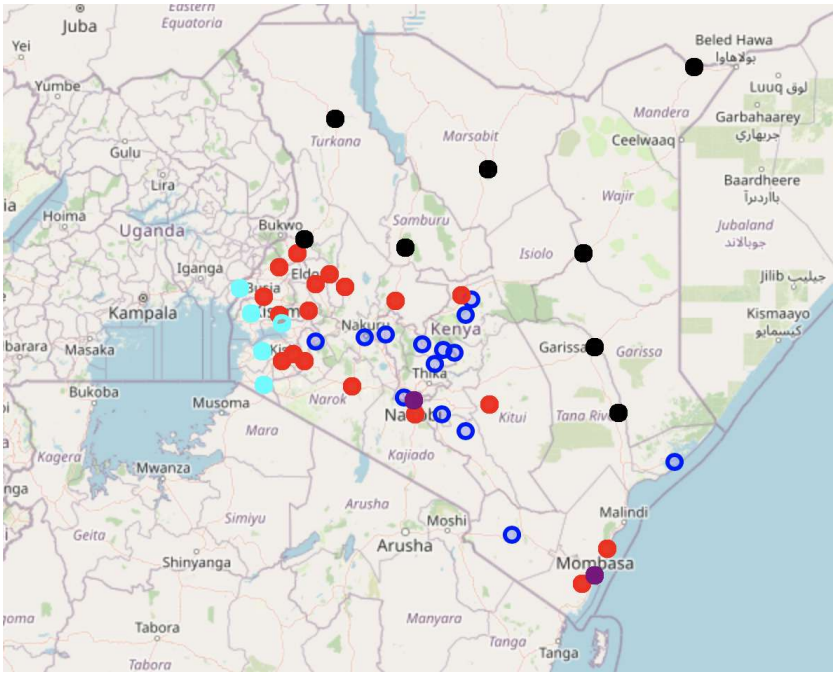


Figure 6: Visualizing the county's socioeconomic and health indicators and their geographic context using Folium. The 47 Kenyan counties were divided into five clusters based on the hierarchical and K-Means algorithm. This map is generated using the clusters from the K-Means method. The link to the generated geographic map: `http://localhost:8888/view/map.html`.

low GDP have high mortality rates, fertility rates, household sizes, and early child marriages.

Cluster analysis 1 – underdeveloped counties. This cluster includes counties as shown in Tables. 3 and 2. These are counties in the northern part of Kenya (Fig. 6). Most importantly, these counties are located in arid and semi-arid areas whose main economy is livestock farming. What is noteworthy is that these regions have a low GDP level of less than 1 percent. Kenya as a country relies on agriculture as its main industry and these regions cannot support most agricultural activities, which is why the area is sparsely populated. Therefore, the counties report lower productive economic activity as food poverty is widespread in this cluster and the counties average more than 60 percent food poverty. Therefore, these counties also have higher poverty rates compared to other counties. For example, the poverty rate in Mandera, Wajir, and in Garissa is high. Since there is a strong positive correlation between household size and poverty rate, the counties in this cluster have larger household sizes as indicated in Fig. 3.

Cluster analysis 2 – counties with a high level of development. This cluster consists of only three counties: Nairobi, Mombasa and Kiambu (Fig. 4). The cluster includes counties with large cities and towns and can be described as developed. Mombasa is the oldest and largest city after Nairobi (the capital) and it benefits hugely from its location as it is located on the Eastern coastline of Kenya bordering the Indian Ocean, a popular beach destination (Fig. 6). On the other hand, Kiambu borders Nairobi County and has benefited enormously from the urbanization and growth of the city. Counties in this cluster are the most urbanized; Nairobi and Mombasa have an urbanization rate of 100 percent, with Kiambu in third place. The three counties are home to thousands of Kenyan businesses and are well connected to the electricity grid. It is also evident that the counties have well-developed industries and a robust manufacturing sector, which has created productive employment opportunities at the national and county levels and increased the counties' GDP. Nairobi contributes a quarter of the total GDP, with Kiambu and Mombasa contributing the second and third largest shares.

Cluster Analysis 3 – High Mortality Counties. This cluster entirely comprises counties of Nyanza Province in southwestern Kenya around Lake Victoria (Fig. 6). The place is predominantly inhabited by the Bantu speakers Luo and Kisii. Since these counties are located in the lake region, their main industry is fishing. The water from the lake is not used optimally as the region still suffers from food shortages due to low commercial irrigation and low

agricultural activities. Communities in these clusters have strong social and cultural traditions. HIV/AIDS rates are widespread in the region. Kenya as a country has a high HIV burden; In 2013, an estimated 1.6 million people were HIV positive [8]. Women are more susceptible to the disease. The epidemic geographically affects these counties as Homabay, Siaya, Kisumu and Migori have high HIV prevalence rates. The high HIV rates correlate strongly with mortality rates, child mortality rates, and under-five mortality rates. Homabay County has the highest mortality rates, followed by Siaya. All counties in this cluster have an infant mortality rate of more than 100 per 1,000 live births, with Siaya County having the highest rate at 142. It can also be seen that the mortality rate among children under five years of age is high, with the highest rate recorded in Siaya County 3.

Cluster analysis 4 – medium developed counties. This cluster includes counties that geographically belong to the Western, Coast and Rift Valley provinces (Fig. 6). The counties have no major cities and operate agriculture and trade. The counties have fewer cities and are looking for markets for their agricultural products in other counties in clusters 2 and 5. Their GDP is still limited by the lack of significant manufacturing industries in the regions. However, the counties perform significantly better economically in terms of socio-economic indicators than the counties in cluster 1. They perform equally better on health indicators than Cluster 3 3.

Cluster analysis 5 – Economically stable counties. This is the largest cluster as seen in Fig. 4. The counties come from all parts of Kenya except the northern region, which is dominated by the counties in Cluster 1 (see Fig. 6). These counties are in the process of urbanization, but rely on agriculture as their main economic sector alongside trade and commerce. They have significant cities and towns, but not enough to create a large and viable market for agricultural products. However, major cities have attracted larger populations and led to inter-county migration that has marginalized some regions of the country. These counties are undeniably still in the process of development to reach the levels of Mombasa, Kiambu and Nairobi. They still lack significant manufacturing industries that can improve the county's GDP. This further explains the limited regional change at the county level. A robust manufacturing, agricultural and service sector is expected to create productive county-level employment opportunities for residents. The cluster performs above average and better on socioeconomic indicators than the counties in Cluster 1 and on health indicators in Cluster 2. They have good school enrollment rates, better health care, good electricity connections, bet-

ter urbanization rates and manageable fertility, mortality, household size and child marriage prevalence rates 3. However, we still have peripheral regions in this cluster and this applies to the other clusters as well.

Our work has some further limitations. First, there is a data gap as the socio-economic indicators were not collected in the same year. Demographic information such as the county's population was based on the year 2019, education level (2018), poverty rate (2016), infant and child mortality, crime index and gender index (2009). Therefore, there is a possibility that the indicators are not representative of the current situation. Second, the hierarchical and K-means clustering techniques are sensitive to noise and outliers and therefore form slightly different clusters. Cluster 2 has developed counties that may impact the optimal grouping of the other remaining counties. Third, using few principal components to explain the raw features may lead to difficulties in interpreting their meaning and labeling. Although the first principal component indicates the direction of maximum variance, it may not represent the component of interest for the study because it only provides a general profile of the data. This also applies to other selected components. Fourth, the socioeconomic indicators selected were solely due to data availability. The 4 components used for the analysis are based on the literature and are subject to change. However, the characteristics of this study provided quite good and reliable results. The reported socioeconomic and health situation may change as the new national government introduces new policies and development programs. The results should therefore be interpreted with caution and future studies will be required to update the clusters to capture the changes in the socio-economic situation of the county.

6. Conclusion

Using readily available variables, we used PCA to reduce the dimensionality of the data and used a hierarchical and K-means clustering technique to stratify counties in Kenya into five clusters. The grouped counties were then projected onto a geographic map to understand the relationship between their location and socioeconomic and health indicators. The cluster analysis shows significant differences according to socioeconomic and health indicators. The stacked heatmap provided insights into the county's performance and associated variables. The results obtained may be useful to the county and state governments in future plans to promote inclusive and sustainable economic development.

7. Ethics approval and consent to participate

Not applicable.

8. Consent for publication

Not applicable.

9. Data Availability

Code and county-level data used for this analysis are available on Github: `https://github.com/Evanskorir/kenya-counties`

10. Competing interests

The authors declare that they have no competing interests.

Funding

No financial support was received for this study.

11. Acknowledgement

The Stipendium Hungaricum Scholarship Program with Application No. 118250 supported the author.

References

[1] P. K. Ngenoh, Challenges of implementing devolution and planning objectives by the ministry of devolution and planning in kenya, Ph.D. thesis, University of Nairobi (2014).

[2] A. K. Mwenda, Economic and administrative implications of the devolution framework established by the constitution of kenya (2010).

[3] N. Mose, Determinants of regional economic growth in kenya, African Journal of Business Management 15 (1) (2021) 1–12.

[4] T. Achoki, M. K. Miller-Petrie, S. D. Glenn, N. Kalra, A. Lesego, G. K. Gathecha, U. Alam, H. W. Kiarie, I. W. Maina, I. M. Adetifa, et al., Health disparities across the counties of kenya and implications for policy makers, 1990–2016: a systematic analysis for the global burden of disease study 2016, The Lancet Global Health 7 (1) (2019) e81–e95.

[5] J. M. Muthoka, E. E. Salakpi, E. Ouko, Z.-F. Yi, A. S. Antonarakis, P. Rowhani, Mapping opuntia stricta in the arid and semi-arid environment of kenya using sentinel-2 imagery and ensemble machine learning classifiers, Remote Sensing 13 (8) (2021) 1494.

[6] U. M. Wiesmann, B. Kiteme, Z. Mwangi, Socio-economic atlas of Kenya: Depicting the national population census by county and sub-location, Kenya National Bureau of Statistics, Centre for Training and Integrated ..., 2014.

[7] S. Chae, T. Ngo, The global state of evidence on interventions to prevent child marriage (2017).

[8] Z. A. Kwena, S. W. Njuguna, A. Ssetala, J. Seeley, L. Nielsen, J. De Bont, E. A. Bukusi, L. V. C. for Health Research (LVCHR) Team, Hiv prevalence, spatial distribution and risk factors for hiv infection in the kenyan fishing communities of lake victoria, PloS one 14 (3) (2019) e0214360.

[9] H. Lowe, L. Kenny, R. Hassan, L. J. Bacchus, P. Njoroge, N. A. Dagadu, M. Hossain, B. Cislaghi, 'if she gets married when she is young, she will give birth to many kids': a qualitative study of child marriage practices amongst nomadic pastoralist communities in kenya, Culture, health & sexuality 24 (7) (2022) 886–901.

[10] D. M. Nyariki, D. A. Amwata, The value of pastoralism in kenya: Application of total economic value approach, Pastoralism 9 (1) (2019) 1–13.

[11] J. K. Nyoro, Agriculture and rural growth in kenya, Tech. rep., Tegemeo Institute (2019).

[12] P. M. Macharia, E. Mumo, E. A. Okiro, Modelling geographical accessibility to urban centres in kenya in 2019, PLoS One 16 (5) (2021) e0251624.

[13] O. Denis, J. M. Kilonzo, Resource allocation planning: Impact on public sector procurement performance in kenya, International Journal of Business and Social Science 5 (7) (2014) 1.

[14] J. A. ONYANGO, V. KERARO, A. IRUNGU, M. ALUOCH, Adequacy of the commission on revenue allocation parameters for equitable revenue sharing with counties in kenya, International Journal of Innovative Finance and Economic Research 3 (4) (2015) 1628.

[15] M. Çağlar, C. Gürler, Sustainable development goals: A cluster analysis of worldwide countries, Environment, Development and Sustainability 24 (6) (2022) 8593–8624.

[16] A. Merzouki, J. Estill, E. Orel, K. Tal, O. Keiser, Clusters of sub-saharan african countries based on sociobehavioural characteristics and associated hiv incidence, PeerJ 9 (2021) e10660.

[17] S. A. Rizvi, M. Umair, M. A. Cheema, Clustering of countries for covid-19 cases based on disease prevalence, health systems and environmental indicators, Chaos, Solitons & Fractals 151 (2021) 111240.

[18] N. Yakovenko, I. Komov, R. Ten, Cluster approach in assessing the level of socio-economic development of the municipal districts (voronezh region), in: International Science and Technology Conference" FarEaston"(ISCFEC 2019), Atlantis Press, 2019, pp. 201–203.

[19] B. Sadeghi, R. C. Cheung, M. Hanbury, Using hierarchical clustering analysis to evaluate covid-19 pandemic preparedness and performance in 180 countries in 2020, Bmj Open 11 (11) (2021) e049844.

[20] C. Mongi, Y. Langi, C. Montolalu, N. Nainggolan, Comparison of hierarchical clustering methods (case study: Data on poverty influence in north sulawesi), in: IOP Conference Series: Materials Science and Engineering, Vol. 567, IOP Publishing, 2019, p. 012048.

[21] E. K. Korir, Z. Vizi, Clustering of countries based on the associated social contact patterns in epidemiological modelling, arXiv preprint arXiv:2211.06426 (2022).

[22] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2) (2001) 411–423.

[23] G. H. Dunteman, Principal components analysis, Vol. 69, Sage, 1989.

[24] E. K. Korir, Z. Vizi, Clusters of african countries based on the social contacts and associated socioeconomic indicators relevant to the spread of the epidemic, arXiv preprint arXiv:2303.17332 (2023).

[25] R. M. Carrillo-Larco, M. Castillo-Cara, Using country-level variables to classify countries according to the number of confirmed covid-19 cases: An unsupervised machine learning approach, Wellcome open research 5 (2020).

[26] C. Nicholson, L. Beattie, M. Beattie, T. Razzaghi, S. Chen, A machine learning and clustering-based approach for county-level covid-19 analysis, Plos one 17 (4) (2022) e0267558.

[27] G. Gan, C. Ma, J. Wu, Data clustering: theory, algorithms, and applications, SIAM, 2020.

[28] T. Hastie, R. Tibshirani, J. H. Friedman, J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, Vol. 2, Springer, 2009.