Clustering Countries- Improving Measurement in Comparative Political Research[1]

Brooke Luetgert

Research and Computing Center, University of Chicago
luetgert@uchicago.edu

Abstract

Global development indicators are widely used to estimate the macroeconomic and institutional stability of a country. However, as the number of different indicators continues to expand, it has become difficult to interpret this high dimensional data as a whole. Identifying overall trends and assessing how indicators are related to each other can be further obscured by qualitative indices of broad concepts such as democratic stability. While these data continue to increase in dimensionality, empirical research often struggles to explain complex and nuanced societal change from a comparative perspective. Classic approaches to comparative democratic research frequently opt for additive indices like Polity IV, or level measures such as GDP; however, individual indicators often suffer from limited temporal variance. As social scientists, our theoretical work is more advanced than our indicators. The goal of this study is to better understand how we can measure variation across countries with regard to their social and economic division. Employing dimension reduction and subsequent clustering to the vast pool of indicators currently available, the aim is to derive a richer means of describing countries in their development over time. The implications of this explorative, data driven approach are promising: As we better understand the underlying structure of our country level data, we move toward building more effective, objective socio-economic measures and even predictive social science research.

Introduction

The goal of this paper is to take a large data approach to assessing global development. Rather than creating yet another composite index to reduce the dimensionality of our many macroeconomic and socio-political indicators at hand, the purpose of this study is to embrace the richness of the many readily available indicators to best understand patterns and changes across countries over time. We will use principal component analysis (PCA) and K-means clustering to explore how countries of the world are best classified relative to measures of conflict, economic stability and institutional presence. The greater goal is to understand the underlying dimensionality of these global indicators through visualization to better discuss difference through clustering and dimension reduction. This marks a clear departure from the

[1] Please note, this is a draft based on preliminary findings of a highly complex data set. Do not quote or circulate without my permission.

use of composite indices that show very little variation over time and seem to crudely capture the complexity of the abundant data available to comparative scholars.

When dealing with complex data, machine learning offers a means of uncovering underlying patterns and trends that an analyst might otherwise miss. As the number of dimensions and observations increases, standard plotting techniques are insufficient for capturing patterns of variation over time. Machine learning algorithms allow us to obtain these deep insights and build generalizable models that yield highly accurate predictions. For political science, the power of machine learning is a double edged sword, giving us tremendous potential for expanding empirically driven prediction, yet at the same time setting a high standard for data collection and data quality. Where we have big, clean data, we have a powerful method for clustering, multi-class classification, regression, anomaly detection and even predictive recommendation.

In the field of comparative politics, we are frequently faced with building groups of countries that follow a certain pattern or propensity. Indeed, we often want to identify macro (country level) indicators that explain why certain groups of countries behave in a similar fashion, experienced a similar outcome or at risk of a common future trend. As our compute capacity has increased, we have the potential to look deeper into high-dimensional data for previously unobserved trends and even future risk analysis. Clustering is an unsupervised technique for exploring the structure and composition of a given data set. The data are clumped into clusters to observe what groupings, if any, emerge. Each cluster contains a set of data points and a cluster centroid. This centroid represents the average of all data points across all features in that cluster. In contrast to supervised applications of machine learning algorithms, clustering does not involve splitting or labelling the data set. The goal is to learn from the data rather than predict outcomes. This approach is particularly useful when we consider the breadth of time series indicators recorded by large NGOs and international organizations that are intended to capture various dimensions of country performance and division.

Defining the Problem

Trying to quantify how countries differ has busied comparative social scientists since the 1950s. Societal cleavages are a central concept in democratic consolidation theory, parliamentary research on party formation and electoral politics. These cleavages refer to the axes of political conflict within a society. Historically these divisions mark a cultural line dividing citizens within a society into groups with differing political interests. Lipset and Rokkan, (1967) suggest that cleavages define preference constellations that reflect class divides, urban-rural divides, religious and ethnic cleavages as well as the tension between secular and clerical views of state. Within a parliamentary context, when the institutions of representative democracy and rule of law are upheld, cleavages can form party bases or be used to explain preference formation and voting behavior within a given policy space (Gallagher, et al. 2006). While cleavages have been argued to be frozen, others have argued that new conflict lines continue to emerge and have potential for politicization and changing the party system, such as conflicts over migrants (nationalist) and environmental concerns (post-materialist) or even a divide between libertarians and authoritarians (Kriesi, 1998).

For our purposes, we are not concerned with normative claims about the implications of these cleavages. Given the power of clustering methods and interactive visualization, it would be exciting to see if these cleavage patterns are reinforcing or cross-cutting in nature. The more social cleavages there are in a country and the more that these cleavages are cross-cutting, the greater the demand for distinctive, proportional representation and the greater the demand for political parties. Learning more about patterns in these cleavages could also reveal a great deal about very different countries that have surprising similarities in the dimensionality of their societal tensions. Some have suggested that, at least for Western Europe, the classic four cleavages can be reduced to two- with the center/periphery and state/church cleavages reduce to a cultural divide dominated by religion, and the rural/urban and owner/worker cleavages best represented by socio-economic class divisions (Kriesi, 1994). Others contend that the cultural dimension continues to shift periodically to include new social movements such as environmentalism, gender equality and multi-culturalism (Bornschier, 2008 and Inglehart and Norris, 2016). Given the right data, we could apply methods of dimension reduction to address how countries are similar or different in how these conflict dimensions evolve over time, and which dimensions align or stand orthogonally to each other.

Moving beyond the question of societal cleavages, a second discussion related to measurement persists in the measurement of effective political institutions and democratic quality. While many indices have been constructed, the literature often criticizes the validity of these indicators, the limited sample and temporal coverage of these indices and their crude, additive construction that vary little over time. Classic indices created by Freedom House and Polity seek to capture a more nuanced spectrum of regime types scaled from democracy to autocracy, but newer projects such as the Economist Intelligence Unit (EIU, 2012), the Varieties of Democracy Project (Coppedge et al., 2011) and the Democracy Barometer (Bühlmann et al., 2013) are focused on evaluating institutional weaknesses within consolidated democracies.

While indices and measures abound, we have no consensus on how to measure democratic success or identify tendencies toward decline before we reach very large steps such as state interference in media freedom, court appointment or election results. In other words, the attempt to classify regimes and empirically identify subtle changes over time is limited by the quality of measurement.

We clearly observe change in consolidated democracies, or at the very least, shifts within society that strain the stability of well-established, representational forms of government: We have seen large migrant flows from Syria feeding criticism from the radical right across Europe. Growing ethnic-cultural heterogeneity and widening social inequality are urging greater tensions in race relations in the US and UK. Global financial turbulence and the current COVID-19 crisis are adding to dissatisfaction with government provision and efficacy. We have experience a strong rise in populist politics in Russia, through Turkey, Europe, Latin American and even under Trump in the US. The question becomes, when we see these concerning tendencies, is it useful to move away from the strict classification of democracy according to institutions and process? Can we turn toward a view of increasing heterogeneity within societies? What would baseline for this change be?

Where populist politics take root in the political arena, and we see tendencies toward exclusionary politics that embrace autocratic decision-making, I would argue that we have democracy in decline. The gradual backsliding within consolidated democracies reduces representation of group interests and threatens fractures within society that undercut even our minimalist understanding of democracy: That is the guarantee of participation and contestation (Dahl, 1971). There is no consensus on what democracies should 'deliver and produce' (Campbell et al., 2015), but expanding the concept of a good and functioning democracy to include rule of law, vertical and horizontal accountability, participation, competition, responsiveness, freedom and equality (as suggested by Diamond and Morlino, 2004) would suggest that the quality of democracy in consolidated systems may be in decline. Most importantly, this gradual fractionalization can occur long before a classic democratic index captures change at the macro-level. Recent findings by Morlino and Quaranta (2016) suggest that economic crisis in Europe has impacted the quality of democracy. They observe a general deterioration of rule of law, citizens becoming more critical of government policy response and detachment from channels of representation. When we view democracy as more than a "procedural procedure", aspects of performance, welfare production and sustainability could be viewed as output linked to the quality of democracy (Geissel et al. 2016, 575).

The concept of cross-cuttingness and horizontal inequality is that reinforcing, aligned cleavages create deeper divisions and more societal conflict (Lijphart, 1977). For example, if we have a country with Protestants and Catholics that are regionally dispersed, linguistically close, ethnically uniform and equal with regard to public goods provision and economic potential, then this cleavage is less intense because individuals feel “cross-pressured” and belong to multiple identity groups. However, where these cleavages are no longer orthogonal, the reinforcement or shared identity, creates ranked societies with bitter divisions.

For economists, ‘fractionalization’ typically refers to ethnic conflict that leads to political instability, badly designed economic policy and unequal political representation. Among political economists, development (that is economic growth as measured by change in per capita GDP) has been shown to be inversely related to ethnolinguistic fractionalization (Easterly and Levine, 1997). La Porta et al. (1999) find that the quality of government is directly related to underlying ethnic fragmentation in their large scale, cross national research. Economists have also found that localities within the US that suffer from more fragmented societies indicate less efficient public goods provision, lower trust in government and less economic success (Alesina et al., 1999 and Alesina and La Ferrara, 2000). Yet, identifying heterogeneity in a cross-cultural, international context is difficult. Ethnic differences, linguistic divisions and religious heterogeneity may be key to understanding economic growth and political stability, but quantification is challenging (Alesina et al. 2002). Regression analysis is highly problematic in many of these studies because of the pronounced, direct correlation between ethnic fractionalization and latitude.

Assuming that geography does not provide a convincing explanation for corruption, political freedom and democratic institution building, it may seem reasonable to drop the latitude variable from any regression analysis. However, the point here is also a methodological criticism. Regression analysis hinges on the structure of the data. A variable such as latitude functions as a static indicator for group membership to the global South. While structurally critical to the covariance matrix, it points to problems categorizing and analyzing cross-country variation over time.

Measuring ethnic diversity is difficult. Fearon (2003) suggests that creating a cross-national list of ethnic groups is challenging without careful survey results. Where we do not have aggregated micro-level data about identity, we are left to construct diversity indices. This is particularly difficult in a historical context for regions where we may not have reliable and recorded data on minority groups and individual identification. Gurr (1996) provided the first large-N, cross sectional study on group oppression, protest and rebellion in 115 countries, commonly called the Minorities at Risk (MAR) study , but the data have faced criticism due to a perceived sample selection bias. Alesina et al. (2002) also attempted to distinguish between ethnic, linguistic, and religious groups in a sample of about 190 countries and then use their lists to construct measures of ethnic, linguistic, and religious fractionalization. Unfortunately, the coding distinction among these groups is muddied by overlapping identities.[2] Ethnic fractionalization is understood as the probability of two individuals selected at random from a country will be from entirely different ethnic groups. Fearon (2003) establishes a measure of cultural fractionalization, constructed analogously to his ethnic fractionalization indicator, which should capture drawing two people at random from a country and computing their cultural resemblance. In a country with one language group, or set of ethnic groups, that all speak unrelated languages (not dialects), the resemblance is close to 1. As cultural fractionalization increases, the indicator approaches 0.

Do religious, ethnic and linguistic divisions shift over time? Given recent examples such as the brutal disintegration of Yugoslavia and Rwanda, or the ongoing conflict between the government and ethnic Kurds in Turkey, the answer appears to be that the salience of these identities may shift more than the underlying identities. We may observe migratory flows as exogenous system shocks, but the underlying linguistic, ethnic and religious divides appear to be slow changing. Ethnic differences make it harder for people to cooperate. Linguistic differences reduce communication and often demarcate regions or pockets of similar linguistic groups. Without microlevel survey analysis, we cannot estimate ideological distance between these groups (although that would arguably be a better indicator for cleavage), but for now, we must turn to the abundance of available macro indicators for division and fractionalization.

Describing the data

Our goal is not to identify one or two indicators that capture fractionalization, cleavage or even institutional performance over time. Instead, we want to maximize variance over time and across countries by drawing in as many variables as we can find with dense converge across the data panel. We then want to identify clusters of similar countries and break down what

[2] In an ideal case, we might choose to identify ethnic fractionalization as those that have been legislated into existence and do not have a naturalized history, meaning that despite lacking a common language, shared customs or common religious practice, we may observe self-identification with an “ethnic” group. Here Fearon (2003) uses reference to identification as “Jewish” to support the distinction between ethnic identity and religious identity, although the overlap is clear. Furthermore, the identification of ethnic groups includes only those larger than 1% of the country population.

drives these clusters (particularly if we observe change over time). In contrast to regression analysis, where we must limit our explanatory variables and hope that we have a dependent variable with enough variance to make a reasonable claim about any underlying hypotheses, here we don't have hypotheses or a dependent variable. For purposes of this study, having more indicators for each observation is clearly preferred.

Despite multiple APIs and a multitude of online repositories, the assembly of data remains a formative task. Very briefly summarized, the data for this study were extracted from the World Bank/IMF, Freedom House, WHO, UN, ILO, IDEA, Polity IV, Gapminder and CREG[3]. From these sources, we begin with more than 112 indicators for child mortality ages 0-5 per 1000 born, average number of children born to each female, civil liberties, co2 emissions, the Polity IV[4] democracy score, the percentage of population with a fixed telephone line, GDP per capita, the Gini index for poverty, the average schooling of women, the number of medical doctors per 1000, the number of murder deaths per 1000, political rights index, population total, ethnic fractionalization, average life expectancy, availability of clean water, unemployment, human development index, agricultural land % of territory, rural population %of total, primary school completion rate, inflation, percentage of females in total labor force, index of religious freedom, index of judicial independence, social group equality, civil society participation, electoral participation, protection of fundamental human rights and level of military expenditure. This brief selection of thirty indicators represent the most prominent among the data set for there exceptional coverage over time and across the world. The indicators are intended to capture both level difference across countries as well as differences in resources and structural cleavages or fractionalization over time. The data have been assembled for 184 countries for the longest possible time frame of each indicator. While a handful of indicators are available since 1800 (projected and compiled by Gapminder), the data for the period prior to 1960 is only sparsely populated. For this initial exploratory study, we will begin with 35 central indicators over a period from 1975 to 2018.

---

[3] Drazanova (2020) recently released a new measure for ethnic fractionalization that varies over time for 162 countries from 1945-2013. This Historical Index of Ethnic Fractionalization (HIEF) measures the degree of ethnic fractionalization based on the annual percentage of ethnic groups in each country. This corresponds to the probability that two randomly drawn individuals within a country are not from the same ethnic group. The original data were collected within the Composition of Religious and Ethnic Groups (CREG) project led by the Cline Center for Democracy, University of Illinois at Urbana-Champaign.

[4] Polity IV is designed to measure states on a uni-dimensional scale of autocracy/ democracy, where the component indices have been scored and weighted by hand before being reduced to this singular dimension.

Methodological Approach

PCA has become a widely used statistical technique in the social sciences to reduce the dimensionality of the underlying data matrix by identifying, separating and sorting features explaining the most variance in the data in descending order. When we cannot visualize a multi-dimensional hyperspace, such as the 112 dimensional space created by the indicators described above, PCA yields a low-dimensional representation of the dataset by finding a sequence of linear combinations of the variables that contain maximal variance but are mutually uncorrelated. These principal components are orthogonal to each other, and the loadings on each factor reveal the amount of variance contributed by that factor. PCA removes redundancy present in the dataset by projecting the data to a different vector space. This transformation multiplies the given data by a rotation matrix (Eigen vectors of the covariance matrix). It is a fast and flexible, unsupervised method for dimensionality reduction that preserves maximal variance among the underlying data. With high dimensional data, PCA is often a good first step to understanding the main variance and intrinsic dimensionality of the data. The primary advantage of PCA is the identification of principal components that are clearly defined as underlying the structure of the data. With large data sets, we are able to visualize the structure of the relationships among variables rather than just looking at correlation of pairs. PCA does have a few drawbacks: It can be sensitive to the scaling of variables with vastly different units of measurement, and it tends to be biased in the direction of larger measurement as well as strongly affected by outliers. Further, PCA does not perform well when there are nonlinear relationships within the data. For now, it is enough to say that it is an observational tool for identifying hidden linear correlations among variables within a dataset.[5]

Given the goal of clustering our data, we do not have a target response variable. We want to find unknown subgroups and gain a deeper insight of the variation within our data. The

---

[5] When we observe that x and y values are not fundamental to the relationship that we are exploring in the data, we should consider whether the pair wise distance between each point and other points is fundamental. A distance matrix gives us a representation of our data that is invariant to rotations and translations. Multidimensional Scaling (MDS) uses the distance matrix to create a low-dimensional representation of the data that preserves the distance between every pair of points in the data set. If the underlying data are non-linear in nature and PCA does not yield satisfactory dimensional reduction, MDS can offer a valuable alternative by preserving the distances between each data point. While both algorithms minimize dimensionality, PCA preserves covariance of the data and MDS preserves the distance between data points. These two are the same if the covariance in the data is the euclidean distance between data points in the high dimensional space. Distance can become a decisive underlying feature if we are interested in an ideological policy space, reference to geographic region or conclusion about preferences relative to other actors.

interpretation of these results may be quite subjective, but could also serve as a pre-processing step for supervised learning.

What is the difference between clustering and PCA? We recall that PCA identifies low-dimensional representation of the observations that explains a large portion of the variance in the data matrix. Clustering attempts to identify relatively homogenous subgroups among the observations. The clustering algorithms seek to optimally divide and label groups of data points that are most similar given the properties contained in the underlying data. Because we do not have a target response, we cannot objectively evaluate our solution performance or report errors. In this application, we are data mining using a machine learning clustering algorithm. The feature variables are observed for each data point, but the class variable is unknown. We are running parameter learning on a Bayes net, alternating between estimating the hidden variable based on parameters and computing the parameters based on the observed data. Each iteration in this process brings us closer to the local optimum estimates for the unobserved cluster centroid.

The k-means algorithm (MacQueen, 1967) is designed to search for a preset number of clusters (k) within an unlabeled multidimensional dataset. Here, the cluster center is the arithmetic mean of all points within the cluster. The points within a given cluster are closer to their own cluster center than to other cluster centers. The k-means algorithm is driven by an expectation-maximization algorithm. This involves estimating cluster centers and repeating until convergence is achieved. It is best described as an iterative process, where our expectation about which cluster each point belongs to is updated and the fitness function is maximized by taking the simple mean of the data in each cluster until we achieve convergence.

Applying K-Means Clustering to Country Data

No matter how good the approach, if the data are low quality, small in number or error plagued, the result may not be useful. We must transform, clean and pre-process the data to remove outliers, impute missing values and avoid structural errors that will impede the analysis. Global indicators data typically suffer from a great deal of missing observations. This is particularly problematic when we aim to apply methods for dimensionality reduction and clustering. It is important to first remove any systematic bias by eliminating early time periods and certain countries where our data are particularly weak. Using a threshold of no more than 50% missingness, the data matrix is reduced to 8792 country*year observations and 34 indicators for the period 1975-2018.[6] Here we should note that most missings are present on subjective measures such as the Human Development Index (3850 missings), Social Group Equality (2358 missings) and HIEF ethnic fractionalization index (3620 missings). These missings are often prevalent for a given country and block of years rather than intermittently dispersed throughout the matrix.

After the data were trimmed, imputation remains our best alternative for correcting missing values. Based on multiple attempts with matrix factorization, replacement with median values and a KNN or (K-Nearest Neighbors) approach, imputation with KNN reduces the mean average error and mean square error for the data under study. The KNN approach suggests a prediction of a missing value in one country by averaging the changes of this variable in similar countries (where similarity is measured as Euclidean distance) and estimating the imputed value accordingly. KNN imputation will only work on numerical data; however, the indices in this study have at least ten categories and have been encoded prior to imputation. This algorithm requires specifying the number of nearest neighbors to average.[7]

The second observation regarding the composite data set with imputed values is the vast difference in distributions across the indicators. Some of the variables include scales from zero to one hundred; others contain percentages, and some are very large measures that are heavily skewed like GDP and total population. Each indicator is transformed to optimize the distribution. Finally, we standardize each indicator with the z-score to set the variable means to zero and standard deviations to one. This provides the best level playing field for evaluating the correlation matrix and beginning dimension reduction. PCA results in a feature subspace that maximizes the variance along the axes. Accordingly, the results would be strongly influenced by different scales of non-standardized indicators.

Once the data are clean, transformed and standardized, the next task is feature selection, or dimensionality reduction. This is the first evaluation of features that will be used for estimates of model parameters, their significance and correlation estimates among the features. In the machine learning process, reducing dimensionality reduces model complexity and increases computational speed. It is the process that transforms original predictors into a smaller set of newly defined predictors, which can then be used to fit a model. We begin with a technique of multidimensional scaling known as principle component analysis (PCA).

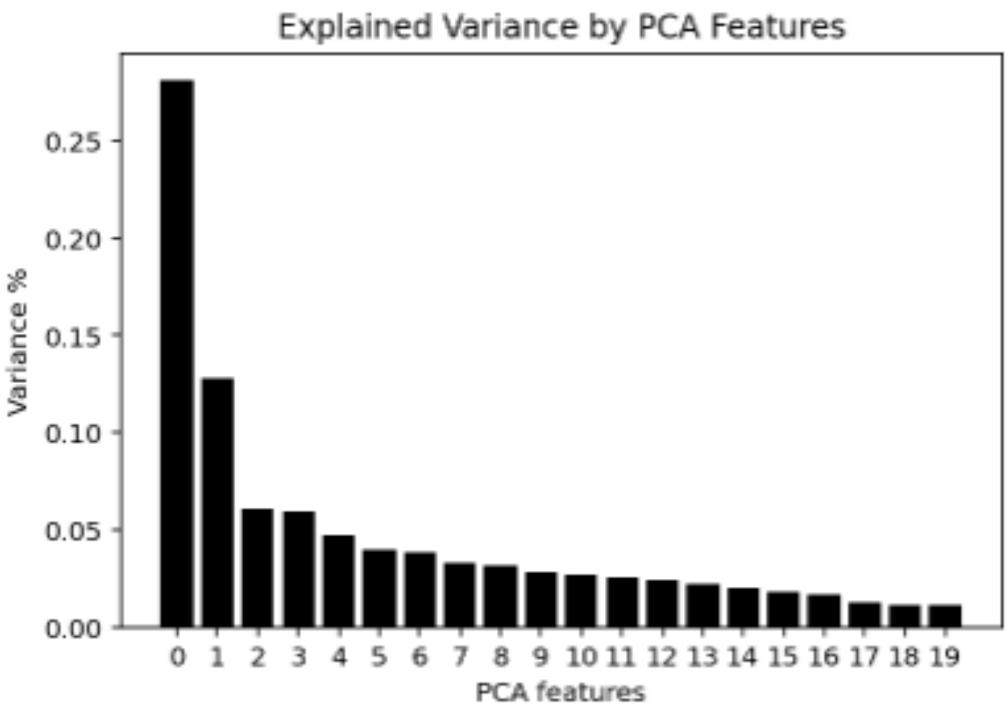
---

[6] The matrix expands to 13,671 observations for the period from 1960-2018.

[7] This value has been set to D=3 in this paper.

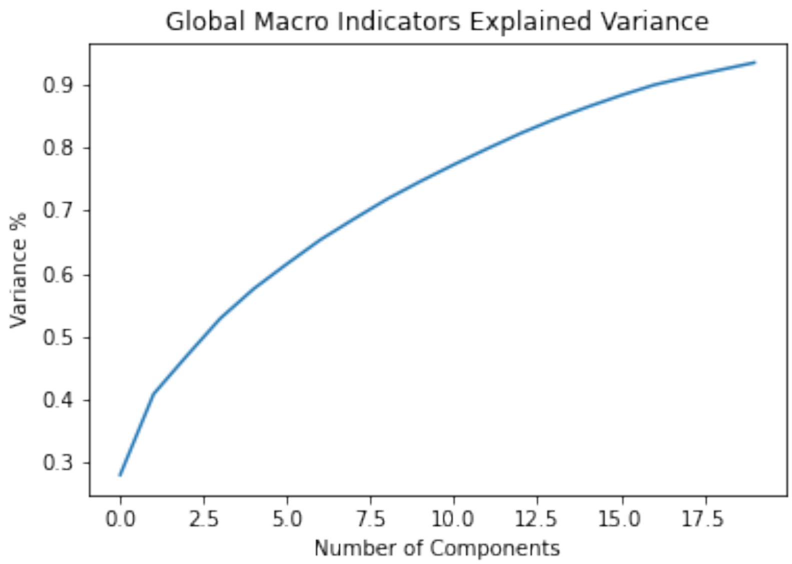
After scaling and running the PCA algorithm, we obtain the eigenvalues measuring the amount of variation retained by each principal component. These eigenvalues are large for the first principal components and smaller for subsequent ones. We determine the number of principal components to consider by looking at the eigenvalues. Below we see that the contribution of the first two components is largest, but that components three and four also add more than 5% each to the overall explained variance in our data.

Figure 1: Variance explained by each PCA component dimension



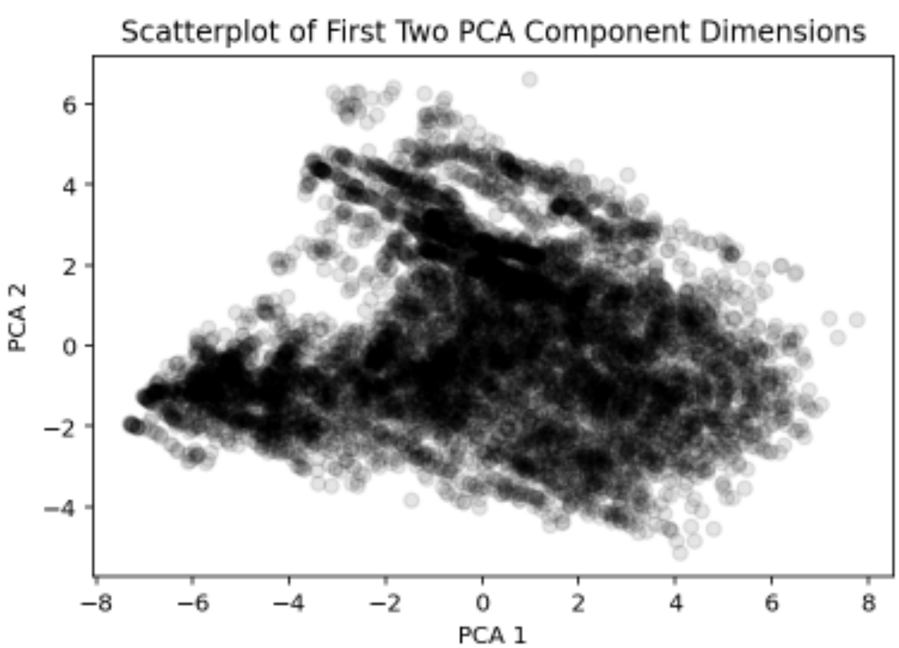
If we want to explore a plot of the cumulative summation of the explained variance by adding each new feature, we can refer to the Figure 2 below. It is important to note that the elbow or bend in the plot reflects the drop off in added value in considering additional features. Here, that elbow shows again a decline in added explained variance just after two PCA features.

Figure 2: Cumulative summation of the explained variance



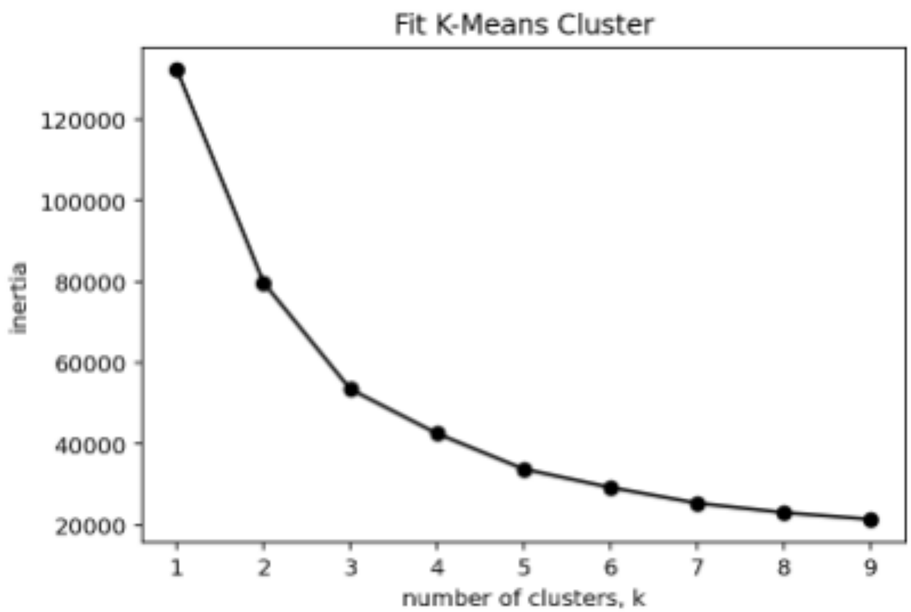
For our purposes, the story about our data becomes most exciting when we refer back to our individual observations. For a two dimensional representation of our countries in PCA projection, we can look at a scatterplot, noting that each data point corresponds to the coordinates of a country*year observation by (PCA feature 1, PCA feature 2). Figure 3 shows this 2D scatter plot, revealing that our 8,279 observations may form clusters with something like three to four clusters of similar data points. We also observe very few outliers, supporting the finding that two dimensions capture a convincing portion of the variance in our underlying data. However, we should note that at first glance, we do not have clear borders between clusters.

Figure 3: Scatter plot of first two PCA component dimensions



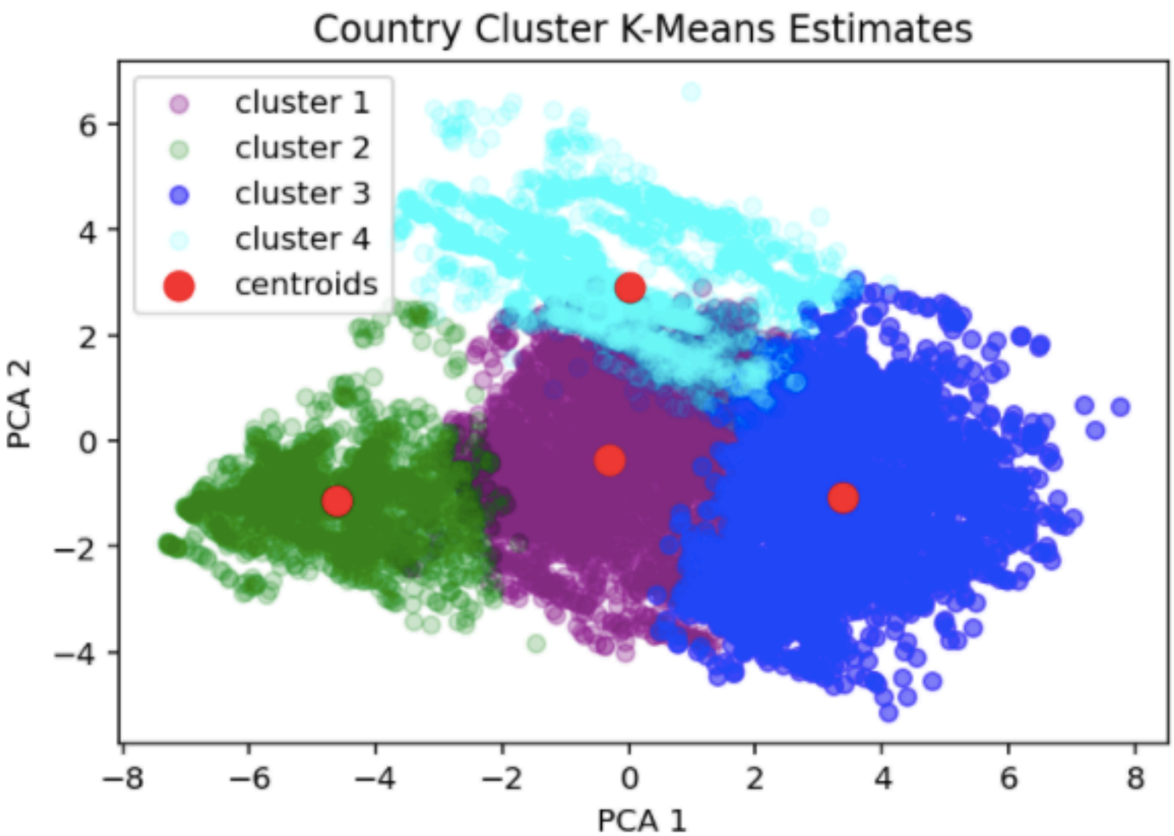
In the second step of our analysis, we need to dig deeper into the clusters of countries that we suspect based on the plot in Figure 3. For clustering, we will take a K-means approach. The k-means approach to clustering is highly attractive for its simplicity, but the non-probabilistic nature and use of simple distance from the cluster center used to assign cluster membership is problematic when we have overlapping or elliptical clusters, lacking firm circular/ spherical boundaries to act as cuts-off points for assignment. Keeping this in mind as we proceed, we first create a k-means instance with k number of clusters. We will fit the model to the sample by referencing the PCA components plotted above. In Figure 4, we consider the number of clusters, k, on the inertia of the fit of the k-means cluster.

Figure 4: Scatter plot of first two PCA component dimensions



The elbow method of selecting the number of clusters will not always work well because the error function is monotonically decreasing for all k; however, it appears that our data include about four clusters for our closer consideration. In this final step, we want to view our clusters first in the two dimensional world of plotting the first two PCA features against each other, and in the final figure, considering this same data cloud with centroid estimates in a three dimensional view of the first three PCA components that we considered above.

Figure 5: K-Means estimate of country clusters in 2D



We know that the K-means algorithm works best with near-spherical clusters, and would not work well when clusters are oddly shaped or severely elongated on one dimension. When there is overlap in the clustering, we do not have a measure of uncertainty for the assignment of each data point. This approach to clustering is highly attractive for its simplicity, but we must look deeper for firm circular/ spherical boundaries to act as cuts-off points for assignment in three dimensions. This 2D plot would suggest that the clusters are more or less circular in 2D, with the exception of the light blue cluster marked 4. This plot also suggests that we do not have optimal border distinction to clearly demarcate the prominent three clusters.

There are two drawbacks to a k-means analysis: First, the algorithm hinges on a preset (determined a priori) number of clusters and cannot learn this value from the data. Second, k-means is limited to linear cluster boundaries. If the data exhibit complicated, geometric clusters, k-means clustering will fail. With low dimensionality data, k-means may not perform well without probabilistic cluster assignment. The final result is sensitive to the initial selection of cluster centers, which can lead to results that are difficult to replicate. Changing the order of the data can also yield different solutions. Possible solutions to these issues include computing k-means for a range of k values, such as varying between 2 and 10. Then choose the best k by comparing the results obtained for these values. Likewise, compute the K-means algorithm several times with different initialization values and select the run with the lowest total within-cluster sum of squares as the final clustering solution.

Figure 6: K-Means estimate of country clusters in 3D

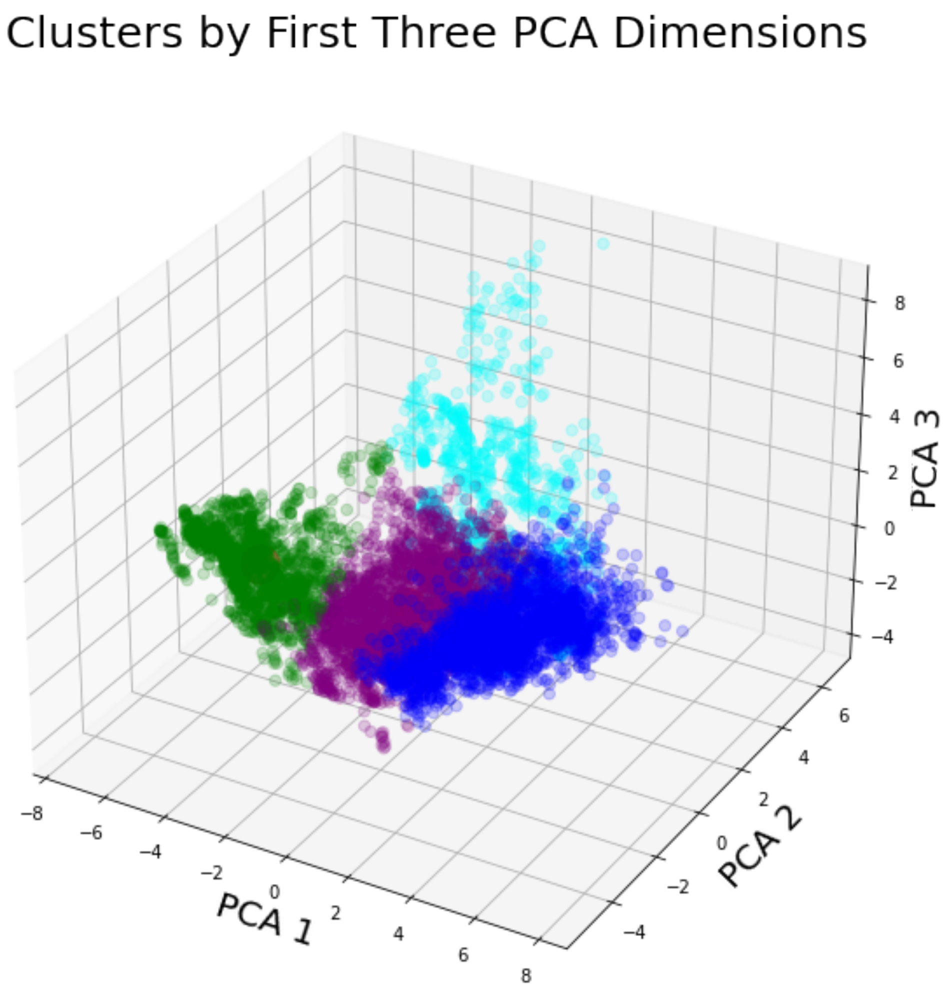
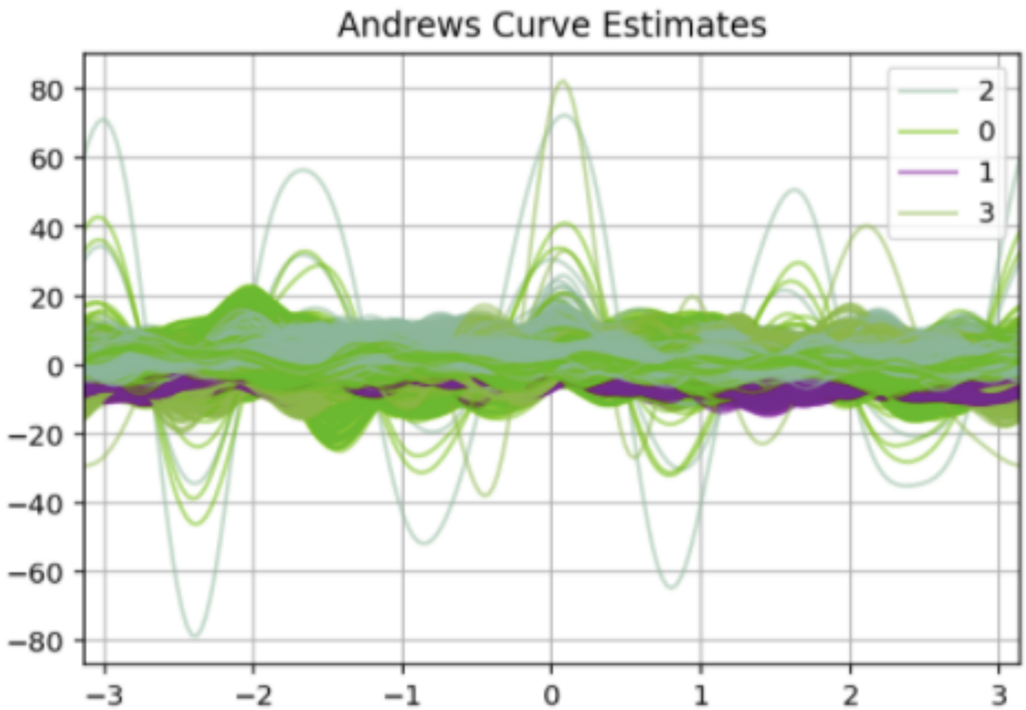


Figure 7: Andrews Curve Estimates of PCA features



On a final note, we must make a few observations about the data and the interpretation of the clustering method. The basic problem of graphical displays of multivariate data is dimensionality. Scatterplots can work well in up to three dimensions, but we see that even here, we lack interaction with the data points in print media. In order to tell our narrative about countries and how they differ, interactive plotting of these clustering results is essential. For now, we can take a last look at the Andrews curve estimates for the latent patterns that we are trying to observe. The Andrews curve is an alternative means of visualizing structure in high-dimensional data. A subgroup of observations appears as a set of similar curves. The order of the variables is important for interpretation, here optimized by PCA. Unfortunately, with 8,279 observations, we have a bad 'signal to ink' ratio, leaving us with many curves overlaid in one plot. The curves outside of the ribbons of data represent outliers. Again, the Andrews curve would confirm that most variation in the underlying data structure can be captured by reducing to four clusters or sub-groups of data. Here, we again see that many of the more extreme observations would be clustered to the same group (indicated as light blue above).

While still preliminary, these first results build on the idea that democracy may be a latent variable (Trier and Jackman, 2008). More specifically, given the richness of the more than 112 global indicators compiled in this study and the tremendously powerful algorithms available for unfolding the underlying variance within our data, we may be moving one step closer to finding objective measures to describe countries similar in the dimensionality of their primary cleavages and fractions as they change over time.

Going Forward

This analysis considers a novel means of exploring the many macro indicators available to us to observe how countries cluster. The goal was to capture the multi-faceted character of societal divisions and view the data in a cross national setting to recognize patterns over several decades. The intention was not to measure democracy, but to gain a sensitivity for how divisions, as they shift, may contribute to representational backsliding or the deepening of asymmetries. The trade-off when working with large data analysis of global indicators is an optimization game of maximizing the number of dimensions used to describe these cross-national settings and the time range and country coverage. Going forward, this rich conglomeration of data needs to be more carefully appended to increase the period of observation or, at least, add more indicators. Again, by reducing the country sample, an increase in the number of usable indicators becomes more feasible as we can draw on regional statistics, particularly as we move away from unstable and poorly recorded countries in the former USSR, the Middle East and Africa. Finally, PCA and K-means clustering are powerful, but what about substantive interpretation? We need to look closer into the issue areas that are captured by the first and second features. We also need to look at changing patterns over time. This study began with a hunt to increase variance among our indicators over time. Society is dynamic. Institutions are sticky. However, certain trends may be visible through careful analysis of these macro indicators before we reach dramatic institutional changes that ultimately only appears as a negligible change in the classic democratic indices. Political science research is moving toward predictive risk assessment, but we will first need to look deeper at country clustering over time and at a level of careful detail that is only available in interactive, multidimensional plots where we can identify each observation and trace patterns in clustering over time.

References

Alesina A., R. Baqir and W. Easterly. 1999. "Public Goods and Ethnic Divisions" *Quarterly Journal of Economics*. 114(4): 1243-1284.

Alesina A. and E. La Ferrara. 2000. "Participation in Heterogeneous Communities", *Quarterly Journal of Economics.* 115(3): 847-904.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, R. Wacziarg. 2002. "Fractionalization." NBER Working Paper 9411.

Bornschier S. 2010. Cleavage Politics and the Populist Right. The New Cultural Conflict in Western Europe, Philadelphia: Temple University Press.

Bühlmann, M., W. Merkel, L. Müller and B. Weßels 2008. "Wie Lässt sich Demokratie am Besten Messen? Zum Forumsbeitrag von T. Müller und S. Pickel." *Politische Vierteljahresschrift* 49(1): 114–122.

Campbell, D., E. Carayannis and S. Rehman 2015. "Quadruple Helix Structures of Quality of Democracy in Innovation Systems: The USA, OECD countries, and EU member countries in global comparison." *Journal of the Knowledge Economy* 6(3): 467–493.

Clarke, H. and A. Kornberg 1971. "A Note On Social Cleavages and Democratic Performance," *Comparative Political Studies* 4(3): 349-360.

Coppedge, M., J. Gerring, D. Altman, M. Bernhard, S. Fish, A. Hicken, M. Kroenig, S. Lindberg, K. McMann, P. Paxton, H. Semetko, S. Skaaning, J. Staton and J. Teorell 2011. "Conceptualizing and Measuring Democracy: A new approach." *Perspectives on Politics* 9(02): 247–267.

Dahl, R. 1956. *A Preface to Democratic Theory*. Chicago: Univ. of Chicago Press.

Dahl, R. 1971. *Polyarchy: Participation and Opposition*. New Haven, CT: Yale University Press.

Diamond, L. and L. Morlino 2005. *Assessing the Quality of Democracy*. Baltimore, MD: Johns Hopkins University Press.

Drazanova, L. 2020. "Introducing the Historical Index of Ethnic Fractionalization (HEIF) Dataset: Accounting for Longitudinal Changes in Ethnic Diversity." *Journal of Open Humanities Data* 6: 6 DOI: https://doi.org/10.5334/johd.16

Easterly W. and R. Levine. 1997. "Africa's Growth Tragedy: Policies and Ethnic Divisions", *Quarterly Journal of Economics*. 111(4): 1203-1250.

EIU 2012. Democracy Index 2012: Democracy at a Standstill. London: The Economist Intelligence Unit.

Fearon, J. 2003. "Ethnic and Cultural Diversity by Country," *Journal of Economic Grouwth*. 8: 195-222.

Gallagher, M., M. Laver, P. Mair 2006. *Representative Government in Modern Europe*. New York: McGraw-Hill, pp. 268-269.

Geissel, B., M. Neuer and H. Lauth 2016. “Measuring the Quality of Democracy: Introduction.” *International Political Science Review*. 37(5): 571-579.

Gurr, T. 1996. *Minorities at Risk III Dataset: User’s Manual.* CIDCM, University of Mayland.

Inglehart R.F., P. Norris 2016. “Trump, Brexit, and the Rise of Populism: Economic Have-Nots and Cultural Backlash”, Harvard Kennedy School, Faculty Research Work-ing Papers (https://research.hks.harvard.edu/publications/getFile.aspx?Id=1401).

Kriesi, H. 1994. Les Démocraties Occidentales. Une Approche Comparée, Paris: Economica.

Kriesi, H. 1998. "The Transformation of Cleavage Politics. The 1997 Stein Rokkan lecture.” *European Journal of Political Research*. 33: 165–185.

La Porta R., F. Lopez de Silanes, A. Shleifer and R. Vishny. 1999. “The Quality of Government”, *Journal of Law, Economics and Organization.* 15(1): 222-279.

Lijphart, A. 1977. *Democracy in Plural Societies: A Comparative Exploration*. New Haven: Yale University.

Lipset, S. 1959. "Some Social Requisites of Democracy: Economic Development and Political Legitimacy." *American Political Science Review* 53(2): 69-105.

Lipset, S. and S. Rokkan.1967. "Cleavage Structures, Party Systems, and Voter Alignments: an Introduction". In Lipset, S. And S. Rokkan (eds.). *Party Systems and Voter Alignments: Cross-National Perspectives*. The Free Press. pp. 1–64.

MacQueen, J. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” In *Proceedings of the Fifth Berkley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281-97. Berkeley, CA.: University of California Press.

McDoom, O. And R. Gisselquist (2016) “The Measurement of Ethnic and Religious Divisions: Spatial, Temporal and Categorical Dimensions with Evidence from Mindanao, the Philippines.” *Social Indicator Research.* 129: 863-891.

Morlino, L and M. Quaranta (2016) “What is the Impact of the Economic Crisis on Democracy? Evidence from Europe.” *International Political Science Review* 37(5): 618-633.

Rae, D and M. Taylor. 1970. *The Analysis of Political Cleavages*. New Haven: Yale Univ. Press.

Trier, S. and S. Jackman. 2008. “Democracy as a Latent Variable,” *American Journal of Political Science*. 52(1): 201-217.