1_METHODS

1.1. Data

The DHS has been gathering nationally representative data at regular intervals since its inception in 1984 (USAID, 2020b). The data span social, behavioral, geographic, economic, and health related issues, including HIV-related data and sexual behavior, especially since the emergence of HIV/AIDS as a public health crisis (Curtis and Sutherland 2004). This has naturally made the DHS the primary source of data for this study.

Data was collected with the intent of including all indicators possibly influencing the risk of acquiring HIV for individuals aged 15-49 years old, or possibly affecting the course of HIV epidemics on local, regional, and national levels. We used the StatCompiler tool (USAID, 2020a) to collect the surveys of 29 SSA countries between 2000 and 2018, that were available up to September 2020. While we aimed to go as far back in time as possible to include the emergence of HIV, earlier surveys lacked reliability in the specific indicators of interest, resulting in a total of 80 surveys collected, with each of the 29 countries having between 1 and 6 surveys over that time period.

Similarly as Merzouki et. al. (2021), we retained only those variables that varied significantly across regions and did not strongly correlate with other variables. The data were represented as percentages, with mean number of sexual partners in lifetime scaled using min-max normalization. The resulting 46 attributes are shown in Table 1 and create a 46 dimensional image of each survey collected, in turn each representing a snapshot of a country at a specific point in time.

While most of these attributes were collected using the DHS StatCompiler tool ("STATcompiler" 2020), latent variables were either replaced with alternative sources in the cases of male circumcision healthdata.org, urban/rural divide World Bank, Gini coefficient World Bank, ART coverage World Bank, and

religion data.world correlates of war dataset. Other latent variables for which alternative sources were unavailable were imputed using additional topic-related variables available from the DHS using multiple iterative chained equations – while the detailed analysis of the imputation technique lies outside the scope of this study, all details can be found in Annex 1. Finally, estimates for national HIV incidence and prevalence between the years 1990 and 2019 were obtained from the latest (2020) UNAIDS' AIDSinfo UNAIDS.

1.2. Ethics approval

This study uses publicly accessible data aggregated through StatCompiler, the World Bank, healthdata.org, Correlates of War, and UNAIDS' AIDSinfo. No ethics approval was needed from our side.

1.3. Analysis

§ 1.3.1 – Sociobehavioral characteristics

We reduced the dimensionality of each survey using Principal Component Analysis (PCA) (Pearson 1901). PCA allows the transformation of the 46 original dimensions of each survey into a smaller subset of uncorrelated indices called principle components (PCs) along which the variation of data is maximized and information loss is minimized (Jolliffe and Cadima 2016). As the PCs consist of a linear combination of the initial 46 dimensions, they can be interpreted in terms of the original demographic, socioeconomic, and behavioral characteristics, and can thus be used to construe sociobehavioral heterogeneity or homogeneity across SSA, both spatially and over time. To capture the longitudinal trends in our dataset, and in an effort to facilitate comparison with the results obtained by Merzouki et. al. (2021), we obtained the rotation matrix using only the most recent surveys i.e. only those collected since 2015. This rotation matrix was then used to project all other surveys onto the dimensionally reduced PCA space.

While the first 2 PCs, which explain the most variance, are used to represent the axes on a 2D-space to provide a visual perspective of similarity between the surveys, we used a consensus clustering method to provide a more robust means of identifying groups of similar surveys in terms of sociobehavioral characteristics. We predefined 3 as the number of clusters to obtain as was obtained by Merzouki et. al. (2021). The consensus clustering method consisted of three hierarchical clustering and three k-means clustering (Lloyd 1982) partitions, each run on a different number of PCs: the first with only the 2 main PCs, the second with the n PCs that accounted for 95% of the overall variance, and the third with the entire 46 dimensional data set. We identified longitudinal trends and determined the key sociobehavioral characteristics driving change by visualizing both the trajectories of countries and of the cluster centroids on the 2 dimensional PCA space over time. Finally, we compared the HIV incidence of countries between these clusters by visualizing the HIV incidence using box plots.

§ 1.3.3 . Course of HIV epidemics and effective contact rate β

We calculated for each country their effective contact rate β from the estimates of the basic HIV indicators of incidence and prevalence - details of the derivation of this indicator can be found in Annex 2. The effective contact rate β is a measure of incidence given the prevalence (i.e. the number of new infections given the local context of the HIV epidemic) and is effectively both a proxy for the sociobehavioral characteristics of the population while also directly gauging the transmission dynamics of HIV within that population, doing so independently of the local context of the HIV epidemic. We compared the progression of the effective contact rates since 1990 across SSA to that of HIV incidence and compared them across the clusters obtained previously and visualized them again with box plots. ("Sub-Saharan Africa Male Circumcision Geospatial Estimates 2000-2017" 2020)

Lastly, we identified clusters of countries based on the evolution of their effective contact rates by first scaling the time-series such that they had zero mean and unit variance (following the assumption that their amplitudes are not as informative as their relative shapes) and used k-means clustering (Lloyd 1982) with a dynamic time warping barycenter averaging algorithm (Petitjean, Ketterlin, and Gançarski 2011). We measured the quality of the clustering results using a silhouette score (Rousseeuw 1987) and determined the optimal number of clusters. We compared this new set of clusters with the ones previously obtained using sociobehavioral characteristics and compare their respective HIV incidence and effective contact rates.

1.4. Tools and software

We used the open source Python language (Van Rossum and Drake Jr 1995), complemented by Pandas (McKinney 2010), NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020), TSLearn (Tavenard et al. 2020), Scikit-learn (Pedregosa et al. 2011) for all data aggregating, data wrangling, and data analysis, and used Plotly (Inc. 2015) for data visualizations. All code, data, and other material are available on RenkuLab (https://renkulab.io/gitlab/jeffrey.post/ssa_hiv_ml).

References

Curtis, S L, and E G Sutherland. 2004. "Measuring Sexual Behaviour in the Era of HIV/AIDS: The Experience of Demographic and Health Surveys and Similar Enquiries." *Sexually Transmitted Infections* 80 (suppl 2): ii22–27. https://doi.org/10.1136/sti.2004.011650.

Harris, Charles R., K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585: 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Inc., Plotly Technologies. 2015. "Collaborative Data Science." Montreal, QC: Plotly Technologies Inc. 2015. https://plot.ly.

Jolliffe, Ian T., and Jorge Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2065): 20150202. https://doi.org/10.1098/rsta.2015.0202.

Lloyd, S. 1982. "Least Squares Quantization in PCM." *IEEE Transactions on Information Theory* 28 (2): 129–37. https://doi.org/10.1109/TIT.1982.1056489.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 51–56.

Merzouki, Aziza, Janne Estill, Erol Orel, Kali Tal, and Olivia Keiser. 2021. "Clusters of sub-Saharan African countries based on sociobehavioural characteristics and associated HIV incidence." *PeerJ* 9 (January): e10660. https://doi.org/10.7717/peerj.10660.

Pearson, Karl. 1901. "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11): 559–72. https://doi.org/10.1080/14786440109462720.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (85): 2825–30. http://jmlr.org/papers/v12/pedregosa11a.html.

Petitjean, François, Alain Ketterlin, and Pierre Gançarski. 2011. "A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering." *Pattern Recogn.* 44 (3): 678–93. https://doi.org/10.1016/j.patcog.2010.09.013.

Rousseeuw, Peter J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20: 53–65. https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7.

"STATcompiler." 2020. https://www.statcompiler.com/en.

"Sub-Saharan Africa Male Circumcision Geospatial Estimates 2000-2017." 2020. Institute for Health Metrics; Evaluation (IHME). https://doi.org/10.6069/42Y8-WX60.

Tavenard, Romain, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, et al. 2020. "Tslearn, a Machine Learning Toolkit for Time Series Data." *Journal of Machine Learning Research* 21 (118): 1–6. http://jmlr.org/papers/v21/20-091.html.

Van Rossum, Guido, and Fred L Drake Jr. 1995. *Python Tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17: 261–72. https://doi.org/10.1038/s41592-019-0686-2.