

Title: Using Cluster Analysis to Group Countries Based on Non-Communicable Disease Risk Factors

Authors: Nicholas Peckham, supervised by Dr James Bentham

Background: Classifying countries based on non-communicable disease (NCD) risk factors helps international agencies, such as the World Health Organisation, to plan effective health strategies and interventions. We explored whether cluster analysis could robustly and effectively group countries by NCD risk factors.

Methods: Pooled country-level estimates were collected for 30 NCD risk factors. The data were standardised, and principal component analysis was used to reduce the dimensions and explain underlying patterns in the data. K-mean, k-medoid, hierarchical, fuzzy, density, and distribution-based clustering techniques were applied on the principal components, using the Euclidian distance measure where appropriate. Clusters were compared using 5 different quantitative validation measures, and were assessed using prior knowledge of the data.

Results: Most of the clustering techniques produced relevant, homogeneous clusters based on the NCD risk factors. However, fuzzy and density-based clustering did not sufficiently address outliers and grouped countries less reliably. K-means and model-based t-distribution clustering algorithms handled outliers more effectively than the other methods, producing groups with the lowest inter-cluster variance and largest intra-cluster differences.

Conclusion: Cluster analysis can be a useful and effective way to group similar countries by NCD risk factors. However, as some countries did not fit easily into defined clusters, the limitations of clustering must be considered when investigating a broad range of NCD risk factors. This project found that using k-means and t-distribution based cluster algorithms produced the most homogenous groups, and therefore would be recommended to be explored further as effective methods of grouping countries by disease risk factors.