International Journal of Hygiene and Environmental Health 217 (2014) 379–385

Contents lists available at ScienceDirect

International Journal of Hygiene and Environmental Health

journal homepage: www.elsevier.com/locate/ijheh

Country clustering applied to the water and sanitation sector: A new tool with potential applications in research and policy

Kyle Onda, Jonny Crocker, Georgia Lyn Kayser, Jamie Bartram*

The Water Institute, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, 148 Rosenau Hall, CB#7431, Chapel Hill, NC 27599, USA

ARTICLE INFO

Article history:
Received 24 December 2012
Received in revised form 13 July 2013
Accepted 24 July 2013

Keywords:
Water and sanitation
Cluster analysis
Country groupings

ABSTRACT

The fields of global health and international development commonly cluster countries by geography and income to target resources and describe progress. For any given sector of interest, a range of relevant indicators can serve as a more appropriate basis for classification. We create a new typology of country clusters specific to the water and sanitation (WatSan) sector based on similarities across multiple WatSan-related indicators. After a literature review and consultation with experts in the WatSan sector, nine indicators were selected. Indicator selection was based on relevance to and suggested influence on national water and sanitation service delivery, and to maximize data availability across as many countries as possible. A hierarchical clustering method and a gap statistic analysis were used to group countries into a natural number of relevant clusters. Two stages of clustering resulted in five clusters, representing 156 countries or 6.75 billion people. The five clusters were not well explained by income or geography, and were distinct from existing country clusters used in international development. Analysis of these five clusters revealed that they were more compact and well separated than United Nations and World Bank country clusters. This analysis and resulting country typology suggest that previous geography- or income-based country groupings can be improved upon for applications in the WatSan sector by utilizing globally available WatSan-related indicators. Potential applications include guiding and discussing research, informing policy, improving resource targeting, describing sector progress, and identifying critical knowledge gaps in the WatSan sector.

Introduction

Inadequate access to water and sanitation are major causes of millions of deaths and billions of cases of diarrheal and respiratory morbidity each year. This has direct costs to households and health systems, contributing to the poverty trap for billions of people worldwide (Pruss-Ustun and Corvalan, 2006; Waddington et al., 2009). There is, therefore, a concerted effort to address the lack of basic water and sanitation (WatSan) services worldwide, including initiatives by the World Bank, the UN, and bilateral agencies (About the Water and Sanitation Program, 2013; United Nations, 2011). Despite these initiatives, there remain 783 million people without access to improved water sources and 2.5 billion people without access to improved sanitation, according to the Joint Monitoring Program (UNICEF and WHO, 2012). Some have suggested that causes of this situation include poorly invested resources, misaligned country investments (Saleth and Dinar, 2000), and other donor failings (Radele and Levine, 2010).

In the WatSan sector, countries are often clustered in order to inform efforts to address the lack of access to WatSan services. Specifically, country clusters are used to explain trends, target investments, and generalize findings. Existing approaches in the WatSan sector for clustering countries have focused on geographic-, economic-, or health-based indicators. Geography-based country clusters group countries based on proximity to each other, while economic- and health-based clusters group countries based on similarity across indicators of country-wide economic performance or health, respectively. The Millennium Development Goals country grouping is geography-based (excepting developed countries, which are put in a separate group) (UNICEF and WHO, 2012). The WHO country grouping is also geography-based, with sub-groupings based on patterns of child and adult mortality (Definition of region groupings, 2013). The World Bank has three systems for grouping countries: geographic region, income (gross national income per capita), and lending category (determined based on income group and credit rating) (How we classify countries, 2013).

* Corresponding author.
E-mail addresses: konda@live.unc.edu (K. Onda), jonny.crocker@unc.edu (J. Crocker), gkayser@unc.edu (G.L. Kayser), jbartram@unc.edu (J. Bartram).

1438-4639/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ijheh.2013.07.017

The WatSan sector is complex: institutional arrangements, levels of service, and program and project experiences do not align by geography, economics, or health alone. The WatSan sector is also impacted by and connected to political, economic, social, and environmental factors. This paper presents a cluster analysis which classifies countries into groups designed to be representative of WatSan sector arrangements, performance, and progress.

Background

Previous applications of cluster analysis

Cluster analysis denotes a family of methods in applied statistics used to identify groups in data, where the groups are formed so as to be as homogenous as possible, while maximizing the differences between groups (Kaufman and Rousseeuw, 1990). Cluster analysis has been applied across a wide range of disciplines, including chemistry, genetics, and marketing to create interpretable classifications using multiple variables (Downs and Barnard, 2003; Eisen et al., 1998; Zandpour and Harich, 1996). Cluster analysis has been used to group countries along national-level indicators in political economy and international business contexts (Ketchen and Shook, 1996; Wolfson et al., 2004). In the international development sector, previous work has shown that applying cluster analysis to countries over a broad set of relevant economic indicators can yield more informative country classifications than can traditional geographic- and income-based demarcations (Berlage and Terweduwe, 1988). Cluster analysis has also been used to group countries across multiple dimensions of environmental sustainability (Esty, 2002).

Major approaches in cluster analysis

There are many methods of cluster analysis, of which Everitt et al. (2001) offers an overview. Major clustering approaches include hierarchical, optimization, and model-based methods. Hierarchical methods connect data points based on a measure of distance between the data points to form clusters. Such methods produce a hierarchy of clusters that can be judged to merge together as a distance threshold increases, and can be expressed visually as a dendrogram. Optimization methods produce a partition of the data into a number of groups k that must be pre-specified by the analyst, by choosing k data points as pre-assigned "cluster centers," and then assigning data points to those centers in a way that minimizes the squared distances between members within that cluster. Model-based methods generally use an expectation-maximization algorithm that assigns data points to a fixed number of Gaussian distributions.

An important limitation to any of these methods is that there is no internal mechanism to distinguish between important and unimportant indicators. As such, the resulting clusters are sensitive to the indicators included in the analysis; therefore, indicators must be chosen carefully based on conceptual underpinnings highly dependent on sector context.

Methods

Data sources

After a literature review and consultation with experts (academics and practitioners) in the WatSan sector, we chose indicators with which to cluster countries based on their relevance and suggested influence on national water and sanitation service delivery, as well as data availability so as to capture a majority of the world's population in the country groupings.[1] Suggested influence here means that the indicator has a hypothesized mechanism by which it influences the level of water and sanitation service delivery, or has been associated with levels of service delivery in previous studies. Previously cited influences on access to water and sanitation service and the quality of that service provided include: investment, aid, governance, education, human capital, inequality and water availability (Fry, 2008). Table 1 presents the indicators we chose for the cluster analysis and the rationale for their inclusion. Seven of the indicators were chosen for their influence on the WatSan sector's capacity and arrangements and two indicators were chosen to represent levels of WatSan service delivery.

There are many other indicators for which data is available and also may have influence on national water and sanitation service delivery. Such indicators were excluded either because of missing data for many countries and a relatively large proportion of the global population, or to avoid co-linearity which would distort the results of the cluster algorithm. Examples of indicators excluded due to data availability are domestic water and sanitation infrastructure investment and bacteriological water quality. National figures for domestic public and private investment in water and sanitation infrastructure are available from a limited number of national government expenditure reports, as well as from 26 public expenditure reviews conducted by the World Bank (Manghee and van der Berg, 2012). Nationally representative drinking water quality data is only available for the five countries covered by the Rapid Assessment of Drinking Water Quality project of WHO and UNICEF (Onda et al., 2012). We avoided constructed indices such as the human development index (HDI) and water poverty index (WPI) in order to focus on the underlying data and to avoid unnecessary colinearity among these indices and chosen indicators, many of which are constituents of such indices (Anand and Sen, 1994; Sullivan, 2002). We did not include the health indicator under-five diarrheal incidence because the available figures were produced using a regression that included WHO region and per capita national income as covariates (Walker et al., 2012). We included years of education completed over other education indicators because we deemed it to be relevant for a range of WatSan arrangements and outcomes.

Clustering method

A cluster analysis was conducted to classify countries into groups based on similarity across the nine selected WatSan indicators. We used a hierarchical clustering method to allow for the natural number of relevant clusters to emerge, since optimization- or model-based approaches involve pre-determining the number or shape of clusters. All computations were performed with R version 2.15.1 (R Core Team, 2012).

We elected to use the squared Euclidean distance as the distance metric, which is the square of the simple geometric distance in multidimensional space (Kaufman and Rousseeuw, 1990). This requires that all indicator variables be on the same scale in order for each variable to be given equal weighting in the overall distance calculation.

We elected to use the raw variables rather than to use a data reduction technique such as principal components analysis because the use of such synthetic variables in cluster analysis is generally considered poor practice (Kettenring, 2006). Highly right-skewed variables were transformed by taking the natural logarithm. Highly

[1] The consulted experts included: faculty in Environmental Engineering at the University of North Carolina, the University of Cape Town, Bristol University, senior water and sanitation experts at WHO and a Political Economy research consultant at the World Bank.