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

Article *in* International journal of hygiene and environmental health · August 2013

DOI: 10.1016/j.ijheh.2013.07.017 · Source: PubMed

4 authors:

Kyle Onda
University of North Carolina at Chapel Hill
8 PUBLICATIONS **462** CITATIONS
SEE PROFILE

Jonny Crocker
University of Washington Seattle
39 PUBLICATIONS **139** CITATIONS
SEE PROFILE

Georgia Lyn Kayser
University of California, San Diego
15 PUBLICATIONS **163** CITATIONS
SEE PROFILE

Jamie Bartram
University of North Carolina at Chapel Hill
403 PUBLICATIONS **11,451** CITATIONS
SEE PROFILE

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 co-linearity 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.

Table 1
Description of indicators included in cluster analysis and data sources.

Indicator	Indicator description	Indicator rationale with sources	Source and date
GDP per capita	Gross Domestic Product (GDP) per capita reflects the general amount of resources available for investment, and specifically for water and sanitation infrastructure and associated programs.	Research reveals that increased per capita income results in improved access to water and sanitation services (Rudra, 2011; Shafik, 1994; Wagstaff, 2002).	World Development Indicators (2013)
Water and Sanitation ODA (5-year average per capital)	Official Development Assistance (ODA) reflects a large proportion of foreign capitalthat is invested in water and sanitation infrastructure and programs.	While access to development assistance may not be sufficient on its own, it can increase investmentand influence the direction ofwater and sanitation programs (Clemens et al., 2007; Gomanee et al., 2005).	Query Wizard for International Development Statistics (2011)
Gini-index	The Gini index reflects inequality of resource distribution across a society.	Research suggests income inequality influences access to water and sanitation, especially for the poorest in the most unequal countries (Monteiro et al., 2010; Rudra, 2011).	World Development Indicators (2013)
Governance Effectiveness (GE)	Government Effectiveness (GE) reflects government commitment and effectiveness in implementing programs.	Studies have pointed to the influence of effective governance in the provision of WatSan services (Bakker et al., 2008; Meeting the Water Governance Challenge, 2012; Nunan and Satterthwaite, 2001; Rogers and Hall, 2003).	Worldwide Governance Indicators (2011)
Expected Years of Education per capita	Expected years of education reflects household knowledge of WatSan as well as the availability of qualified professionals for the design, operation, and maintenance of WatSan infrastructure and programs.	Numerous studies show that education is a key determinant of WatSan-related household decision making (Jalan et al., 2009; Rogers et al., 2007).	Education for All Global Monitoring Report (2011)
Renewable Fresh Water Availability per capita	Renewable freshwater per capita reflects the availability of water for WatSan services.	Environmental water supply can assist or constrain countries as they try to expand access to water and sanitation services (Fry, 2008; Saleth and Dinar, 2000).	World Development Indicators (2013)
% Urban	The percent of the population that is urban reflects the portion of the WatSan sector that has historically had greater access to large WatSan infrastructure projects as compared to small community-managed systems, and household technologies.	The WHO/UNICEF Joint Monitoring Program consistently shows that water and sanitation access patterns and progress vary between rural and urban settings (UNICEF and WHO, 2012; Wolf, 2009).	World Urbanization Prospects, the 2011 Revision (2011)
Improved water source	The percent of the population that has access to improved water sources (private tap, public tap, borehole, protected dug well, protected spring, rainwater).	The percent of the population that has access to improved water sources reflects the performance of the drinking water sector (Pruss-Ustun and Corvalan, 2006; Saleth and Dinar, 2000; UNICEF and WHO, 2012; Waddington et al., 2009).	Data resources and estimates of the WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation (2012)
Improved sanitation source	The percent of the population that has access to improved sanitation (flush toiler, piped sewer, septic tank, VIP latrine, Pit latrine with slab, composting toilet).	The percent of the population that has access to improved sanitation reflects the performance of the sanitation sector (Pruss-Ustun and Corvalan, 2006; UNICEF and WHO, 2012; Waddington et al., 2009).	Data resources and estimates of the WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation (2012)

left-skewed variables were reflected, transformed by the natural logarithm, and reflected back to preserve the original rank-order. Then, all variables were standardized to a mean of 0 and standard deviation of 1, and the distance matrix for all country pairs was calculated. We used Ward's agglomerative clustering algorithm, which at each step combines clusters so as to minimize the resulting within-cluster variance (Kaufman and Rousseeuw, 1990).

The number of clusters into which to divide the resulting dendrogram was determined as the number of clusters corresponding to the first local maximum of the gap statistic. The gap statistic is a measure of the difference between the within-cluster dispersion (sum-of-squares around the cluster means), and the expected value of this dispersion over a null reference distribution for a given number of clusters. The magnitude of the gap statistic represents the degree to which the proposed clusters are internally homogenous and well separated. The first local maximum (subject to a confidence-interval check) of the gap statistic represents the point at which further joining of clusters would require combining clusters that are significantly separated relative to the previous combination (Tibshirani et al., 2001).

We internally validated the resulting clusters by calculating the Dunn and Davies-Bouldin Indices, which are both measures of the degree to which a clustering is compact and well separated (Davies and Bouldin, 1979; Dunn, 1973). These indices were also calculated for the World Bank, MDG, and WHO groupings to evaluate whether our analysis produced a superior clustering.

Data analysis

We included 156 countries, representing 6.75 billion people, in our dataset. We excluded 40 countries, representing 99.7 million people, from an initial dataset of 196 countries, due to missing data. The most common missing values were the Gini index, water availability, and GDP per capita; missing from 30, 25, and 16 countries, respectively. These countries and their populations are shown in Table 2. The GDP, freshwater availability, and ODA indicators were highly positively skewed, and were log-transformed accordingly. The water and sanitation access indicators were highly negatively skewed, and were reflected, log-transformed, and reflected again. We utilized the gap statistic to cluster countries. The first attempt resulted in a clustering for which the gap statistic yielded only two clusters, one with 33 countries characterized by high incomes and governance effectiveness, no aid inflow, and near-universal water and sanitation access that generally corresponded to the

Table 2
Countries with population (millions) excluded from analysis due to missing data.

Andorra	0.09	Monaco	0.04
Antigua and Barbuda	0.09	Montenegro	0.63
Aruba	0.11	Northern Mariana Islands	0.06
Bahamas	0.34	Occupied Palestinian Territory	4.04
Bahrain	1.26	Palau	0.02
Barbados	0.27	Puerto Rico	3.75
Brunei Darussalam	0.40	Saint Kitts and Nevis	0.05
Cayman Islands	0.06	Saint Lucia	0.17
DPR Korea	9.93	Saint Vincent and the Grenadines	0.11
Dominica	0.07	Samoa	0.18
Eritrea	5.25	San Marino	0.03
Fiji	0.86	Saudi Arabia	27.45
French Polynesia	0.27	Serbia	9.86
Grenada	0.10	Seychelles	0.09
Guam	0.18	Solomon Islands	0.54
Kiribati	0.10	Somalia	9.33
Kuwait	2.74	Tonga	0.10
Libra	3.99	Turks and Caicos Islands	0.04
Marshall Islands	0.05	Tuvalu	0.01
Micronesia (Federated States of)	0.11	Vanuatu	0.24
Total		99.7	

Table 3
Water and sanitation sector country clusters.

Cluster	Country members
1 (*n* = 33)	Australia, Austria, Belgium, Canada, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Japan, Latvia, Lithuania, Luxembourg, Malta, Netherlands, New Zealand, Norway, Portugal, Singapore, Slovak Republic, Slovenia, South Korea, Spain, Sweden, Switzerland, United Kingdom, United States
2 (*n* = 15)	Argentina, Belarus, Brazil, Bulgaria, Chile, Colombia, Cuba, Iran, Kazakhstan, Mexico, Oman, Russia, Ukraine, Uruguay, Venezuela
3 (*n* = 28)	Albania, Algeria, Armenia, Azerbaijan, Bosnia and Herzegovina, Costa Rica, Croatia, Dominican Republic, Egypt, El Salvador, FYR Macedonia, Georgia, Iraq, Jordan, Kyrgyzstan, Lebanon, Maldives, Mauritius, Moldova, Mongolia, Sri Lanka, Syria, Tajikistan, Tunisia, Turkey, Turkmenistan, Uzbekistan, Vietnam
4 (*n* = 24)	Belize, Bhutan, Bolivia, Botswana, China, Ecuador, Gabon, Guatemala, Guyana, Honduras, India, Indonesia, Jamaica, Namibia, Nicaragua, Panama, Paraguay, Peru, Philippines, Sao Tome and Principe, South Africa, Suriname, Thailand, Trinidad and Tobago
5 (*n* = 51)	Afghanistan, Angola, Bangladesh, Benin, Burkina Faso, Burundi, Cambodia, Cape Verde, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Democratic Republic of the Congo, Djibouti, Equatorial Guinea, Ethiopia, The Gambia, Ghana, Guinea, Guinea-Bissau, Haiti, Kenya, Laos, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Morocco, Mozambique, Myanmar, Nepal, Niger, Nigeria, Pakistan, Papua New Guinea, Rwanda, Senegal, Sierra Leone, Sudan, Swaziland, Tanzania, Timor-Leste, Togo, Uganda, Yemen, Zambia, Zimbabwe

World Bank's High-income OECD grouping, plus such wealthy small countries as Singapore and Malta. We elected to remove this group as its own cluster, and perform another cluster analysis on the remaining 124 countries in order to take advantage of the country variability that was otherwise overwhelmed by the difference between these wealthier countries and the rest of the world.

Results

A gap statistic analysis resulted first in two distinct groups, with the first being composed of countries typically considered developed or newly industrialized. The second analysis of the second group resulted in four country clusters. Greater subdivision would have resulted in clusters that were statistically similar, while greater aggregation would have resulted in highly internally heterogeneous clusters. The dendrogram resulting from this cluster analysis is shown in Fig. 1, and the corresponding gap curve yielding four clusters is shown in Fig. 2. Table 3 shows the resulting five water and sanitation country clusters. Fig. 3 shows the cluster country membership on a world map. The cluster indicator variable means are summarized in Table 4. The Dunn and Davies-Bouldin Indices for our clusters, as well as the World Bank, WHO, and MDG groupings are shown in Table 5. Our clusters have both the largest Dunn index and smallest Davies-Bouldin index, indicating that they represent a superior grouping along the chosen relevant indicators.

Cluster 1 consists of 33 developed and recently industrialized countries from North America, Europe, and the wealthier Asian countries. Cluster 1 countries are characterized by the highest GDPs per capita and GE scores, relatively low levels of inequality as measured by the Gini index, the highest expected years of education, high urbanization, and near-universal water and sanitation access. Example members are Canada, Germany, and Japan.

Cluster 2 consists of 15 geographically disparate countries. Cluster 2 countries are characterized by the second-highest GDPs per capita, second-highest GE scores, and second-highest expected years of education. They are urbanized to the same degree as Cluster 1 countries. They also have very high, though not quite universal water and sanitation access, and receive the lowest amount of WatSan ODA of the groups that receive such aid. Example members are Argentina, Russia, and Iran.

Cluster 3 consists of 28 mostly eastern European and Middle Eastern countries, as well as some Latin American and Asian

Table 5
Internal cluster validation indices.

Grouping	Dunn[a]	Davies-Bouldin[b]
Clusters	1.04	1.57
World Bank Income Groups	1.02	1.64
WHO Regions	0.84	1.84
MDG Regions	0.78	1.87

[a] Higher values indicate a better clustering.
[b] Lower values indicate a better clustering.

Table 4
Means (std. dev.) for clustering variables by country clusters.

Variable	1	2	3	4	5
GDP pc (2005 USD)	34,857 (12,400)	13,290 (4700)	7040 (3530)	7400 (5150)	2400 (4800)
GE	1.4 (0.48)	−0.12 (0.63)	−0.33 (0.54)	−0.2 (0.46)	−0.93 (0.46)
GINI (%)	32 (4.8)	40 (10.3)	38 (5.2)	50 (9)	43 (8.5)
Expected yrsedu (yrs)	16 (1.2)	14 (1.3)	12 (1.1)	12 (1.1)	8.8 (1.7)
% pop urban	78 (12)	78 (10)	54 (16)	53 (17)	36 (15)
% Improved water	99 (1)	96 (3)	92 (8)	91 (4)	69 (15)
% Improved San	99 (4.4)	91 (9)	90 (10)	65 (22)	36 (19)
Annual renewable freshwater pc	26 (89)	15 (17)	4.2 (5.7)	4.3(7.3)	10 (20)
5-year avg annual WASH ODA ($/person)	0.0001 (0.0006)	0.00018 (0.00025)	0.0083 (0.0085)	0.0043 (0.0055)	0.0042 (0.0054)

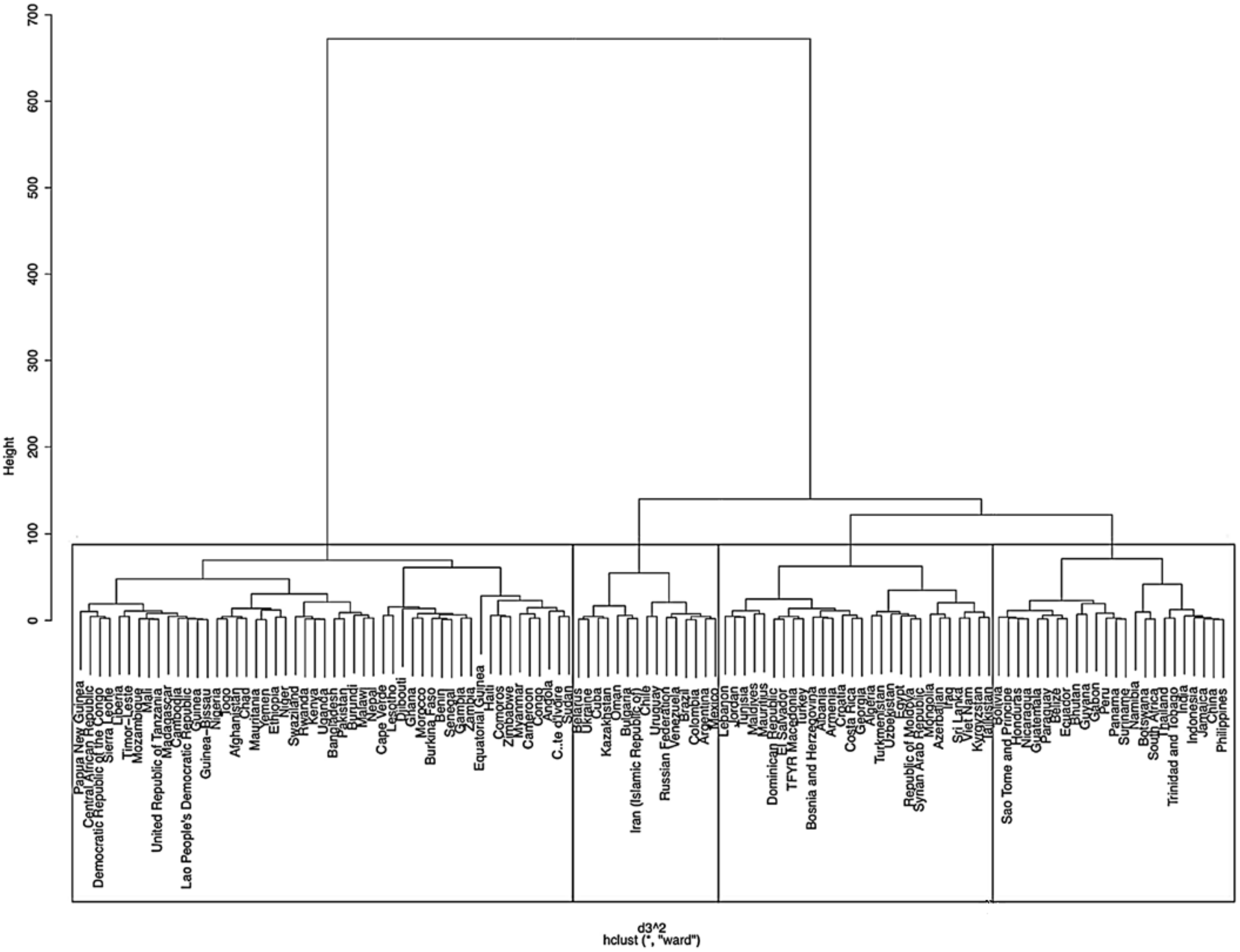

Fig. 1. Dendrogram of agglomerative clustering of countries along WASH indicators.

countries. Cluster 3 countries are middle-income, have moderate expected years of education, are roughly 50% urban, and have similar water and sanitation access to countries in Cluster 2. Cluster 3 countries stand out as having the lowest renewable freshwater resources per capita, and the highest WatSan ODA per capita of all the clusters. Example members are Algeria, Armenia, El Salvador, and Sri Lanka.

Cluster 4 consists of 24 countries mostly from Latin American and Southern African, and some Asian countries. Cluster 4 countries are characterized by similar GDP per capita, expected years of education, urbanization, and water access to those in Cluster 3, though half the ODA. Cluster 4 countries have higher income inequality, lower sanitation access, and more renewable freshwater resources than Cluster 3. Example members are Honduras, South Africa, and the Philippines.

Cluster 5 consists of 51 Sub-Saharan African and South/South-East Asian countries plus Haiti. Approximately 75% of cluster 5 countries are from Africa. Cluster 5 countries are characterized by the lowest GDP per capita, GE scores, expected years of education, urbanization, and water and sanitation access of all clusters; as well as relatively high WatSan ODA and renewable freshwater resources. Example members are Cambodia, Mozambique, and Pakistan.

The 40 countries representing 99.7 million people which were excluded for missing data are primarily small island states or fragile states. However, they should not be treated as a sixth cluster, as, despite some similarities, they were not grouped using the clustering algorithm. Example countries are the Bahamas, Serbia, and Somalia.

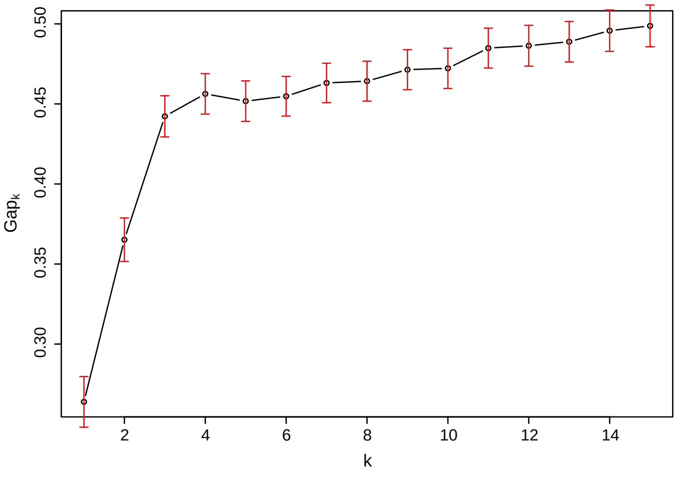

Fig. 2. Gap statistic by number of clusters (k) for country-level WASH indicator data.

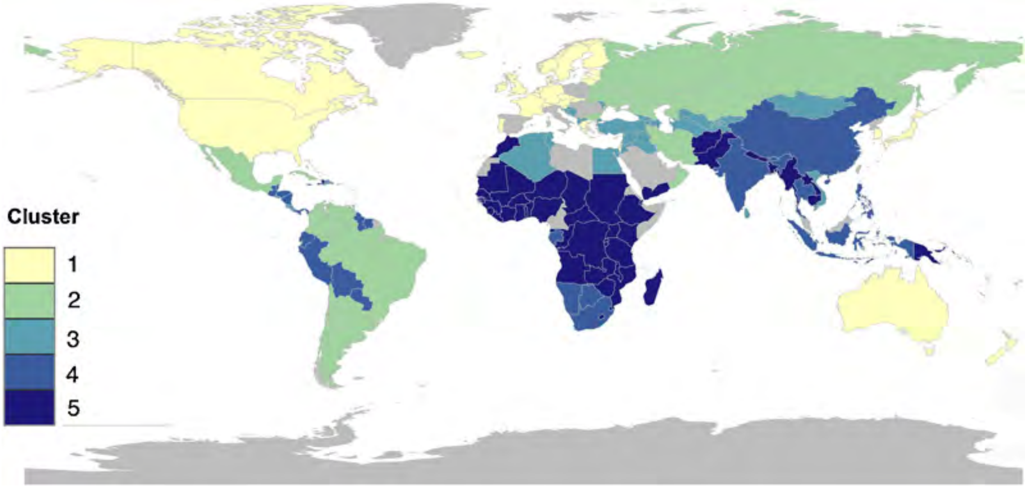

Fig. 3. Geographic distribution of WASH country clusters.

Discussion

Our novel application of cluster analysis to the WatSan sector classifies countries into groups designed to be representative of WatSan sector arrangements, performance, and progress. The Dunn and Davies-Bouldin cluster indices shown in Table 5 indicate that these comparative country clusters are more compact and well-separated across nine WatSan-related variables indicators than are the geography based MDG and WHO country clusters, and the World Bank income based country clusters. These WatSan clusters demonstrate that, for example, El Salvador, Georgia, and Egypt (Cluster 3); and Afghanistan, Laos, and Uganda (Cluster 5) are more similar along the WatSan indicators selected than Cambodia, Vietnam, and Thailand (each in a different cluster, though in the same geographic region) or South Africa, Algeria, and Chad (also in different clusters though within the same region). MDG, WHO, World Bank, and other existing country groupings have been used by others for three major applications which could be improved upon in the WatSan sector by using our country clustering model: (1) research, (2) investments and policy, and (3) discussing and communicating progress.

Use of existing country groupings in research has included stratified sampling of countries for multi-country studies, discussing findings (Fry, 2008), and extrapolating study outcomes across countries (Hutton, 2012; Walker et al., 2012). Clustering countries along critical WatSan indicators could improve these specific research applications in the WatSan sector. For example, our country clusters could be useful for stratified sampling. When the goal of stratified sampling is to better characterize or account for the variability or breadth of WatSan practices and outcomes, rather than selecting a country from each region, one could select a country from each of our clusters, which would have a higher likelihood of representing a range of WatSan practices. The clusters generated for this paper could also be used to extrapolate data sets to countries with data gaps. Many studies in water quality, water consumption, and water system sustainability exist for small sets of countries. These small data sets are not always generalizable across existing country groupings. For datasets inclusive of enough countries to show low variance within clusters, data could be extrapolated across our country groupings.

Geographic country groupings have been used to target concessional lending (The World Bank, n.d.), and to guide investment and development strategies (Sachs, 2005). WatSan specific needs do not inherently correlate with geography, but logically relate to the nine indicators used to cluster countries in this study. WatSan-specific investments and strategies may be better aligned with need if the country clusters presented in this table were used.

The most commonly referenced WatSan global database is the JMP data on water and sanitation access. The 2012 JMP report summarizes water and sanitation progress using the MDG regional groupings: Oceania, Southern Asia, Eastern Asia, South-Eastern Asia, Caucasus and Central Asia, Western Asia, Sub-Saharan Africa, Northern Africa, Latin America and the Caribbean, and Developed Regions. These eleven regions are frequently used to discuss and generate debate on levels of water and sanitation access and progress trends (UNICEF and WHO, 2012; Pruss-Ustun and Corvalan, 2006; United Nations, 2011). Our cluster analysis shows that some members from these regional groups are distinct from their geographic neighbors. Our country clustering could be used alongside geographic groupings to compare water and sanitation progress and generate debate on progress and trends in the WatSan sector.

There are several limitations to this study. The proposed clusters should not be thought of as definitive, as they are highly dependent on the variables indicators chosen, and indeed, the years of the data included. Changes in the figures from year to year would necessarily change cluster membership. It would be expected that countries undergoing significant changes along the chosen indicators would change cluster membership. Changes in variable indicator choice, chosen distance metric, clustering algorithm, or cluster number decision metric would all change the results.

Further research on the WatSan country cluster typology could involve cluster validation, or creating targeted clusters for specific regional or sub-national applications. Cluster validation could be done by selecting a non-JMP WatSan indicator to check for similarity within clusters and difference between clusters if and when they become available. Application of cluster analysis where policies and strategies demand a regional or sub-national focus could be done by applying the clustering approach to the member states of relevant regional political groupings such as the African Ministers' Council on Water (AMCOW). Creating clusters within groups such as AMCOW would be superior for applications such as organizing regional conference deliberations, or building human resource capacity.

Conclusion

Clustering countries along relevant WatSan indicators produces a comparative country typology that is more relevant to the WatSan sector than country groupings that are currently used throughout the sector. WatSan indicators included in our cluster analysis were based on their influence on national water and sanitation service delivery. These WatSan clusters push beyond geographic country groupings to suggest that country-level WatSan outcomes can be accounted for and compared across factors that are more relevant than geography or income alone, the two most commonly used methods of stratification in the international development field. This WatSan country typology could help the sector to better propose policies, track progress, target funds, and plan research in ways that account for country similarities and differences across the major outcome drivers in the sector.

Acknowledgements

Funding for this study was provided by the National Institute of Health, Water Wisdom (Bentley, PI), the Aquatest Research Consortium (University of Bristol, UK), and the International Association of Plumbing and Mechanical Officials (IAPMO). The conclusions reported are those of the authors and do not necessarily reflect the views of the funding organizations.

References

About the Water and Sanitation Program [WWW Document], WSP. 2013. http://www.wsp.org/wsp/about

Anand, S., Sen, A.K., 1994. Human development index: methodology and measurement. Human Development Occasional Papers 19922007 370.

Bakker, K., Kooy, M., Shofiani, N.E., Martijn, E.-J., 2008. Governance failure: rethinking the institutional dimensions of urban water supply to poor households. World Dev. 36, 1891–1915.

Berlage, L., Terweduwe, D., 1988. The classification of countries by cluster and by factor analysis. World Dev. 16, 1527–1545.

Clemens, M., Kenny, C., Moss, T., 2007. The trouble with the MDGs: confronting expectations of aid and development success. World Dev. 35, 735–751.

Data resources and estimates of the WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation [WWW Document], 2012. http://www.wssinfo.org/data-estimates/introduction

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 1, 224–227.

Definition of region groupings [WWW Document], WHO, 2013. http://www.who.int/healthinfo/global_burden_disease/definition_regions/en/index.html

Downs, G.M., Barnard, J.M., 2003. Clustering methods and their uses in computational chemistry. Rev. Comput. Chem. 18, 1–40.

Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Cybern. Syst. 3, 32–57.

Education for All Global Monitoring Report, 2011. UNESCO, Paris.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Nat. Acad. Sci. 95, 14863–14868.

Esty, D., 2002. Environmental Sustainability Index. Yale Center for Environmental Law and Policy, New Haven.

Everitt, B.S., Landau, S., Leese, M., 2001. Cluster Analysis. Arnold, London.

Fry, L.M., 2008. Water and non-water related challenges of achieving global sanitation coverage. Environ. Sci. Technol. 42, 4298–4304.

Gomanee, K., Morrissey, O., Mosley, P., Verschoor, A., 2005. Aid, government expenditure, and aggregate welfare. World Dev. 33, 355–370.

How we classify countries [WWW Document], World Bank 2013. http://data.worldbank.org/about/country-classifications

Hutton, G., 2012. Global Costs and Benefits of Drinking-Water Supply and Sanitation Interventions to Reach the MDG Target and Universal Coverage [WWW Document].

Jalan, J., Somanathan, E., Chaudhuri, S., 2009. Awareness and the demand for environmental quality: survey evidence on drinking water in urban India. Environ. Dev. Econ. 14, 1–28.

Kaufman, L., Rousseeuw, P., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Ketchen, D., Shook, C., 1996. The application of cluster analysis in strategic management research: an analysis and critique. Strateg. Manage. J. 17, 441–458.

Kettenring, J.R., 2006. The practise of cluster analysis. J. Classif. 23, 3–30.

Manghee, S., van der Berg, C., 2012. Guidance Note: Public Expenditure Review from the Perspective of the Water and Sanitation Sector. World Bank, http://water.worldbank.org/sites/water.worldbank.org/files/publication/Water-Guidance-Note-PER.pdf

Meeting the Water Governance Challenge In: Meeting the Reform Challenge. OECD Publishing.

Monteiro, C.A., Benicio, M.H.D., Conde, W.L., Konno, S., Lovadino, A.L., Barros, A.J.D., Victora, C.G., 2010. Narrowing socioeconomic inequality in child stunting: the Brazilian experience, 1974–2007. Bull. World Health Organ. 88, 305–311.

Nunan, F., Satterthwaite, D., 2001. The influence of governance on the provision of urban environmental infrastructure and services for low-income groups. Int. Plann. Stud. 6, 409–426.

Onda, K., LoBuglio, J., Bartram, J., 2012. Global access to safe water: accounting for water quality and the resulting impact on MDG progress. Int. J. Env. Res. Public Health 9, 880–894.

Pruss-Ustun, A., Corvalan, C., 2006. Preventing Disease Through Healthy Environments. Towards an Estimate of the Environmental Burden of Disease. World Health Organization, Geneva.

Query Wizard for International Development Statistics [WWW Document] 2011. http://stats.oecd.org/qwids/ (accessed 06.01.12).

R Core Team, 2012. R: A Language and Environment for Statistical Computing.

Radele, S., Levine, R., 2010. Can we build a better mousetrap? Three new institutions designed to improve aid effectiveness. In: Easterly, W. (Ed.), Reinventing Foreign Aid. EBSCO, pp. 431–460.

Rogers, A.F., et al., 2007. Characteristics of latrine promotion participants and non-participants; inspection of latrines; and perceptions of household latrines in Northern Ghana. Trop. Med. Int. Health 12, 772–782.

Rogers, P., Hall, A.W., 2003. Effective Water Governance (TEC Background Paper). Global Water Partnership, Stockholm.

Rudra, N., 2011. Openness and the politics of potable water. Comp. Pol. Stud. 44, 771–803.

Sachs, J.D., 2005. Investing in Development: A Practical Plan to Achieve the Millenium Development Goals [WWW Document]. United Nations Millenium Project.

Saleth, R.M., Dinar, A., 2000. Institutional changes in global water sector: trends. Water Policy 2, 175–199.

Shafik, N., 1994. Economic development and environmental quality: an econometric analysis. Oxford Econ. Pap. 46, 757–773.

Sullivan, C., 2002. Calculating a water poverty index. World Dev. 30, 1195–1210.

The World Bank, n.d. A History of Operational Guidelines [WWW Document]. http://data.worldbank.org/about/country-classifications/a-short-history

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. – Series B: Stat. Methodol. 63, 411–423.

UNICEF, WHO, 2012. Progress on Drinking Water and Sanitation. 2012 Update.

United Nations, 2011. Goal 7: Ensure Environmental Sustainability [WWW Document], http://www.un.org/millenniumgoals/environ.shtml (accessed 11.30.11).

Waddington, H., Snilstveit, B., White, H., Fewtrell, L., 2009. Water, sanitation, and hygiene interventions to combat childhood diarrhoea in developing countries. International Initiative for Impact Evaluation, Delhi, pp. 119.

Wagstaff, A., 2002. Poverty and health sector inequalities. Bull. World Health Organ. 80, 97–105.

Walker, C.L.F., Aryee, M.J., Boschi-Pinto, C., Black, R.E., 2012. Estimating diarrhea mortality among young children in low and middle income countries. PLoS ONE 7, e29151.

Wolf, S., 2009. Water and sanitation for all? Rural versus urban provision. Int. J. Serv. Econ. Manage. 1, 358–370.

Wolfson, M., Madjd-Sadjadi, Z., James, P., 2004. Identifying national types: a cluster analysis for politics, economics, and conflict. J. Peace Res. 41, 607–623.

World Development Indicators [WWW Document] 2013. http://data.worldbank.org/data-catalog/world-development-indicators (accessed 06.01.12).

World Urbanization Prospects, the 2011 Revision, 2011. United Nations, New York.

Worldwide Governance Indicators [WWW Document] 2011. http://data.worldbank.org/data-catalog/world-development-indicators

Zandpour, F., Harich, K.R., 1996. Think and feel country clusters: a new approach to international advertising. Int. J. Advert. 15, 325–344.