Anonymous Referee #2 comment

Received and published: 15 April 2012

This manuscript describes the development and preliminary analysis of a dataset combining data for multiple variables from a wide range of sources for developing countries. This is a challenging task and the authors should be recognised for attempting it. The paper is well organized. It could use some work by an editor to address consistent grammatical issues that probably derive from English as a second language. I have two general concerns about this manuscript: thematic and technical.

Thematic: This paper is organized around a technical task that assumes some future useful application, rather than a research question. If pretty much stays on this technical task. Those results that it produces simply confirm what is already known about the relationships among variables and indicators of development. As such this interest of this paper is limited. These is potential, however, to further develop some of the outcomes of this work so as to inform development efforts, to add to theory about development, to direct international development or aid efforts at the macro scale, etc.. For example the categorization of counties briefly explored in this manuscript could be further developed to further inform models such as the demographic transition, the mobility revolution, etc..

Technical:

1 Principle components analysis is related to factor analysis (it is basically factor ´analysis with commonalities equal to 0). Why do both?

2 For a PCA to be considered useful it should explain around 70% or more of the ´variance in the data set. For subset of African countries only about 50% of variance was explained with 3 components. This is rather low. Were there other components that explained enough variance to include? How were the number of components to include determined (eigenvalues alone? Scree test?).

3 Nations are not consistent in such things as definitions of key concepts and terms, ´methods of data collection, etc. For example what is considered urban in some countries would be considered rural in others. This will be a profound problem in using measures that are not standardized across countries. How was this issue addressed in this work?

4 I don't see how BOD could be used in this dataset. BOD is not reported as a ´aggregate variable at a country scale, and if it was it would be meaningless. BOD will vary within and across water bodies and streams. Perhaps there is an explanation that is missing from this paper. Otherwise this should be removed from the analysis.

5 If I read this correctly some of the variables used in the analysis are aggregates (e.g., indices) of other variables that are also used in this analysis. This is like using the same variable, same measurements in the same analysis, leading to a problem of multicollinearity. On or the other should be removed.

6 Why was hierarchical clustering chosen? Is it better for this application that some- ´thing like K-means clustering?

7 What is meant by "coherency" and "robustness" of the data is not well enough ´explained.

Comment 1

It could use some work by an editor to address consistent grammatical issues that probably derive from English as a second language

Answer 1 : ok

Comment 2 This paper is organized around technical task that assumes some future useful application, rather than a research question […]Those results that it produces simply confirm what is already known about the relationships among variables and indicators of development.

Answer 2:

In order to answer and monitor the objectives of the Millennium Development Goals (MDGs), the international community and the states decided to develop several monitoring processes. For this purpose, several international institutions in charge of monitoring the Access to Water Supply and Sanitation (WSS) developed two indicators. Although the WSS indicators are proving that the access to water and sanitation are improving, today it is difficult to say what are the main factors involved in this improvement and what are the relationships between the different factors. This is the main objective of this research.

The work proposed in this paper is one of the first steps to answer these questions involving the MDG targets. An important research effort has been done to structure and analyse all the variables that could have a direct and/or an indirect influence on WSS indicators.

To build the dataset was a critical part of this research as it's the result of the analysis of the variables and the normalisation of a huge amount of data. The data processing and methodologies used will be of interest for future researchers in the domain, not only because of future analysis of the dataset but also because the methodologies as proposed in the paper can be applied to other data.

The preliminary analyses of this dataset have allowed two different things:

1) To verify that the relationships among the variables are coherent with literature and common knowledge in the domain. This allows us to say that a strong base has been developed to go further in the analysis.

2) It also allows the creation of country profiles based on these indicators opening new interpretation possibilities in developing countries.

Based on this exploration between variables and countries, we will build a tool to create probabilistic scenarios. The latter will be the subject of the next paper as we considered that this submitted paper is coherent as such, separated from the second part.

Comment 3: The categorization of counties briefly explored in this manuscript could be further developed to further inform models such as the demographic transition, the mobility revolution, etc.

Answer 3: We focused on the database validation which calls already for quite a number of concepts and theories. Therefore, we tried to keep their description short and clear preferring to provide references for further reading (to avoid too long reading and to stick to the objective). The second foreseen article of this research is more related more with interpretation and scenarios. These concepts will be more detailed when necessary. The same reason applies to country profiles on which we did deeper analysis.

Comment 4: Principle components analysis is related to factor analysis (it is basically factor ´analysis with commonalities equal to 0). Why do both?

Answer 4 : It's true that Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA) are both variable reduction techniques but they are used (or should be) for different purposes in data analyses. In PCA, all of the observed variance is analysed, while in factor analysis it is only the shared variances that is analysed. In this paper, Principal components analysis is used to find optimal ways of combining variables into a small number of subsets, while factor analysis is used to identify the structure underlying such variables and to estimate scores to measure latent factors themselves.

Differences between PCA and FA (source: http://www2.sas.com/proceedings/sugi30/203-30.pdf):
Principal Component Analysis :

- Principal Components retained account for a maximal amount of variance of observed variables
- Analysis decomposes correlation matrix
- Ones on the diagonals of the correlation matrix
- Minimizes sum of squared perpendicular distance to the component axis
- Component scores are a linear combination of the observed variables weighted by eigenvectors

Factor Analysis :

- Factors account for common variance in the data
- Analysis decomposes adjusted correlation matrix
- Diagonals of correlation matrix adjusted with unique factors
- Estimates factors which influence responses on observed variables
- Observed variables are linear combinations of the underlying and unique factors

Comment 5: For a PCA to be considered useful it should explain around 70% or more of the ´variance in the data set. For subset of African countries only about 50% of variance was explained with 3 components. This is rather low. Were there other components that explained enough variance to include? How were the number of components to include determined (eigenvalues alone? Scree test?)

Answer 5:

70% of the variance is usually acceptable when working in well known systems. When working in developing countries, taking into account the important amount of variables, heterogeneity, missing data, … it's usually accepted to explain at least 55% of the variability. For the extended dataset (101

countries – developing countries), we increased this level up to around 64% confirming the conclusions of the subset of African countries.

To determine the number of components to be considered we looked at the eigenvalue (value above 1) first but also the factor loadings. In our case, the four first components gather the maximum loading for all variables. For the extended database, we provided the matrix for the four representative components but figures display only 3 components for readability reasons.

Comment 6: For example what is considered urban in some countries would be considered rural in others. This will be a profound problem in using measures that are not standardized across countries. How was this issue addressed in this work?

Answer 6: This issue on how to define urban areas is an old and endless question if considering a **quantitative approach**. The latter suggests that countries agreed on what is urban or not, independently of their context... This work could not reach such quantitative level of analysis but only qualitative aspects of the variable (such as a qualitative scale “low,medium ,high level” as mentioned lines 13-14 of p490). We considered the data from the World Bank and UN: they use “the national definition and proceed to corrections to keep a minimum coherency between countries”. Being international reference data, we have chosen to first include the data in the analysis and see a posteriori if it shows incoherencies.

Metadata on urban population rates:

“Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects’ In details, they do not use their own definition of "urban" population but follow the definition that is used in each country. The definitions are generally those used by national statistical offices in carrying out the latest available census. When the definition used in the latest census was not the same as in previous censuses, the data were adjusted whenever possible so as to maintain consistency. In cases where adjustments were made, that information is included in the sources listed online.”

Comment 7: I don’t see how BOD could be used in this dataset. BOD is not reported as a ´aggregate variable at a country scale, and if it was it would be meaningless. BOD will vary within and across water bodies and streams.

Answer 7 BOD is an aggregated variable (expressing organic pollutant discharge in water resources in average at country level) provided by World Bank. It could have been an indicator of the global water quality of waters and is only included for Africa. Following the same pragmatic approach as for urban population (including a priori indicators in PCA and removing it if not relevant), It has been removed from the extended database because found not suitable in characterising water quality (see lines 11-19 p498 and figure 2 p537) For information, I put the reference of the first 1998 study on BOD done by Hemamala Hettige, Muthukumara Mani, and David Wheeler, "Industrial Pollution in Economic Development: Kuznets Revisited" (available at www.worldbank.org/nipr). The data were updated by the World Bank's Development Research Group using the same methodology as the initial study.

Comment 8: If I read this correctly some of the variables used in the analysis are aggregates (e.g., indices) of other variables that are also used in this analysis. This is like using the same variable, same measurements in the same analysis, leading to a problem of multicollinearity.

Answer 8: We have excluded from our statistical analysis the composite indictors as for example Water Poverty index or ESI(mentioned line 6-12 p 496).

To avoid the multicollinearity issues, we used the single variables for the analysis but we have only re-projected of composite indicators for improving the interpretation. This is common when analysing multivariate data. It excludes in a first step the composite indices from the PCA performed on non-composite variables and re-projects, in a second step, the composite variables within the PCA avoiding bias and getting their real position within the projection.

Comment 9 Why was hierarchical clustering chosen? Is it better for this application that some- ´thing like K-means clustering?

Answer 9: Both K-mean and agglomerative clustering (tested using various distances) were performed on the dataset providing similar results however AHC was found slightly more suitable and precise that k-means regarding the objective of this clustering process. In other to avoid a long paper we decided not to mention these details.

Comment 10: What is meant by “coherency” and “robustness” of the data is not well enough ´explained.

Answer 10: OK. We meant by coherency the relation between variables and countries behaviours are conformed to scientific or field experience to go beyond the statistical validity of the various analyses (PCA- FA or OLS regression). A clear definition should be included in page 495.