*2011 2nd International Conference on Environmental Science and Development*
*IPCBEE vol.4 (2011) © (2011) IACSIT Press, Singapore*

# Inter-relationships among water, governance, human development variables in developing countries: a database coherence analysis

Céline Dondeynaz, César Carmona-Moreno
Global Environment Monitoring (GEM)
Joint Research Centre
ISPRA, Italy
e-mail: celine.dondeynaz@jrc.ec.europa.eu

Andrea Leone and Daoyi Chen
Department of Engineering
Liverpool University
LIVERPOOL, UK

***Abstract*—A deep understanding of the water sector in developing countries relies on studying complex interactions between different environmental, socio-economic, governance and other human development factors. The work aims to obtain a preliminary analysis of the interactions between these variables at the country level over Africa. This paper described how data, collected on a worldwide basis, have been processed using the Expectation Maximization (EM) algorithm, hot deck imputation methods, and normalization methods to format the datasets. Statistical analysis is performed using Principal Component Analysis to verify the coherence of the database.**

***Keywords: indicators, principal component analysis, water supply and sanitation, governance, human development, water resources, human activity pressure***

## I. INTRODUCTION

The experience of international cooperation over the last 50 years[1] indicates that improving the understanding of the inter-relations among different related factors linked with economic and human development is an essential baseline in the design of development cooperation policies and strategies at national, regional and continental levels. Regarding the water sector, the adoption of the integrated water resources management (IWRM) approach[2] has led to a shift in the main effort for providing sustainable WSS (Water and Sanitation Services) from infrastructure development to effective management of water resources. The level of efficiency and development of water and sanitation services can be considered, indeed, as the result of many factors.

Applying a cross analysis approach, this work aims to identify the key elements explaining the various levels of access to WSS. Using the standard MDG indicators[3] (the percentages of the population having access to improved water supply and sanitation), the objective is to build a methodological framework to analyze these variable behaviours and thus to map the variables impacting and/or being influenced by the WSS level.

This paper presents the methods and analyses performed to validate the dataset coherence. It also includes the description of the data sources and selection criteria. As a testing phase, we restricted our present analysis to the African countries for the year 2004. Further analysis across countries worldwide will be performed in the next phase after solving normalization issues associated with very diverse behaviours found among countries.

## II. METHODOLOGY

### A. Dataset construction

#### 1) Logical framework and indicators

The data were chosen considering all the variables that can both result in and influence (double-way relationships) the WSS levels. These variables have been clustered under four main areas or pillars:

- **Environment**: indicators on quality and quantity of water resources.
- **Human pressure**: indicators on demographic, human activity pressure.
- **Governance**: indicators measuring stability, government effectiveness, rule of law, control of corruption, democratic conditions, regulatory quality and environmental governance.
- **Human development**: indicators on social and health to measure a country's level of well-being.

In addition to these four pillars, as developing countries are the current focus, the **Official development Aid delivery** (ODA) in the water sector has been included in the database. This indicator represents the global disbursed official aid provided to the developing countries.

#### 2) The data

Indicators have been collected from official providers such as the World Bank, OECD, FAO, WHO, UN DESA, UNDP, UNSD, UN-HABITAT and research institutions such as Universities, NGO and Institutes.

The compatibility and consistency of this dataset, in terms of geographical and temporal scales, is a major constraint in the analysis process. The national scale was chosen as most of the data were supplied at this level. Data sets for 2004 were used in the present study because the last release of the Joint Monitoring Programme (JMP[4]) report on

[1] Easterly, W, 2001 [1]

[2] Principles laid down at the International Conference on Water and the Environment held in Dublin in January 1992

[3] Millennium Development Goals indicators Provided by the United nations Statistic department for monitoring the progress toward the Target 3 of the Objective 7 about Water supply and sanitation services.

[4] Joint Monitoring Programme [12]

WSS access level at the beginning of this research was also based on data collected for year 2004.

The data collection covers countries **worldwide**. 132 indicators have been examined based on the following main criteria:

[1] Relevance: the indicator has a potential role regarding the water supply or sanitation level,

[2] Data availability: The dataset has enough observations (less than 100 missing data over the world).

[3] Reliability: the indicator has been produced by trustful providers and described methods.

After this first filter, **53 variables** were finally selected and transformed into a normal distribution. Complementary normalization tests were performed to verify the statistical stability of the variables.

The errors and incoherence of the dataset (the relationships between the variables and magnitudes of the values) were tracked through their Principal Component Analysis (PCA) performance. The final database contained a list of 48 variables. The data obtained provide somewhat **raw estimates** (qualitative estimations) than exact quantitative values, mainly because of the nature of the indicators themselves and the context of developing countries.

In addition to the variables to be explained (water supply and sanitation coverage and proportion of water house connection) the database[5] also considers:

- Education, health, and well-being aspects: fertility rates, children mortality under 5 years, life expectancy at birth, health expenditure, malaria prevalence, gross enrolment at school (primary to university), percentage of children having diarrhoea, literacy rates for youth (15-24), gross domestic product per capita (GDP), female economic activity, poverty rate, Human development Index (HDI), Human Poverty Index (HPI)
- Human activities and demographic pressures: agricultural area, withdrawal by sectors (total withdrawal which could be assimilated as Agriculture demand, industrial and domestic withdrawals) agricultural production index, proportion of irrigation area, water use intensity in agriculture, urban population and both rural and urban population growths
- Environmental conditions: Environmental Sustainability Index (ESI)[2], water bodies total surface, amount of precipitations, proportion of arid lands, estimation of water resources, dam capacity, Biologic Oxygen Demand (BOD), National Biodiversity Index (NBI), Water Poverty index(WPI),
- Governance aspects: Corruption Perception Index (CPI), environmental governance, participation into international environmental agreements, Worldwide Governance Indicators (WGI) related to Voice and Accountability (W&A), Political stability and absence of violence (PS&AV), Government effectiveness (GE), Rule of Law (RofL), Regularity Quality(RQ), Governance Index for Africa (GI Afr)[4],
- Aid delivery with both global official aid (ODA) per capita and the ODA specific to WSS

*3) Missing data treatment*

The missing data (m) treatment aims to obtain realistic values for missing data rather than to have accurate values that take into account the nature of the indicators that we have collected. With the characteristics of our dataset, we applied multiple imputation methods [6] that compare country observations based on several indicators in order to impute missing data without modifying the general statistical behaviour of the variables.

The imputation method used in this study was the Expectation-Maximization Algorithm (EM)[6]. The algorithm combines the classic EM algorithm with a bootstrap approach to draw samples from a second processing stage. For each draw, the algorithm bootstraps the data to simulate estimation uncertainty and then run the EM algorithm [5] to find the mode of the posterior result for the bootstrapped data.

We completed our dataset incrementally by imputing missing data for variables with less missing data before processing more incomplete ones. To improve imputations, this process was done on the worldwide dataset starting with 170 observations (where small states were removed).

Where there were few missing values (m<5) the Hot deck imputation method [7] was used. This method compares country observations on several indicators in order to impute missing data (the median distance) according to the "nearest neighbour" rule. (Table 1)

| Country | GDP-PPP per capita[1] | Industrial Withdrawal in % | Corruption Perception Index | BOD emission per capita[2] |
|---|---|---|---|---|
| Kyrgyzstan | 1.721502 | 3 | 2.2 | 2.210746 |
| Sudan | 1.719801 | 1 | 2.2 | NULL → 1.1<x <2.2 **x = 1.65** |
| Uzbekistan | 1.712442 | 2 | 2.3 | 1.02716 |

[illegible]

TABLE 1: EXAMPLE OF HOT DECK IMPUTATION METHOD

## B. Statistical analysis methods

This preliminary analysis aims to test and validate the methodology as well as verifying the coherence and robustness of the dataset. It also provides an initial view of the possible relationships among the variables, thus indicating areas for further analysis. The dataset has been focused on Africa as it was possible to carry out data normalisation, enabling us to perform Principal Component

[5] See Online database section in the references section

[6] Imputation have been made using Amelia II software [3] provided by Honaker James, King Gary, Blackwell Matthew

Analysis. However, the objective is to extend this approach to the full dataset using the lessons learnt.

## III. DATASET VALIDATION FOR AFRICA

### A. PCA performance (Figure 1)

#### 1) PCA parameters

Composite Indicators (namely the Worldwide Governance Indicators WGI, Governance Index GI Africa, Human development Index HDI, Environmental Sustainability Index ESI, Human Poverty Index HPI, Water Poverty) have been re-projected within the PCA to avoid bias caused by the partial overlap of their sub pillars with active variables.

The cumulated variability of the first three factors is equal to **50.386%.** If we take into account the high heterogeneity and the nature of the variables, the first three principal components can be considered as suitable for a first interpretation. However, caution is needed when interpreting the maps, as some information might be hidden in subsequent factors.

#### 2) Analysis of the correlation between the variables

The general representation of the dataset on the different axes is coherent with what was expected, allowing us to consider the dataset as consistent. On the F1 axis, the positive development indicators (Group1) are negatively correlated with poverty indicators (Group 2). The F2 axis represents the indicators about water resources and water demand, especially agricultural pressure (Groups 3 and 4).

In particular, the following relationships can be deduced from the PCA (figure 1) and the correlation matrix.

##### a) Within the groups

Governance indicators (Group 1), even calculated through various methods and data sources, are coherent and very highly correlated.

The Environmental Sustainability Index is a cross-cutting and very complex index which demonstrates the capacity of a nation to manage its environment in a sustainable way. Therefore, as expected, it is highly correlated with many variables, but in particular with the water resources' availability, the level of urbanization (Domestic demand in water) and agriculture pressure (water use intensity in agriculture).

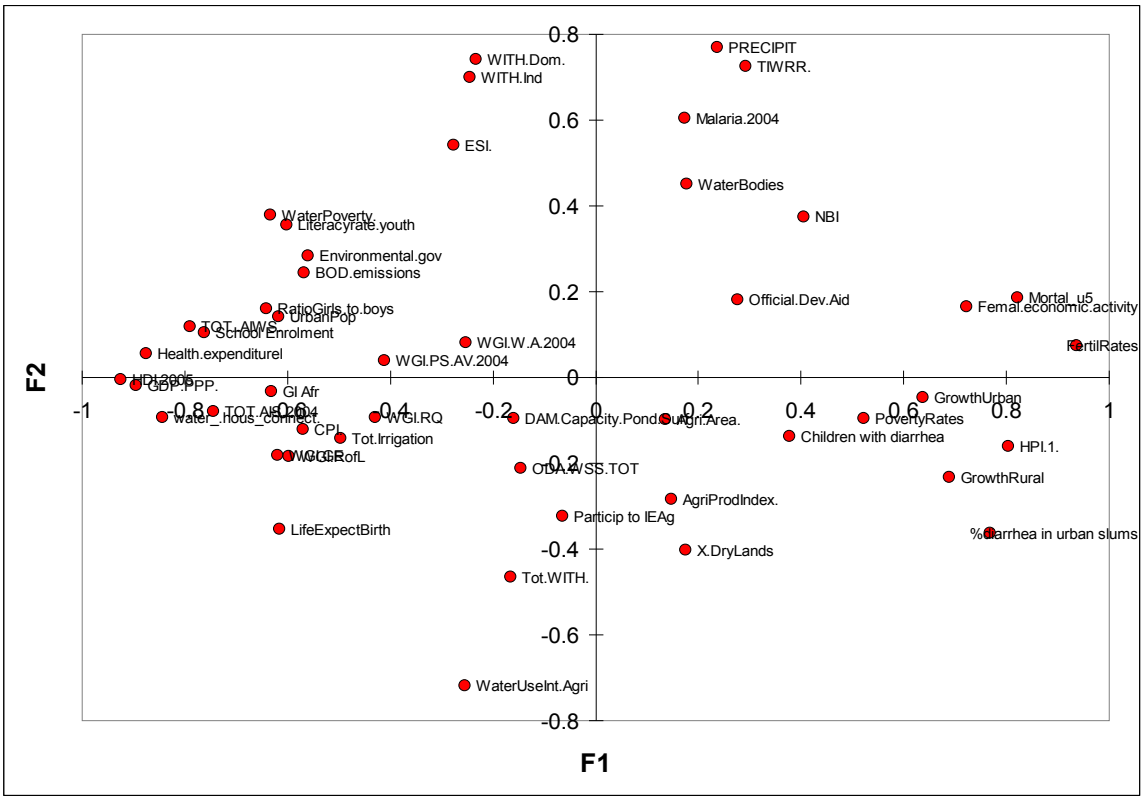


Figure 1: the first two PCA factors of variables, (accumulated variability equal to 43,02%)

The Biological Oxygen Demand (BOD), an indicator of water quality covering surface and groundwater, is correlated with the level of urbanization (the proportion of the population in urban settlements, Urban pop). This index is highly correlated with the industrial and urban

development of the country and cannot be taken as an indicative value of water quality.

*b) Relationships between groups*

Governance indicators are positively correlated with the development of the country, the income (GDP-PPP, health expenditure, education rate), WSS access are negatively correlated with child mortality and fertility rates.

The variable on participation in international environmental agreements shows a negative correlation with the water resources available (TWRR) and the amount of precipitation.

*c) Variables near the centre of the graph*

Centralised variables in PCA do not allow us to identify significant relationships but the following remarks can be made:

- Regarding the dam capacity, no significant correlations can be observed with any other variables except the agricultural area. It means that the capacity of countries to have reservoirs and dams neither impacts directly nor indirectly on the level of access to WSS.
- The agricultural surface area is correlated with National Biodiversity (0.414) in addition to the dam capacity. No conclusions can be made because of overlapping bias due to the nature of the national biodiversity index.
- The national biodiversity index correlates with variables belonging to several groups (Groups 2 and 4 and for a smaller part of Group 1) contributing with the same approximate weight.
- The Children Diarrheal Prevalence behaviour is mainly caused by an inverse correlation with the variables in the Group 1. Lower correlations are observed with Group 2.
- Financial flow (either as global aid or finances specific to the WSS) has few significant correlations and, in addition, is distributed among the three axes. However, stability and the absence of violence could be one of the criteria for Aid delivery. This needs to be verified while looking for other missing explicative variables.

## IV. Remarks and Limits

For Africa, the dataset shows coherence as the variables are correlated as expected regarding common knowledge [7] in the water sector. By employing these statistical methods (Multiple Imputation methods and PCA) we have shown that it is possible to readily draw comparisons between many water and development indices. However, several points should be made regarding the results:

Firstly, the data are reliable if considered as **qualitative estimates**.

Secondly, concerning the interpretation of the results, relationships between variables through the PCA are clearly depicted but more analysis and investigations are needed with complementary fieldwork to improve and validate interpretations.

The next step will involve modelling the dataset to extract the essential key variables influencing, or being influenced by, the level of WSS in a given country. The number of variables considered should be clustered or reduced, too, to obtain key and relevant indicators that will facilitate the interpretation of the results.

[7] "Common knowledge" refers to expertise, documents that have been consulted to verify the variables behaviors observed within the PCA

## References


[1] W.Easterly, "The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics."The MIT Press, 2001.

[2] D.C Esty, M. Levy, T. Srebotnjak, A. de Sherbinin, "Environmental Sustainability Index: Benchmarking National Environmental Stewardship". NewHaven: Yale Center for Environmental Law & Policy, 2005. http://www.yale.edu/esi/

[3] J. Honaker, G. King, M. Blackwell, "Amelia II: A Program for Missing Data," Version 1.2-17. http://gking.harvard.edu/amelia/

[4] R. I. Rotberg, R. M Gisselquist "Strengthening African Governance: Index of African Governance, Results and Rankings" Cambridge, MA, 2009. http://www.worldpeacefoundation.org/africangovernance.html

Article in a journal

[5] J. Honaker, G. King, "What to Do about Missing Values in Time-Series Cross-Section Data," American Journal of Political Science, Vol. 54, No. 2, April 2010, pp. 561–581. http://gking.harvard.edu/files/pr.pdf

[6] N. J Horton, S. R Lipsitz," Multiple Imputation in practice: Comparison of Software Packages for regression models with missing variables." The American Statistician, 55(3), August 2001, pp 244-254.

[7] M. Reilly "Data analysis using hot deck multiple imputation" The Statistician, 42, 1993, 307-313.

Online databases

[8] AQUASTAT database. a global information system on water and agriculture http://www.fao.org/nr/water/aquastat/main/index.stm

[9] CPI database from Transparency International http://www.transparency.org/policy_research/surveys_indices/cpi/2004

[10] Earth trends, the environmental information portal,http://earthtrends.wri.org/searchable_db/index.php?action=select_theme&theme=

[11] GEO Portal, Datasets used by UNEP, http://geodata.grid.unep.ch/#

[12] Joint Monitoring Programme (JMP) Portal http://www.wssinfo.org/datamining/tables.html

[13] OECD database http://stats.oecd.org/qwids/

[14] World bank database http://databank.worldbank.org/ddp/home.do#ranking