

# Toward a Methodology for Computing a Progress Composite MDG Index

---

Maurice Mubila<sup>1</sup> and Achille Pegoue<sup>2</sup>

## **Abstract**

*The proposed methodology for computing a progress composite MDG index (P-CMI) should be considered as a work in progress. It is still undergoing peer review and will likely incorporate further revisions. The final methodological proposal is scheduled for publication some time in 2009. The methodology involves the following steps: selection of countries if the percentage of indicators with a missing value is 25% or below; selection of indicator if the percentage of total population of countries without missing data is 75% or above; multiple imputations for missing data; setting of two categories for data normalization; an equal weighting scheme; and linear additive rule for aggregation. In addition, the aggregation procedure follows the hierarchical MDGs structure. Thus, an aggregate index is computed for each target, each goal, and for the whole set of MDGs to obtain the P-CMI. The P-CMI is considered as the probability of achieving the MDGs and it ranges between 0 and 1. A P-CMI of around 1 indicates that a country or group of countries is likely to achieve the MDGs by 2015. A P-CMI of around 0 indicates that a country or group of countries is not making progress in achieving the MDGs by 2015. The sensitivity of the P-CMI is carried out by the computation of Sobol's indices for five factors of uncertainty identified and M-estimators for the overall P-CMI.*

**Key Words:** Composite indicator, Millennium Development Goals (MDGs), Normalization, Aggregation, Sensitivity analysis

## **Résumé**

*La méthodologie proposée pour le calcul d'un indicateur synthétique de l'avancée des pays africains dans l'atteinte des Objectifs du Développement pour le Millénaire (ODM) doit être considérée comme en cours de développement. Son processus de validation par les différents partenaires se poursuit, pouvant conduire à une future version révisée. La proposition d'une version définitive de la méthodologie est prévue pour l'année 2009. La présente méthodologie est assise sur: sélection du pays si l'ensemble des indicateurs présente moins de 25% de données manquantes, sélection d'un indicateur si la population des pays sans données manquantes représente plus de 75% de la population de l'Afrique, l'imputation multiple pour l'estimation des données manquantes, la définition*

---

<sup>1</sup> Chief Statistician, Economic and Social Statistics Division, Statistics Department, AfDB, Tunis. E-mail: m.mubila@afdb.org.

<sup>2</sup> International Consultant, Yaoundé, Cameroon. E-mail: apegoue@yahoo.com.

*de deux catégories pour la normalisation; l'équi-pondération et l'agrégation complète additive. De plus, la structure hiérarchique des ODM est respectée dans la procédure d'agrégation. Ainsi, un indicateur synthétique est calculé pour chaque cible, chaque objectif et l'ensemble des ODM afin d'obtenir l'indicateur synthétique global. Cet indicateur synthétique, compris entre 0 et 1, est considéré comme la probabilité d'atteindre les ODM. Une valeur proche de 1 indique que le pays ou le groupe de pays réalise de substantiels progrès pour l'atteinte des ODM alors qu'une valeur proche de 0 indique que les progrès sont insuffisants. La sensibilité de la méthodologie de calcul de l'indicateur synthétique est réalisée à l'aide des indices de Sobol et des M-estimateurs pour cinq facteurs d'incertitude identifiés.*

**Mots clés :** *Indicateur synthétique, Objectif du Développement pour le Millénaire (ODM), normalisation, agrégation, analyse de sensibilité*

## 1. INTRODUCTION

Most of the assessments that monitor the progress countries are making toward the attainment of the MDGs are based on the indicator tracking technique. Ideally, however, the goals are interlinked, as each of the first seven goals addresses an aspect of poverty. As such, it is safe to argue that they should be viewed together because they are mutually reinforcing. This paper therefore presents a summary of a methodology to compute a Composite MDG Index (CMI). The CMI in this case is a composite indicator based on the aggregation of MDG indicators. The indicator can be used to assess a country or group of countries in terms of the progress being made in attaining the MDGs, hence it is referred to as a Progress CMI (P-CMI). The P-CMI should be seen as a way of interpreting how a set of MDG indicators are evolving for a country with respect to other countries, in terms of achieving the MDGs by 2015 if the observed trend were to continue. It shows “at a glance” whether or not a country or group of countries will achieve the entire set of MDGs by 2015.

The methodology for computing the P-CMI is primarily based on the guidelines for constructing composite indices developed by the Organization for Economic Cooperation and Development (OECD). The development of the methodology has also benefited from other work carried out in this area (e.g. Rouzier 2003; ESCAP et al. 2005, 2006, 2007).

There are five main steps involved in the methodology to compute a P-CMI, namely: selection of countries and indicators; computation of the

expected years of achieving the targets; imputation of missing data; data normalization; and weighting and aggregation schemes. A summary of these steps is presented in the remainder of this paper. The full detailed version of the methodology is available from the Statistics Department (ESTA), African Development Bank. The methodology, however, is not yet finalized as it is still undergoing peer review and will likely incorporate further revisions.

## **2. METHODOLOGY FOR COMPUTING THE PROGRESS CMI**

### **2.1 Selection of countries and indicators**

The first step in the methodology involves the selection of countries and indicators based on the following criteria. A country is selected if the percentage of missing values is 25% or below; an indicator is selected if the percentage of total population of countries without missing data is 75% or above. For a given indicator, a missing observation occurs for a given country whenever there are no data at all, or where the data available are such that two data points with at least 3 years apart cannot be found. Box 1 below illustrates the case of missing data.

**Box 1: Illustration of Missing Data**

Let us consider an indicator and four countries (C1, C2, C3 and C4). In the table below, ‘Yes’ means a data point is available and ‘No’ means a data point is not available.

Country	Years			
	1990	2000	2001	2002
C1	Yes	Yes	No	No
C2	No	Yes	Yes	Yes
C3	No	Yes	No	No
C4	No	No	No	No

C1 has two data points (i.e. 1990 and 2000) and the number of years apart is 10 (2000-1990=10). Hence, a trend can be computed.

C2 has three data points (i.e. 2000, 2001, and 2002). The number of years apart between the baseline and the latest year is 2 (2002-2000 = 2). Hence, in this case, C2 has a missing value for this indicator.

C3 has one datum and C4 has no data. Hence C3 and C4 have missing data.

**2.2 Computation of the expected years of achieving the targets**

In the second step, the expected year of achieving a target is computed based on the assumption that the upward trend or increasing indicator (e.g. Net primary school enrollment rate) follows a linear model and the downward trend or decreasing indicator (e.g. Under-five mortality rate) follows a geometric model. The relation between the earliest value and the latest value is expressed in a linear model by:

$$Y_{Lst} = Y_{Fst} + q(Lst - Fst) \quad (1)$$

In a geometric model this relation is expressed as:

$$Y_{Lst} = Y_{Fst} (1 + r)^{Lst - Fst} \quad (2)$$

The following notations are used in this relation:  $Lst$  and  $Fst$  are respectively the earliest and latest year,  $Y$  is the indicator value, and  $q$  and  $r$  are respectively the average yearly increase and the average growth rate. In addition to the model assumption, when the value of the 1990 baseline is unknown, an estimation of the 2015 value is carried out, based on the earliest value, by assuming that the required trend crosses three points: the 1990 baseline, the 2015 target, and the earliest point (supposed to be after the 1990 baseline). Therefore, three equations are required: two for computing the expected year of achieving a target. The equations and their solution for upward trend and downward trend are presented in the table below with the following notation:  $T_0$  is the 1990 baseline year,  $T$  is the 2015 target year,  $Fst$  is the earliest year,  $Lst$  is the latest year,  $\alpha$  is the authoritative trend,  $q_a$  and  $q_r$  while  $r_a$  and  $r_c$  are the actual trend and the required trend to reach the target for upward and downward trend indicator, respectively;  $\lambda = 1 / (T - Lst + 1)$  is the weight of the required trend.

Type of trend for indicator	Upward	Downward
Initial equation	$\begin{cases} Y_{Fst} = Y_{T_0} + q(Fst - T_0) \\ Y_T = Y_{Fst} + q(T - Fst) \\ Y_T = (1 + \alpha)Y_{T_0} \end{cases}$	$\begin{cases} Y_{Fst} = Y_{T_0}(1 + r)^{Fst - T_0} \\ Y_T = Y_{Fst} + q(T - Fst) \\ Y_T = (1 + \alpha)Y_{T_0} \end{cases}$
Expected value in 2015	$Y_T = Y_{Fst} \left( \frac{(1 + \alpha)(T - T_0)}{T - (1 + \alpha)Y_0 + \alpha Fst} \right)$	$Y_T = Y_{Fst} (1 - \alpha)^{\left(\frac{T - Fst}{T - T_0}\right)}$
Expected year of achieving the target	$T = Lst + \frac{Y_T - Y_{Fst}}{q_d}$	$T = Lst + \frac{\log\left(\frac{Y_T}{Y_{Fst}}\right)}{\log(1 + r_d)}$
Where	$q_d = \lambda q_r + (1 - \lambda)q_a$	$r_d = (1 + r_r)^\lambda (1 + r_a)^{(1 - \lambda)} - 1$

The derivation of the above equations is presented in Annex 1.

### 2.3 Imputation of missing data

The third step involves the imputation of data using the multiple imputations (MI) method. The “aregImpute” function of the library “Hmisc” in the R software is used to carry out the imputations. The procedure is that

for each missing expected year of achieving the target, five values are provided and the mean of these five values is considered as the final estimated missing value. The assumption of the multiple imputations is that: under uncertainty of the missing values, its imputed value is an average of N estimates. The N estimates are obtained using a regression model or a multinomial distribution. The Markov Chain Monte Carlo (MCMC) method is used for each of the N estimates. The MCMC assumes that the distribution of the current element depends on the value of the previous one; the first value is estimated from the dataset without the missing values; the expected maximum (EM) algorithm is run to select the other value.

## 2.4 Data normalization

In the fourth step, the categorization rule is used as a normalization scheme for computing the P-CMI. For each indicator, a country is 1 if the year of achievement is by 2015 and 0 otherwise. Category 1 may be referred to as “likely to achieve the target by 2015” and category 0 as “not likely to achieve the target by 2015”. The use of two categories is helpful in interpreting the P-CMI as a proportion.

## 2.5 Aggregation and weighting schemes

The fifth and last step involves the aggregation and weighting procedure that is carried out for each country. This procedure is carried in three stages. First, an aggregate index for each target is computed by averaging category values of indicators (these categories are 0 and 1 as defined at the normalization step). Therefore, for a given target, the aggregate index is a proportion of indicators with category 1 (i.e. likely to achieve the goal by 2015), which is an estimate of the probability of achieving the target. Second, an aggregate index for each goal is computed as the average indices for targets. For a given country, this index is the estimated probability of achieving the goal. Third and finally, *the P-CMI is computed as the average of indices for goals. For a given country, this index is the estimated probability of achieving the MDGs by 2015.* At each stage of aggregation mentioned above, equal weights are applied and the linear additive aggregation method is used. This procedure can be summarized by the following formula:

$$CMI^c = \sum_{j \in MDG} W_j \sum_{k_j \in G_j} W_{k_j} \sum_{i_{k_j} \in T_{k_j}} W_{i_{k_j}} F_{i_{k_j}} \quad (3)$$

In the formula (3) above, the following notations are used:

- $MDG$  is the set of the selected MDGs or all goals selected
- $G_j$  is a set of targets relating to a given goal  $j$
- $T_{k_j}$  is a set of indicators relating to target  $k_j$  of goal  $j$
- $F_{i_{k_j}}$  is, for country  $C$ , the normalized value of relating to target  $j$
- $W_{i_{k_j}}$  is the weight of indicator  $i_{k_j}$ . In our equal scheme  $W_{i_{k_j}} = \frac{1}{\#T_{k_j}}$  where  $\#T_{k_j}$  is the number of indicators in  $T_{k_j}$ .
- $W_{k_j}$  is the weight of target  $k_j$ . In our equal scheme  $W_{k_j} = \frac{1}{\#G_j}$  where  $\#G_j$  is the number of targets in  $G_j$ .
- $W_j$  is the weight of goal  $j$ . In our equal scheme  $W_j = \frac{1}{\#MDG}$  where  $\#MDG$  is the number of goals in  $MDG$

Let us identify terms of the formula (3)

- o The quantity  $\sum_{i_{k_j} \in T_{k_j}} W_{i_{k_j}} I_{i_{k_j}}^c$  is an unbiased estimator of the probability of achieving target  $T_{k_j}$  given the set of available information (which is the subset of indicators) since  $I_{i_{k_j}}^c$  can be modeled as Bernoulli variable.
- o The quantity  $\sum_{k_j \in G_j} W_{k_j} \sum_{i_{k_j} \in T_{k_j}} W_{i_{k_j}} I_{i_{k_j}}^c$  is an unbiased estimator of the probability of achieving goal  $G_j$  given the set of available information which is the subset of targets where each target has its own probability to be achieved estimated by the quantity  $\sum_{i_{k_j} \in T_{k_j}} W_{i_{k_j}} I_{i_{k_j}}^c$ .
- o Finally, the quantity  $CM^c$  is the probability of achieving the MDGs for country  $C$  based on the set of available information which is made of the selected goals where each goal has its own probability to be achieved estimated by the quantity  $\sum_{k_j \in G_j} W_{k_j} \sum_{i_{k_j} \in T_{k_j}} W_{i_{k_j}} I_{i_{k_j}}^c$

### 3. SENSITIVITY ANALYSIS OF THE METHODOLOGY

In order to assess the robustness of the methodology, a sensitivity analysis was performed on the basis of *five factors of uncertainties*<sup>3</sup> identified as: the inclusion/exclusion of indicators one by one; methods used for imputation of missing data; normalization rules; weighting schemes; and aggregation rules. For each factor, the alternative choices were assessed, for example: multiple imputations versus unconditional mean imputation for imputing missing data; categorization versus standardization (using z-scores) for normalization method; equal weights versus Principal Components Analysis and indicator variance for weighting schemes; and additive aggregation versus Principal Components Analysis for aggregation rule. A uniform distribution is assigned to each factor to carry out the selection procedure.

Sobol's indices are computed using "brute force approach" on the basis of the methodology as measures of sensitivity analysis. Furthermore, to assess the robustness of the overall P-CMI to outliers,<sup>4</sup> two M-estimators are computed: the *trimmed mean* and the *Winsorized mean*. The results obtained from the sensitivity analysis were adequate to confirm the robustness of the methodology.

### 4. CONCLUSION

A key issue that has been of concern and raised through some ongoing peer reviews, is the assumption to use two models, i.e. a linear model for upward indicator and geometric model for downward indicator. This assumption has, however, been justified and used by UNESCAP (2007). It may, however, be argued that normally, for each indicator, the selection of a model should be based on the data profile of each country. Therefore, there may be a need to carry out a cross-section study for identifying the underlying model for each indicator and each country in view of the data limitation.

---

<sup>3</sup> Uncertainty refers to the error due to the fact that for a given factor, one possibility is chosen for the reference methodology among several. The sensitivity analysis assesses this error for each factor.

<sup>4</sup> Outliers are extreme P-CMI.

## REFERENCES

Dudewicz, E. J. and S. N. Mishra (1998), *Modern Mathematical Statistics*, Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons.

ECA – STATCOM-AFRICA I, (2008), *Millennium Development Goals Monitoring: Challenges and Opportunities for African Countries*. Presented at the first Meeting of the Statistical Commission for Africa (STATCOM-AFRICA-I). Addis Ababa: UNECA. Available online at: <[http://www.uneca.org/statistics/statcom2008/documents/mdgs\\_monitoring.pdf](http://www.uneca.org/statistics/statcom2008/documents/mdgs_monitoring.pdf)>

ESCAP, UNDP, and the Asian Development Bank (2005), *A Future within Reach: Reshaping Institutions in a Region of Disparities to meet the Millennium Development Goals in Asia and the Pacific*. Bangkok: ESCAP. Available online at: <[http://www.mdgasiapacific.org/files/shared\\_folder/documents/Regional\\_MDGs\\_report\\_2.pdf](http://www.mdgasiapacific.org/files/shared_folder/documents/Regional_MDGs_report_2.pdf)>.

ESCAP, UNDP, and the Asian Development Bank (2006), *The Millennium Development Goals: Progress in Asia and the Pacific 2006*. Available online at: <[http://www.mdgasiapacific.org/files/shared\\_folder/documents/MDG-Progress2006.pdf](http://www.mdgasiapacific.org/files/shared_folder/documents/MDG-Progress2006.pdf)>.

ESCAP, UNDP, and Asian Development Bank (2007), *The Millennium Development Goals: Progress in Asia and the Pacific 2006*. Available online at: <<http://www.unescap.org/stat/mdg/MDG-Progress-Report2007.pdf>>.

Government of Papua New Guinea and United Nations in Papua New Guinea (2004), *Millennium Development Goals: Progress Report for Papua New Guinea*. Available online at: <[http://www.undp.org/documents/mdgs/National\\_MDG\\_Progress\\_Report\\_2004.pdf](http://www.undp.org/documents/mdgs/National_MDG_Progress_Report_2004.pdf)>

Harrell Jr., F. E. et al. (2006), “Hmisc: Harrell Miscellaneous”. Available online at: <<http://cran.r-project.org/web/packages/Hmisc/index.html>>

Nardo, M., M. Saisana, A. Saltelli, S. Tarantola, A. Hoffman and E. Giovannini (2005), *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD Statistics Working Paper. Paris: OECD. Available online at: <[http://www.oilis.oecd.org/oilis/2007doc.nsf/LinkToFrench/NT0000109A/\\$FILE/JT03226900.PDF](http://www.oilis.oecd.org/oilis/2007doc.nsf/LinkToFrench/NT0000109A/$FILE/JT03226900.PDF)>

Rouzier, P. (2003), *A Composite Index for Assessing Progress towards the MDGs*.

Saltelli, A. (2004), *Global Sensitivity Analysis: An Introduction*, European Commission, Joint Research Centre of Ispra, Italy.

## ANNEX 1: COMPUTATION OF THE TARGET VALUE

Notation:  $T_0$ ,  $Fst$  and  $T$  are respectively the 1990 baseline year, the earliest and the 2015 target year,  $Y$  is the value,  $q$  is the average yearly increase,  $r$  is the annual growth rate and  $\alpha$  is the required rate of change specified by the MDGs ( $\alpha = \frac{1}{2}$  for Share of poorest quintile in national consumption,  $\alpha = \frac{2}{3}$  for Proportion of births attended by skilled health personnel).

### 1.1 Computation of the target value for upward trend indicator

Let us consider the following system (where  $Y_T$ ,  $Y_{T_0}$  and  $q$  are unknown variables).

$$\begin{cases} Y_{Fst} = Y_{T_0} + q(Fst - T_0) & (1) \\ Y_T = Y_{Fst} + q(T - Fst) & (2) \\ Y_T = (1 + \alpha)Y_{T_0} & (3) \end{cases}$$

We want to derive  $Y_T$  based on  $Y_{Fst}$  and  $\alpha$ .

From (1) and (2), it comes that

$$q = \frac{Y_{Fst} - Y_{T_0}}{Fst - T_0} = \frac{Y_T - Y_{Fst}}{T - Fst}$$

Therefore

$$Y_{T_0} = Y_{Fst} - (Y_T - Y_{Fst}) \frac{Fst - T_0}{T - Fst} \quad (4).$$

Using (4) in (3), it comes that:

$$Y_T = (1 + \alpha) \left( Y_{Fst} - (Y_T - Y_{Fst}) \frac{Fst - T_0}{T - Fst} \right)$$

Grouping terms in  $Y_T$  and  $Y_{Fst}$ , it comes that

$$Y_T \left( 1 + (1 + \alpha) \frac{Fst - T_0}{T - Fst} \right) = Y_{Fst} (1 + \alpha) \left( 1 + \frac{Fst - T_0}{T - Fst} \right)$$

which is equivalent to

$$Y_T (T - Fst + (1 + \alpha)(Fst - T_0)) = Y_{Fst}(1 + \alpha)(T - T_0)$$

Finally,

$$Y_T = Y_{Fst} \left( \frac{(1 + \alpha)(T - T_0)}{T - (1 + \alpha)T_0 + \alpha Fst} \right)$$

## 1.2 Computation of the target value for downward trend indicator

Let us consider the following system (where  $Y_T$ ,  $Y_{T_0}$  and  $r$  are unknown variables):

$$\begin{cases} Y_{Fst} = Y_{T_0} (1 + r)^{Fst - T_0} & (1) \\ Y_T = Y_{Fst} (1 + r)^{T_0 - Fst} & (2) \\ Y_T = (1 - \alpha)Y_{T_0} & (3) \end{cases}$$

We want to derive  $Y_T$  based on  $Y_{Fst}$  and  $\alpha$ .

From (1) and (2), it comes that<sup>5</sup>

$$\log(1 + r) = \frac{\log(Y_{Fst}) - \log(Y_{T_0})}{Fst - T_0} = \frac{\log(Y_{T_0}) - \log(Y_{Fst})}{T - Fst}$$

Therefore,

$$\log(Y_{T_0}) = \log(Y_{Fst}) - (\log(Y_T) - \log(Y_{Fst})) \frac{Fst - T_0}{T - Fst} \quad (4).$$

Using (4) in (3), it comes that:

$$\log(Y_T) = \log(1 - \alpha) + \left( \log(Y_{Fst}) - (\log(Y_T) - \log(Y_{Fst})) \frac{Fst - T_0}{T - Fst} \right)$$

---

<sup>5</sup> Computations are carried out after applying the logarithm function on equation 1, 2 and 3 to obtain a linear system

Grouping terms in  $Y_T$  and  $Y_{Fst}$ , it comes that

$$\log(Y_T) \left(1 + \frac{Fst - T_0}{T - Fst}\right) = \log(1 - \alpha) + \log(Y_{Fst}) \left(1 + \frac{Fst - T_0}{T - Fst}\right)$$

which is equivalent to:

$$\log(Y_T)(T - T_0) = (T - Fst)\log(1 - \alpha) + \log(Y_{Fst})(T - T_0)$$

which is equivalent to:

$$\log(Y_T) = \frac{T - Fst}{T - T_0} \log(1 - \alpha) + \log(Y_{Fst})$$

which is equivalent to:

$$\log(Y_T) = \log(1 - \alpha)^{\frac{T - Fst}{T - T_0}} + \log(Y_{Fst})$$

which is equivalent to:

$$\log(Y_T) = \log \left[ Y_{Fst} (1 - \alpha)^{\frac{T - Fst}{T - T_0}} \right]$$

Finally,

$$Y_T = Y_{Fst} (1 - \alpha)^{\frac{T - Fst}{T - T_0}}$$