

The aggregation problem in its hystorical perspective: a summary overview*

G. Lutero[†]

ISTAT, National Accounts Directorate
Methods Development of Quarterly National Accounts

Abstract

This document covers the problem of aggregation in the context of statistical and econometric linear modelling under rural development. A brief updated overview on this topic, much debated and approved, will be given in different conceptual frameworks: longitudinal or spatial, temporal and contemporaneous. It will be shown the possible introduction of bias in aggregate macroeconomic models and the stochastic characteristics change from higher frequency models to lower ones. The concept of multi-dimensionality as a means to overcome the one-dimensional analysis of development and poverty and also as an alternative measure of welfare will be discussed in its empirical and methodological aspects. A focus will be dedicated to the specific steps necessary to identify the weights and the aggregation methods in the construction of composite indices.

Keywords: aggregation theory, aggregation bias, temporal aggregation, composite indicators, multidimensionality.

1 Introduction

Has long been an established practice of national governments and international institutions to make decisions about the conduction of their policies

*I am greatly indebted to Edoardo Pizzoli and Adriano Pareto for their suggestions and criticism. Any errors and shortcomings remain my own responsibility. Opinions expressed in this paper are exclusively those of author and do not reflect the official position of ISTAT.

[†]Contact author. Via A. Depretis 74/B, Rome. Email: lutero@istat.it.

(i.e. agricultural policies and development) on the basis of suggestions coming from national statistical agencies or economic analysis. Although it is not sufficiently explained to the public, the construction of published statistical data is a long and complex process involving institutional and theoretical constraints, and in which the *aggregation* is one of the most sensitive stage but often overlooked. The aggregation issue is back again the focus of scholars in relation to the necessity of supranational institutions such as central banks, economic and political integrations (such as, for example, the European Union), to manage economic policy in single countries. In such contexts, the availability and accuracy of data become strategic and decisive: the process of aggregation should not be underestimated because it could lead to misinterpretations and consequently to wrong intervention's policies.

A new interest in the theory of aggregation is likely to come from the efforts of the mainstream schools of economics to find a proper and strong foundation of macroeconomic theory and models, based on the reductionist Marshallian agent-based approach. Also, a big impulse could come from new methodological approach HIA (Heterogeneous Interacting Agents) and the agent-based paradigm¹, developed in studies of economic relations interpreted with complex non-linear systems, in which events occur, such as *emergence* and *asymmetric evolutionary dynamics*.

In statistical literature, the aggregation has been studied from many points of view; according to some specific characteristics and to a certain degree of abstraction, could be identified the following three typologies:

- *longitudinal or spatial aggregation*, which concerns the aggregation through geographic space (usually Regions or lower administrative areas) or a combination of physical and institutional units (households, businesses and institutional sectors), or the productive sectors;
- *temporal aggregation*², that concerns passage of time series from an higher frequency data (i.e. quarterly) to lower one (typically annual data);
- *contemporaneous aggregation*, in which the aggregation is done across variables (i.e. the construction of a composite index such as the HDI - Human Development Index, created in 1990 for the purposes of the

¹See Kirman and Timmermann (2001) for an introduction to agent-based approach.

²The inverse operation, temporal disaggregation, is a set of well-known techniques adopted by the Central Offices of Statistics in the production of infra-annual (quarterly or monthly) time series, but is not object of this dissertation, see Bee Dagum and Cholette (2006) for a detailed reference.

UNDP - United Nations Development Programme), or between different prediction models (i.e. the combining or pooling forecasting³).

Actually, these three kinds of aggregation are interrelated: contemporal aggregation might to regard as well time series data⁴ and / or dynamic models; longitudinal aggregation may be associated with temporal aggregation; or, more, building a composite index is made with macroeconomic aggregates and so on.

Several theoretical approaches have discussed the degree of aggregation consistency, developing statistical tests that would identify the conditions for perfect aggregation, while empirical studies are more concerned on which level of aggregation or disaggregation is optimal for analysis. Hence, which are the reasons of working with highly aggregated data? Here are mentioned the most significant ones:

- in some cases the choice between aggregated and disaggregated data is constrained by their availability on the studied phenomena (often, only aggregate data are available);
- simplicity and parsimony; few numbers are very effective in summarizing complex issues in a simple and understandable way to the public and policy-makers;
- statistical processing is faster and clearer; this is a key factor when the data have to be released at fixed deadlines;
- to work on macro variables can be a practical solution if the micro-data are from incomplete or unreliable statistical sources;
- macro-data, in some cases, can solve the problems in micro-data, as well as reduce their impact on model specification, that is the probable introduction of a systematic error;
- the collection and processing of new micro-data from statistical surveys is very expensive and poses a further constraint to their availability: it is therefore necessary to properly evaluate their real contribution in terms of statistical information.

Conversely, working with disaggregated models and micro-data has the following advantages:

³See Timmermann (2006) for an introduction to forecast combinations methodologies.

⁴Composite Leading Indicators, released by OECD, is a very good example of index used for study and dating of business cycle, see OECD (1987) for insights and Marcellino (2006) for a theoretical reference.

- completeness of information; all information content of the source data is preserved and data reduction (aggregation) is just one of several possible choices for the analysis;
- better estimates with respect to macro-models estimates; greater availability of observations involves more degrees of freedom in data processing and, consequently, the estimates are more efficient, the statistical tests are more powerful, etc.;
- alternative model specifications; the availability of micro-data allows to consider various specifications of the model on micro-units depending on the circumstances;
- better forecasts with respect to macro-models estimates; is a consequence of the increasing accessibility of micro-data and a differentiated, and therefore better, specification at the micro level.

The physiological development in applied disciplines result in new and more sophisticated econometric methods, such as panel data models or dynamic factorial analysis, that imply, for example, the possibility to preserve the information content of micro-data and, at same time, to find a common-aggregate representative measure of macroeconomic behaviour of micro-data and a separate, and therefore better, specification at the micro level.

Very often the scientific work on this issue have focused on identifying the best linear predictors for aggregates, whether the forecasts are more efficient if done with aggregate variables or on the aggregation of forecasts by individual micro equations. The first question that scholars should consider is what influence the accuracy of predictors because, unfortunately, several factors could be sources of statistical bias. The sources of error can be summarized as follows:

- model specification identification of variables to include;
 - functional form (linear aggregation vs nonlinear);
 - model selection criteria;
- estimation uncertainty;
- data measurement errors;
- structural breaks over forecast horizon.

The available statistical and econometric procedures (parametric and non-parametric) to address the problem of aggregation may be classified into the following macro-groups:

- regression model-based stochastic framework:
- static models (cross section data);
- dynamic models (time indexed, i.e. typically ARIMA models or OECD's Composite leading indicators);
- nonmodel-based deterministic framework (composite indexes, social indexes).

In recent years has increased a lot, among scholars and public opinion, attention towards the use of composite indexes to measure economic performance, to define a more precise concept of well-being, in order to enable the international comparability of countries not exclusively related to GDP per capita or other main official statistics. The construction of composite indicators involves having to deal with the concept of *multidimensionality*, needed to interpret a more complex and contradictory economic reality: on this subject, the combination of variables involves the use and selection of an optimal aggregation rule (i.e. estimation of parameters in the regression models is a kind of this rule), where the same difficulties arise as in stochastic model approach.

In this paper a brief summary and updated overview on the theory of aggregation is provided. Section 2 will introduce the concept of aggregation bias in the cross-section data and its treatment in a static and in a dynamic time series context. Section 3 provides a brief introduction to the effects of aggregation in dynamic stochastic models at high frequency. Section 4 discusses the concept of multidimensionality and the production of composite indexes for the analysis and measurement of developing countries, illustrating the necessary steps for the normalization of data, the allocation of weights and aggregation in whole. Section 5 concludes with some final comments.

2 Longitudinal aggregation

In the representation of economic facts, the problems arising from the aggregation are the most controversial and discussed in econometrics and statistics, starting from the seminal contributions of Leontief (1947), Theil (1954), Malinvaud (1956), Grunfeld and Griliches (1960), and Zellner (1962). Theil in his famous book introduces for the first time the concept of aggregation bias, which can easily be defined as the deviation of the parameters of the aggregate linear model from the average of corresponding micro parameters in a systematic way.

Formally, the issue of aggregation bias is the possibility that, in relation to the true β value, the asymptotic bias can be greater in the aggregated estimator rather than in the disaggregated one, that is

$$\left| \text{plim} \left(\hat{\beta}_d - \beta \right) \right| < \left| \text{plim} \left(\hat{\beta}_a - \beta \right) \right|$$

For the sake of simplicity, the theory will be exclusively illustrated in relation to univariate linear models and finite sample. Following Theil (1954), let us consider disaggregated models relative to a single micro-unit i (i.e. country, industry, household) at time t

$$y_{it} = x'_{it}\beta_i + u_{it} \quad i = 1, \dots, n \quad t = 1, \dots, T \quad (1)$$

where y_{it} is dependent variable relative to unit i at time t , $x_{it} = (x_{1it}, \dots, x_{kit})$ is a vector of k exogeneous regressors, β_i is vector of coefficient for i -th section and u_{it} is a stochastic error with usual hypothesis (mean equal zero and variance σ_i^2). Assuming linear additive aggregation, let us define the equation relative to aggregated data

$$y_{at} = \sum_{i=1}^n y_{it} = x_{at}\beta_a + u_{at}$$

To identify the statistical correspondance between micro-equations and macro-equation disturbances, Theil adopts a set of auxiliary relations, computing projections of micromodel' predictor variables on the corresponding of aggregated ones

$$x'_{it} = Z_{rt}\delta_i + \varepsilon_{it} \quad (2)$$

where matrix $Z_{rt} = \text{diag}(x_{1at}, x_{2at}, \dots, x_{kat})$ and ε_{it} auxiliary regression shocks. Finally, substituting equation (2) in structural micro equation (1) and summing across units, estimated aggregate disturbances is the sum of two magnitudes, respectevly the *aggregation bias* and sampling disturbances

$$\hat{u}_{at} = \sum_{i=1}^n \hat{u}_{it} = \sum_{i=1}^n (\varepsilon'_{it}\hat{\beta}_i + \hat{u}_{it}) \quad (3)$$

Hence variance of aggregate error is the following

$$\hat{\sigma}_a^2 = n^{-1} \left(\sum_{t=1}^T \varepsilon'_{it}\hat{\beta}_i \right)^2 + \sum_{i=1}^n \hat{\sigma}_i^2 + \sum_{j \neq i} \sigma_{ij} \quad (4)$$

From this last relation is verified that variance of aggregation bias can be computed as a weighted mean, whose weights are represented by estimated

squared micro coefficients. When first sum vanishes, it is obtained the so-called *consistent* or *perfect aggregation* (outputs, in terms of goodness of fit, are approximately the same, in relation to target), that may occur in the following cases⁵:

- *Micro homogeneity*⁶: all microparameters are equal to aggregated one, so they are constant over the longitudinal dimension;
- *Compositional stability*: it regards joint probability distribution of regressors and it occurs when composition of the regressors across units do not change over time.
- *Symmetric distribution*: regressors are white noises.

Theil's analysis is conducted assuming very strong hypothesis of correct specification of both models, aggregate and disaggregates. Relaxing to this last assumption, Grunfeld and Griliches (1960) proposed that grouping data may in some cases lead to an improvement in terms of unbiased estimates and efficiency (*aggregation gain*), evaluated using a goodness of fit criterion based on \bar{R}^2 or considering efficient prediction. These authors conclude that disaggregated models are likely misspecified, i.e. because in micro equations is absent influence of variables representative of macroeconomic behaviors, or because microdata could present measurement errors⁷: in these occasions aggregation is not "necessarily bad".

A very interesting of Pesaran, Pierse and Kumar (1989) resumes state-of-the-art on static linear aggregation and proposes a more general misspecification test, comparing the differences between average of estimated micro coefficients and aggregated estimate. Null hypothesis of perfect aggregation is the following

$$\xi = \sum_{i=1}^n X_i \beta_i - X_a b = 0$$

If $\xi > 0$ authors stress that this statistic test might be taken as measure of misspecification in micro equations.

Stoker (1984) stressed that is necessary to define a *complete* aggregation structure between macro aggregate and micro functions to detect a behavioral interpretation of the former, that's to say there should be a one-to-one

⁵Unfortunately these conditions turn out to be rather stringent: cases here reported are almost exclusively of mathematical interest and scarce economic significance.

⁶This is hypothesis formulated by Zellner for perfect aggregation test, introduced in context of SURE estimator, see Zellner (1962).

⁷See Aigner and Goldfeld (1974).

correspondence between macro function and micro functions. Completeness is a statistical feature which allows to extend and to incorporate analysis of nonlinear micro behavior and to deepen distribution of predictors across units⁸.

Lippi and Forni (1990) expand investigation to dynamic models specification and showed that adopting an ARMAX framework for micro equations as follows

$$\phi(L)y_{it} = \gamma(L)x_{it} + \theta(L)\varepsilon_{it}$$

with usual stochastic assumptions and no common polynomial factors, it is possible to improve study of individual heterogeneity, providing for relations between dynamic structure of several micro equations, represented by polynomials in lag operator $\phi(L), \gamma(L), \theta(L)$. Besides they stress that dynamic shape is altered in the passage from micro to macro equations (*dynamization effects* of aggregation⁹) and, even though aggregated model might possess much richer dynamic specification, well-specified disaggregated models give in general better outcomes.

2.1 A new approach to aggregation

Dynamic factor analysis and panel data models are most recent and refined techniques to treat in a proper way theme of aggregation bias and correlated issue, relating to selection of optimal disaggregation degree.

Lippi e Forni adopted the Factor analysis¹⁰ to determine aggregation error related to microdata. Let us suppose that explanatory variables are collected in matrix $X_t = (x_{1t}, x_{2t}, \dots, x_{nt})$, and it may be split in two distinct, orthogonal processes

$$X_t = WC_t + h_t \tag{5}$$

where $C_t = (c_t, c_t, \dots, c_t)$ is a matrix of non-observable common factors, $h_t = (h_{1t}, h_{2t}, \dots, h_{nt})$ is a matrix of idiosyncratic components and $W =$

⁸For example density classes of exponential family are complete in relation to aggregation.

⁹These dynamization effects may arise from errors in set of available variables, from temporal aggregation (see section 3 of this paper), from omission of relevant variables and finally from adoption of linear framework to interpret nonlinear relationships.

¹⁰Factor analysis representation has been introduced in time series by Lucas (1973), Sargent and Sims (1977) and Geweke (1977). See also Forni and Lippi (1997), chap. 1, for an introduction to this approach.

(W_1, W_2, \dots, W_n) is a matrix of static loadings coefficients: unobserved common movements along time (latent factors) are extracted from observed aggregated data and residual magnitude is regarded as individual specific shock. The micro model expressed in factor analysis approach is the following

$$y_{it} = (W_i c_t + h_{it})' \hat{\beta}_i + \hat{u}_{it} \quad (6)$$

Substituting and summing across units, aggregated model will be

$$y_{at} = \sum_{i=1}^n y_{it} = \sum_{i=1}^n (W_i c_t)' + \sum_{i=1}^n h_{it} \hat{\beta}_i + \sum_{i=1}^n \hat{u}_{it} \quad (7)$$

Similarly to equation (3), estimated aggregate error is a mixture of aggregation bias and misspecification errors in microdata

$$\hat{u}_{at} = \sum_{i=1}^n \hat{u}_{it} = \sum_{i=1}^n (h_{it} \hat{\beta}_i + \hat{u}_{it}) \quad (8)$$

Estimated variance of aggregated factor model is

$$\hat{\sigma}_a^2 = n^{-1} \left(\sum_{t=1}^T h'_{it} \hat{\beta}_i \right)^2 + \sum_{i=1}^n \hat{\sigma}_i^2 + \sum_{j \neq i} \sigma_{ij} \quad (9)$$

In factorial formulation, aggregation bias is directly expressed by idiosyncratic components: Theil's conditions of compositional stability results from absence of micro individual heterogeneity that involves the vanish of first term in equation (9) as for micro homogeneity case.

Later Forni and Lippi (2001) proposed a generalization of factor model, introducing a dynamic relation between variables and unobserved common components and relaxing condition of orthogonality for idiosyncratic components. For every units, there will be following relation

$$y_{it} = b_{i1}(L)c_{1t} + b_{i2}(L)c_{2t} + \dots + b_{iq}(L)c_{qt} + h_{it}$$

To estimate common movements are available several methodologies: static principal component, dynamic principal component and structural state-space approach.

Increasing availability of temporal and longitudinal microdata has allowed to scholars improvement of new analysis instruments like panel data models, with purpose to analyze more complex economic and social phenomenons. Panel approach agrees construction of richer economic models in which heterogeneity is parametrized and final estimates are more linked to individual micro behaviors; this methodology is characterized by estimation of a common

slope parameter $\hat{\beta}$ which summarizes interaction among statistical units at a whole, and considering, at the same time, individual heterogeneity represented with a set of local intercepts as in the following structure

$$\begin{aligned} y_{it} &= \mu_i + \beta' x_{it} + u_{it} & i = 1, 2, \dots, N & \quad t = 1, 2, \dots, T \\ E(u_{it}) &= 0 & E(u_{it}^2) &= \sigma_u^2 \end{aligned}$$

where μ_i (*fixed effects*) are treated as unknown parameters that represent a particular realizations of stochastic processes, or where μ_i are regarded as a random sample derived from a statistical distribution (*random effects*)

$$\begin{aligned} y_{it} &= \alpha + \beta' x_{it} + \mu_i + v_{it} \\ E(\mu_i^2) &= \sigma_\mu^2 & E(v_{it}^2) &= \sigma_v^2 & E(\mu_j v_{it}) &= 0 \\ E(\mu_j v_{it}) &= 0 & E(v_{it} v_{is}) &= 0 & E(\mu_i \mu_j) &= 0 \end{aligned}$$

without loss of information given by not correct aggregation degree or data with only one dimension alone (temporal or sectional). A direct consequence of greater sample size is improvement of estimated model parameters accuracy and less multicollinearity, as a result of a greater degrees of freedom. Finally, another great advantage in use of panel models is opportunity to manage two types of variability, that results from data features and structure (cross-section and temporal dimension), though only one of those is sufficient to estimate all parameters. Therefore, to obtain efficient estimates is advisable adoption of methodology like FGLS (Feasible Generalized Linear Squares) that provide for both sample variances (*within estimator* and *between estimator*).

$$\hat{\beta} = \left(X' \hat{\Omega}^{-1} X \right)^{-1} X' \hat{\Omega}^{-1} y \quad (10)$$

where $\hat{\Omega}^{-1}$ is empirical variance-covariance matrix.

After this short analysis, conclusions are that aggregation bias in static or dynamic aggregated stochastic models could be determined by

- micro parameters values;
- misspecification of equations (omitted variables in micro models or, conversely, models overparametrization);
- measurements errors;
- missclassification errors;
- dynamic interrelation between the behavioral structural micro equations.

At last, maybe it is trivial but useful to stress that in small or finite sample, in several occasions, it is suggested the adoption of a biased but efficient estimators with respect to an unbiased, but less efficient one. Also in this case, a common sense rule suggests to evaluate magnitude of bias and its impact on estimates diagnostics, and remembering that choices should be influenced by goals of analysis and by experience and knowledges of involved scholars.

3 Temporal aggregation

Aggregation process along time dimension covers transformation of short term data to longer one (i.e. passage from quarterly data to annual). Data appear in *flows* or *stocks*: in the first case the aggregation is defined *average sampling* or simply *temporal aggregation*, in the latter *point-in-time sampling*. This brief analysis is limited to univariate context¹¹, but it could be extended to models with seasonality, garch models, multivariate analysis, long memory models, random aggregation, time continuous aggregation, nonlinear aggregation, dynamic factor models, etc.

Let us define a stationary stochastic process y_t at time frequency t . High frequency data are available at aggregation frequency k (i.e. $k=4$ regards quarterly data, $k=12$ monthly data). To obtain low-frequency series is applied a linear transformation to high frequency data, represented by the polynomial function $A(L) = (1 + L + L^2 + \dots + L^{k-1})$, where L is the lag operator such that $y_{t-k} = L^k y_t$

$$y_{lt} = A(L)y_t = \sum_{i=0}^{k-1} \omega_i y_{t-i}$$

Let us start with the most common stochastic process, the autoregressive AR(p), of order p

$$\phi(L)y_t = \xi_t \tag{11}$$

Applying to AR(p) model the polynomial $A(L)$, which transforms the high-frequency data to low-frequency ones

$$A(L)\phi(L)y_t = A(L)\xi_t \tag{12}$$

resulting model will be an ARMA(p,s), where the order s is represented by the following ratio

¹¹See Amemiya and Wu (1972), Brewer (1973), Weiss (1984) for a more deeply treatment of this argument and Granger (1990) for a survey.

$$s = \left[\frac{(p+1)(k-1)}{k} \right]$$

in which square brackets identifies the integer part of ratio. Solution descends from a nonlinear system of equations, taking into accounts the autocovariance structure of MA polynomial.

As well, same transformation operated on Autoregressive Moving Average ARMA(p,q) model

$$\phi(L)y_t = \theta(L)\xi_t \quad (13)$$

leads to an ARMA(p,s), in which magnitude s this time is

$$s = \left[\frac{(p+1)(k-1) + q}{k} \right]$$

If we also considered nonstationary processes, we could to set up into analysis Integrated Autoregressive Moving Average ARIMA(p,d,q), where parameter d is order of integration

$$(1-L)^d \phi(L)y_t = \theta(L)\xi_t \quad (14)$$

In this case the aggregation involves switch of original model to ARIMA(p,d,s), and the following value of parameter s

$$s = \left[\frac{p(k-1) + (d+1)(k-1) + q}{k} \right]$$

To conclude this section, it is possible to accurately characterize aggregated stochastic processes when aggregation scheme and native processes are known. Resuming:

1. the order of AR(p) component is invariant to temporal aggregation, unless that in some particular circumstances;
2. the AR coefficients are very different relating to the primals;
3. MA component is generally introduced in consequences of aggregation while was absent from original processes;
4. an ARMA process is generally introduced by aggregation of not-ARMA or by several ARMA processes;
5. in most cases the aggregated stochastic process is more complicated respect to the native one.

6. It is useful to stress that, however, disaggregated models are information richer and number of observations is k-times greater than aggregated data, although parameters contain all information about disaggregated sample, with all the consequences in terms of consistency and efficiency of estimates.

4 Contemporaneous aggregation

In the last years attention around employ of synthetic indexes in international comparability is increased a lot. Starting from debate around *capability approach* to well-being, concept introduced by Nobel Prize in economics Amartya Sen¹², any scholars point up that maybe economic and unidimensional approach cannot be considered satisfactory to analyze a complex phenomena like society well-being or rural development. Contemporaneous aggregation regards combination of several indicators, relating to a target variable (i.e. the Human Development Index (HDI) a well-being measure, released from 1990 in the scope of United Nations Development Programme). Recently it is born a great discussion between citizens and scholars around these topics, on which has not emerged yet an international consensus: a primary role in this debate has been assumed by international conference “*Beyond the gdp*”¹³, whose “*Report by the Commission on the Measurement of Economic Performance and Social Progress*”¹⁴ is the most cited and famous document.

Use of composite indicators involves the concept of *multidimensionality*, necessary to introduce element like well-being, life satisfaction or sustainable development in evaluation of policy-making. For example OECD’s experts have classified set of indicators in four macro domains to differentiate and to improve their investigation: self-sufficiency, equity, health and social cohesion.

As you would expect, use of social indicators is justified by their power to provide additional useful information respect to GDP per capita or others main official statistics. Way in which statistical figures are reported or used by media or politicians may provide a distorted view of economic phenomena. Great attention is usually put exclusively on GDP of a country: GDP is not a bad thing *in itself*, but rather is incorrect separate it from whole system of national accounts, not considering, for example, magnitude and dynamic of

¹²Sen (1985) and Sen (1992).

¹³See <http://www.beyond-gdp.eu> to explore the goals of this international forum around concept of well-being.

¹⁴See Stiglitz et al (2009).

net national product or the framework of intermediate consumption, or again primary the secondary distribution of national income. When debate is very conflictual, the danger is that also scholars or people who are critics against the instrumental use of GDP, risk to assume interpretative categories typical of mainstream dominant schools, which implies subjective, individualistic and utilitarian approach, that's to say the same theoretical foundations of those who they would criticize.

Saisana and Tarantola (2005) suggest positive features linked to employ of a composite index:

- it summarizes complex and multidimensional issues;
- it is easier to interpret respect to many different indicators;
- it facilitates communication with public opinion (citizens and media);
- it may assess progress of specific country over time.

HDI is one of most famous and representative indicators, and it is very helpful to introduce fully aggregative approach to well-being measures: it is computed as simple arithmetic average of three indexes that measure wealth, education and health, whose second one is in turn a weighted mean. Let us define HDI properly: inserted indicators are *life expectancy at birth* (b), *combined gross enrolment ratio* (s), *adult literacy rate* (a), and finally natural logarithm of *gdp per capita* (y), where subscripts u is for upper value and l for the lower one. Formally it is defined

$$hdi_i = \frac{1}{3} \left\{ \left(\frac{b - b_l}{b_u - b_l} \right) + \left[\frac{1}{3} \left(\frac{s - s_l}{s_u - s_l} \right) + \frac{2}{3} \left(\frac{a - a_l}{a_u - a_l} \right) \right] + \left(\frac{\ln y - \ln y_l}{\ln y_u - \ln y_l} \right) \right\}$$

Each variable is normalised with min-max method (see next subsection for a survey on normalisation methods). Afterwards, it will be illustrated technical difficulties concerning construction of index like that in relation to its implementation in international comparability.

Following Hoffman et al (2008), a more detailed analysis of each singular step in construction of a composite indicator may be synthesized in the following steps:

- Strong theoretical references;
- Variables selection;
- Missing data imputation;

- Multivariate analysis approach;
- Data normalisation;
- Weighting scheme identification;
- Aggregation of indicator variables;
- Check over composite indicators robustness and sensitivity analysis.

The strong relationship among points 5, 6 and 7 will be investigated hereinafter, considering them specific elements of aggregation issue as a whole.

4.1 Data Normalisation

Before aggregation step, it is necessary reduction of data to a common numeraire: without a doubt adopted selection criteria influences observations that will be classified as outliers. This phase assumes a crucial significance for composite indexes because normalisation already introduces an *implicit* or latent system of weights in procedure.

Most common methods of data normalisation are the followings:

- Scale changes: mostly variables are expressed in different measure unit. To apply transformation like logarithmic has many advantages: it allows to reduce data dispersion and to center asymmetric distribution. Besides, logarithmic derivative approximates the percentage growth rate of indicators;
- Standardisation (z-scores): the most used. Each variable is centered over average across countries and normalised with standard deviation across countries. For each country i , at every time t will be

$$z_{it} = \frac{x_{it} - E_i(x_{it})}{\sigma_t}$$

- Min-Max method: data are reported to lower and upper values across countries and assume values between zero and one. For each variable, it will be

$$z_{it} = \frac{x_{it} - \min_i(x_{it})}{\max_i(x_{it}) - \min_i(x_{it})}$$

- Distance from a time benchmark: variables, for each country i , are reported to average value at pre-fixed initial time: in this way indicators are followed in their dynamic evolution

$$z_{it} = \frac{x_{it}}{E(x_{it})_{t=t_0}}$$

- Latent variables (threshold models): it is attributed a score in accordance with a threshold value c (fixed in advance or estimated with i.e. numerical methods)

$$z_{it} = \begin{cases} 1 & \text{if } w > (1 + c) \\ 0 & \text{if } (1 - c) \leq r \leq (1 + c) \\ -1 & \text{if } r < (1 - c) \end{cases}$$

- Growth rates over consecutive years: this transformation cuts out influence of temporal dimension

$$z_{it} = \frac{x_{it} - x_{it-1}}{x_{it}} * 100$$

4.2 The system of weights

In relation to composite indexes formation, set of weights may be classified in three ways:

- *Subjective*: weights are attributed following a subjective probabilities (degree of confidence) that certain events occur;
- *Objective*: system of weights result from application of a statistical rules or methods
 - *exogeneous*, pre-determined: weights are imposed by nature of problem and consequently they are defined in advance, mostly as outcomes of an estimation procedure not directly linked to treated issue;
 - *endogeneous*: system of weights is estimated by means a constrained optimization procedure;
- *equal weighting*: it has been also considered composite indexes that assign an equal system of weights; this is a solution to reduce interferences to a minimum, or it is a consequence of lack of information. In this occasion degree of “neutrality” is reduced by previously adopted

data normalisation methods. A set of weights all equal to one is a special, naive, case.

Instead, evaluating weights in relation to temporal domain we may consider:

- *static weights*: system of weights do not change over time because analysis is involved with evolution of variables;
- *dynamic weights*: this is best solution when weights are output of a parametric optimization procedure.

There are a lot of methodologies available; here is a brief list of proposals:

- *Multivariate estimation techniques*: these methods of data reduction cannot be used when indicators are weakly correlated or when it is impossible to identify latent factors. Most widespread methods are:
 - *Principal component analysis*;
 - *Factor analysis*;
 - *Cluster analysis*.
- *Linear programming* (Data Envelopment analysis): it is based on identification of an efficiency frontier, that represents the optimal benchmark for all countries; distance from frontier measures relative position of a country in a multidimensional framework.
- *Benefit of the doubt approach*: it is computed as ratio between actual performance and a benchmark performance, this last one obtained with a constrained optimization procedure.
- *Unobserved component modelling*: it is assumed that each indicator is related to an unobserved variable plus an error component; this last captures different source of “noise” (misspecifications, measurement errors, aggregation effects, etc.).
- *Budget allocation process*: weights are settled by an experts’pool, who assign a budget score (max 100 points) to each indicators. Outcomes are computed with average budgets; experts cannot be specialists only in individual indicators, but they must possess a wide knowledge of arguments that have been selected to enter in composite index.
- *Public opinion*: as in the previous point, but in this case people express a degree of concern about several issues. The positive matter of these last two methodologies is their transparency.

- *Analytic hierarchy process*: it is a hierarchical methodology, in which opinions are pairwise compared. A matrix of comparison show the ordinal preferences, expressed by relative weights.
- *Conjoint analysis*: preferences are disaggregated and several scenarios are considered in a preference function. Weights are substitution rates (how much the preferences change according to change in indicators) that enter in total differential of objective function.

4.3 Aggregation methods

Final step in the construction of composite indexes is functional aggregation rule. Considerations discussed in section 2 are helpful as well in aggregation of composite indicators, while here is not explicitly formalized a stochastic component in modelization. Besides, it is important to underline that this phase is subsequent to an aggregation operation which is already occurred upstream in the process of collection and treatment-estimation of data. In this phase of process, concept of *compensability degree* between indicators assumes a relevant role in relation to optimal aggregation technique selection; the most known methods are:

- *Linear additive aggregation (LAA)*: most simple use of LAA is based on ordinal information, so that, for each country is considered the simple summation of relative ranking in relation to each individual variable

$$y_{it} = \sum_{j=1}^k rank_{ij} \quad i = 1, 2, \dots, n \quad (15)$$

This index is very easy to compute and is unaffected by outliers but, in aggregated index, are lost absolute values' information. Another method computes times that an indicator is above and below a threshold value c

$$y_i = \sum_{i=1}^k sign \left[\frac{z_{it}}{E_i(z_{it})} - (1 + c) \right] \quad (16)$$

In this aggregate disappears interval level information. The most used linear aggregation involves normalisation of indicators and their linear combination

$$y_i = \sum_{j=1}^k \omega_j z_{ij} \quad \sum_{j=1}^k \omega_j = 1 \quad 0 \leq \omega_j \leq 1 \quad i = 1, 2, \dots, n \quad (17)$$

with weights that not represent importance of associated indicator but quantify substitution or compensation rates. A very recommended and powerful condition in LAA is *preferential independence*: this implies that trade-off ratio between two indicators is independent from the remaining (k-2) variables and, by consequence, it is possible to evaluate separately marginal contribution of each indicator. In LAA, full and constant compensability is advisable, but in many occasion it is unrealistic and it is a cause of bias introduction.

- *Geometric aggregation (GA)*: it is also defined deprivational index and it is a good solution to avoid full compensability, typical of linear additive aggregation. GA is defined

$$\prod_{j=1}^k z_{it}^{\omega_j} \quad (18)$$

Since in GA compensability degree is not constant, because is higher for composite indexes with high values and vice versa, countries with low scores tend to prefer use of linear aggregation, trying to improve their position in ranking.

- *Non-compensatory multi-criteria aggregation (MCA)*: it is a non compensatory approach that allow to interpret weights as “importance coefficients”. In this methodology qualitative and quantitative information are both taken into account and, moreover, indicators do not need of any pre-treatment for comparability. Procedure is articulated in two steps; in the first there is a pairwise comparison of countries: this information is inserted in a *outranking matrix* and total score is obtained with the following formula

$$c_{ab} = \sum_{i=1}^k \omega_i (Pr_{ab}) + \frac{1}{2} \omega_i (In_{ab}) \quad (19)$$

where $\omega_i (Pr_{ab})$ and $\omega_i (In_{ab})$ are, respectively, weights of specific indicator that presents preference and indifference. At the end of aggregation procedure is attributed a score to every possible ranking combination¹⁵.

Resuming, composite indexes are relatively easy to build, to explain and to interpret by professional users and by ordinary citizens, but it is my opinion that this apparent simplicity occults many real pitfalls.

¹⁵See Munda and Nardo (2005) and Hoffman et all (2008) for a numerical example.

5 Concluding remarks

This paper illustrates in a very synthetical way problems related to aggregation procedure in stochastic and deterministic models, considered in its several theoretical frameworks (longitudinal, temporal or contemporaneous). Although it was an issue well known from many years, aggregation is yet largely debated in statistical and econometric disciplines. A great impulse is coming from the birth of new political and supranational institutions like European Union and also from efforts of some scholars to exceed the limits of reductionist and linear analysis and the wish to go over economic and unidimensional analysis of development and poverty, discovering alternative and more nuanced measures of well-being.

Choice between macrodata and microdata and, after that, between aggregated or disaggregated data, cannot be resolved by *a priori* reasoning, but it may find only an empirical solution, in the context of particular applications, because several elements (estimation uncertainty, models misspecification in some level of aggregation, endogeneous dynamic relation between predictors) could be sources of statistical bias: the principal alternative seems to be maximum level of information *versus* a greater readability for citizens and policy-makers. Nevertheless, generally, it is advisable use of microdata for parameters estimation of stochastic models because it is more efficient respect to macrodata, as a result of presence of more degrees of freedom available. On the contrary microdata models might suffer of specification errors or instability problems: in these occasions is more efficient to work with aggregated data (aggregation gain).

As suggest Blundell and Stoker (2005) there are two good reasons to face aggregation topic with a new attitude: 1) the increasing availability of data on elementary units over sequential time periods and 2) the more and more rapid rise in computing power. Increase of data accessibility allowed discovery and application of new methodology like Dynamic Factor models or Panel Data models. These are two of potential partial solution for treatment of bias resulting from longitudinal and temporal aggregation¹⁶ because are considered, in the same model, statistical measures typical of individual events and aggregated behaviours.

In the last years there has been a proliferation of works on temporal aggregation issue. Limiting investigation to univariate and linear stochastic models, conclusions are that time frequency change affects parameters value and structure of autoregressive and moving average components, represented by filters in lag operator.

¹⁶See Blundell (1988) for an empirical reference.

The aggregation process in the calculation of composite indexes involves three steps in succession: data normalisation, identification of a weighting structure and determination of functional form (linear vs nonlinear). The system of weights could be subjective, objective or “neutral” and according to temporal domain, weights could be static or dynamic. Anyway selection of optimal aggregation criteria is always a critical issue, because depends by data correlation configuration and by desired compensability degree between indicators.

Finally, a satisfactory aggregation process, usually necessary for analysis on extended areas or large groups, should be able to suggests the use of statistical techniques, a weighting system or a number of statistics that reshapes the economic and social representation in an unbiased way and should not include final users (policy-makers, public opinion, etc.) toward misleading evaluations and decisions.

References

- Aigner D. J. and Goldfeld S. M. (1974), "*Estimation and prediction from aggregate data when aggregates are measured more accurately than their components*", *Econometrica*, vol. 42, no. 1, pp. 113-134
- Amemiya T. and Wu R.Y. (1972), "*The effect of aggregation on prediction in the autoregressive model*", *Journal of the American Statistical Association*, vol. 67, no. 339.
- Baltagi B. (2008), "*Econometric Analysis of Panel Data*", John Wiley & Sons, 4th edition.
- Barker T. and Pesaran M.H. (1990), "*Disaggregation in econometric modelling*", Routledge, London and New York.
- Blundell R. (1988), "*Consumer behavior: theory and empirical evidence – a survey*", *The Economic Journal*, no. 98, pp 16-65.
- Blundell R and Stoker T.M. (2005), "*Heterogeneity and aggregation*", *Journal of Economic Literature*, vol. 43, no. 2, pp. 347-391.
- Brewer K.R.W. (1973), "*Some consequences of temporal aggregation and systematic sampling for ARMA and ARMAX models*", *Journal of Econometrics*, no. 1, pp 133-154.
- Bee Dagum E. and Cholette P.A. (2006), "*Benchmarking, temporal distribution and reconciliation methods for time series*", Springer.
- Forni M. and Lippi M. (1997), "*Aggregation and the microfoundations of dynamic macroeconomics*", Oxford Clarendon Press.
- Forni M., Lippi M. (2001), "*The generalized dynamic factor model: representation theory*", *Econometric Theory*, vol. 17.
- Forni M., Hallin M., Lippi M., Reichlin L. (2000), "*The generalized dynamic factor model: identification and estimation*", *The Review of Economics and Statistics* vol. 82, pp 540-
- Geweke J. (1977), "*The dynamic factor analysis of economic time series*", in Aigner D.J. And Goldberger A.S. (eds.), *Latent variables in socioeconomic models*, Amsterdam, North Holland, pp 365-383.
- Gorman W.M. (1959), "*Utility and aggregation*", *Econometrica*, vol. 27, no. 3, pp 469-481.
- Granger C. (1990), "*Aggregation of time series variables: a survey*", in *Disaggregation in Econometric Modelling* (edited by Barker, T. and Pesaran, M. H.), pp. 17-34.
- Grunfeld Y. and Griliches Z. (1960), "*Is aggregation necessarily bad?*", *The Review of Economic and Statistics*, vol. 17, no. 1.
- Gupta K.L. (1971), "*Aggregation bias in linear economic models*", *International Economic Review*, no.12, pp 293-305

- Hoffman, A., Giovannini E., Nardo M., Saisana M., Saltelli A. and Tarantola S. (2008), “*Handbook on Constructing Composite Indicators: Methodology and User Guide*”, OECD Statistics Working Paper, Paris.
- Hsiao C. (2002), “*Analysis of panel data*”, Cambridge University Press, 2nd edition.
- Kirman A. and Zimmermann J.B. (2001), “*Economics with interacting agents*”, Springer Verlag, Heidelberg.
- Koopmans T. (1947), “*Measurement without theory*”, The Review of Economic Statistics, vol. 29, no. 3, pp. 161-172.
- Lee K.C., Pesaran M.C. and Pierser R.G. (1990), “*Testing for aggregation bias in linear models*”, The Economic Journal (Supplement), vol. 100, pp 137-150.
- Leontief W. (1947), “*Introduction to a theory of the internal structure of functional relationship*” *Econometrica*, vol. 15, no. 4, pp 361-373.
- Lippi M. and Forni M. (1990), “*On the dynamic specification of aggregated models*”, in *Disaggregation in econometric modelling* (edited by Barker T. and Pesaran M.H.), pp 35-72.
- Lucas R.E. (1973), “*Some international evidence on output-inflation trade-offs*”, *American Economic Review*, no. 63, pp 326-334.
- Malinvaud E. (1956), “*L’aggregation dans le modles conomiques*”, *Cahiers du seminaire d’econometrie*, no. 4, pp 69-146.
- Marcellino M. (2006), “*Leading indicators*”, in *Handbook of economic forecasting* (edited by Elliot G., Granger C. and Timmermann A.), vol. 1, pp. 879-960. Amsterdam, Elsevier.
- Munda G. and Nardo M. (2005), “*Non-compensatory composite indicators for ranking countries: a defensible setting*”, Institute for the Protection and Security of the Citizen.
- OECD (1987), “*OECD Leading Indicators and Business cycles in member countries, Sources and Methods*”, no. 39.
- Orcutt G. H., Watts H. W. and Edwards J. B (1968), “*Data aggregation and information loss*”, *American Economic Review*, vol. 58, no. 4.
- Pektovic A. and Veredas D. (2009), “*Aggregation of linear models for panel data*”, ECARES working paper no. 12, Bruxelles.
- Pesaran M.C., Pierser R.G. and Kumar M.S. (1989), “*Econometric analysis in the context of linear prediction models*”, *Econometrica*, vol.57, no.4, pp 861-888.
- Saisana M. and Tarantola S. (2002), “*State-of-the-art report on current methodologies and practices for composite indicator development*”, EUR 20408 EN, European Commission-JRC: Italy.
- Sargent T.J. And Sims C.A. (1977), “*Business-cycle modelling without pretending to have too much a priori economic theory*”, in Sims C.A. (ed.),

- New methods of business cycle research*, Minneapolis Federal Reserve Bank of Minneapolis, pp 45-109.
- Sen A.K. (1985), "*Commodities and capabilities*", North Holland, Amsterdam.
- Sen A.K. (1992), "*Inequality reexamined*", Russell Sage Foundation, New York.
- Stiglitz J.E., Sen A. and Fitoussi J.P. (2009), "*Report by the commission on the measurement of economic performance and social progress*", Paris, CMESP 2009.
- Stoker T. M. (1984), "*Completeness, distribution restrictions, and the form of aggregate functions*", *Econometrica*, vol. 52, no. 4, pp. 887-907.
- Theil H. (1954), "*Linear aggregation of economic relations*", Amsterdam, North Holland
- Timmermann, A. (2006). "*Forecast combinations*", in *Handbook of economic forecasting* (edited by Elliot G., Granger C. and Timmermann A.), vol. 1, pp. 135-196. Amsterdam, Elsevier.
- United Nations (2001), "*Human development report*", United Kingdom, Oxford University Press.
- Weiss A. (1984), "*Systematic sampling and temporal aggregation in time series models*", *Journal of Econometrics*, no. 26, pp 271-281.
- Zellner A. (1962), "*An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias*", *Journal of the American Statistical Association*, vol. 57, no. 298.