A Generalized Composite Index based on Non-substitutability of Individual Indicators

Matteo Mazziotta and Adriano Pareto

Abstract Composite indices for ranking territorial units are more and more popular in a variety of economic, social and policy domains. The literature offers a wide variety of aggregation methods, all having pros and cons. In this paper, we propose a new and alternative composite index denoted as MPI (Mazziotta-Pareto Index) which, starting from a linear aggregation, introduces penalties for the units with "unbalanced" values of the indicators. As an example of application, we consider the set of individual indicators of the Technology Achievement Index (TAI) and we present a comparison between simple arithmetic mean and MPI.

Key words: dimensionality reduction, ranking

1 Introduction

Social and economic phenomena, like inequality, poverty, development, wellness, malnourishment etc. have been measured, in the past, principally from an unidimensional point of view, that is using only one variable. The more recent literature tends to consider these phenomena as multidimensional or complex since they are characterized by the combination of different variables. The measurement of complex phenomena is a difficult and dangerous operation since it requires simplifications that are inherently somewhat arbitrary, is always constrained by limited resources and time, inevitably involves competing and conflicting priorities, and rests on a foundation of values preferences that are typically resolved by pragmatic considerations, disciplinary biases and measurement traditions. Nevertheless, it is possible to combine, in a dependent way, both the choice of the variables that have to represent the phenomenon and the choice of the "best" aggregation function in order to lose not much statistical

---

Matteo Mazziotta, Italian National Institute of Statistics; email: mazziott@istat.it

Adriano Pareto, Italian National Institute of Statistics; email: pareto@istat.it

information. The aim of this paper is to present a new composite index suitable in the case in which the variables are non-substitutable, that is they have all the same weight (importance) and a compensation among them is not allowed. In the section 2, the steps to implement a composite index are presented; in the section 3, the MPI (Mazziotta-Pareto Index) method and its mathematical properties are shown; finally, in the section 4, an application to real data is proposed.

2 Steps towards the synthesis of indicators

In the scientific literature, there are many studies, by eminent authors, concerning the use of composite indices in order to measure complex, economic and social, phenomena about geographical areas. The main problems, in this approach, are the availability of the data, the choice of the more representative indicators and their treatment in order to compare them (normalization). An other very important step is the definition of the synthesis measure because the aggregation function is the tool used to "add" the normalized indicators; in this delicate phase, from a methodological point of view, the choices of the researcher assume a fundamental role. In fact, in order to define a function, there are many possible choices such as multivariate techniques or distance measures or linear and non linear functions.

It is possible, shortly, to individuate the main steps to tackle: a) defining the phenomenon to measure; b) selecting a group of elementary indicators, usually expressed in different unit of measurement; c) normalizing elementary indicators to make them comparable; d) aggregating the normalized indicators by composite indices (mathematical functions).

For this approach, obviously, there are many problems such as finding data, losing information and researcher arbitrariness for: (i) the selection of indicators, (ii) the normalization, (iii) the choice of the aggregation function. In spite of these problems, how mentioned in the introduction, the advantages of this approach are clear and they can be summarized in: a) unidimensional measurement of the phenomenon; b) immediate availability; c) simplification of the geographical data analysis.

Many works and analysis have won over the critics and the scientific community concluded that it is impossible to obtain a "perfect method" where the results are universally efficient. On the contrary, the data and the specific targets of the work must, time by time, individuate the "best method" in terms of robustness, reliable and consistent solutions.

3 The composite Index

The MPI wants to supply a composite measure of a set of indicators that are considered "non-substitutable", i.e., all the dimensions of the phenomenon must be "balanced" (Mazziotta and Pareto, 2007).

It is designed in order to satisfy the following properties:

- normalization of the indicators by a specific criterion that delete both the unit of measure and the variability effect (Delvecchio, 1995);

- synthesis independent from an "ideal unit", since a set of "optimal values" is arbitrary, non-univocal and it can vary during the time;
- simplicity of computation.

These properties can be satisfied by the following approach. It is known, the distributions of different indicators, measured in different way, can be compared by the transformation in standardized deviations (Aureli Cutillo, 1996). Therefore, it is possible to convert the individual indicators to a common scale with a mean M=100 and a standard deviation S=10: the obtained values will range approximately in the interval (70; 130)[1]. In this type of normalization the "ideal vector" is the set of the mean values and it is easy individuate both the units that are over the mean (values greater than 100) and the units that are under the mean (values less than 100).

In this context, we introduce a penalty coefficient that is function, for each unit, of the indicators variability in relation to the mean value ("horizontal variability"): this variability can be measured by the coefficient of variation. The proposed approach penalizes the score of each unit (the mean of the standardized values) with a quantity directly proportional to the "horizontal variability". The purpose is to favour the units that, mean being equal, have a greater balance among the indicators values (Palazzi, 2004).

This method allows to obtain a "robust" measure and less "sensitive" to inclusion or exclusion of individual indicators (Mazziotta C. et al., 2008).

The steps for computing the MPI are the following.

1) Normalization

Let $\mathbf{X}=\{x_{ij}\}$ be the matrix with n rows (statistical units) and m columns (individual indicators) and let M_{x_j} and S_{x_j} denote the mean and the standard deviation of the j-th indicator:

$$\mathrm{M}_{x_j} = \frac{\sum_{i=1}^{n} x_{ij}}{n}; \qquad \mathrm{S}_{x_j} = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij}-\mathrm{M}_{x_j})^2}{n}}.$$

The standardized matrix $\mathbf{Z}=\{z_{ij}\}$ is defined as follows:

$$z_{ij} = 100 \pm \frac{(x_{ij}-\mathrm{M}_{x_j})}{\mathrm{S}_{x_j}} 10 \qquad (1)$$

where the sign ± depends on the relation of the j-th indicators with the phenomenon to be measured (+ if the individual indicator represents a dimension considered positive and – if it represents a dimension considered negative).

2) Aggregation

Let cv_i be the coefficient of variation for the i-th unit:

[1] On the basis of Bienaymé-Cebycev theorem, the terms of the distribution within the range (70; 130) are at least 89% of the total of terms.

$$\mathrm{cv}_i = \frac{\mathrm{S}_{z_i}}{\mathrm{M}_{z_i}}$$

where:

$$\mathrm{M}_{z_i} = \frac{\sum_{j=1}^{m} z_{ij}}{m}; \qquad \mathrm{S}_{z_i} = \sqrt{\frac{\sum_{j=1}^{m}(z_{ij} - \mathrm{M}_{z_i})^2}{m}}.$$

Then, the generalized form[1] of MPI is given by:

$$\mathrm{MPI}_i^{+/-} = \mathrm{M}_{z_i}(1 \pm \mathrm{cv}_i{}^2) = \mathrm{M}_{z_i} \pm \mathrm{S}_{z_i}\mathrm{cv}_i \qquad (2)$$

where the sign of the penalty (the product $\mathrm{S}_{z_i}\mathrm{cv}_i$) depends on the kind of phenomenon to be measured and then on the direction of the individual indicators.

If the composite index is “increasing” or “positive”, i.e., increasing values of the index correspond to positive variations of the phenomenon (e.g., the socio-economic development), then MPI^- is used. Vice versa, if the composite index is “decreasing” or “negative”, i.e., increasing values of the index correspond to negative variations of the phenomenon (e.g., the poverty), then MPI^+ is used.

The MPI^+ and MPI^- can be written in compact form as follows:

$$\mathrm{MPI}_i^+ = \frac{\sum_{j=1}^{m} z_{ij}^2}{\sum_{j=1}^{m} z_{ij}} \qquad \text{and} \qquad \mathrm{MPI}_i^- = \frac{2}{m}\sum_{j=1}^{m} z_{ij} - \frac{\sum_{j=1}^{m} z_{ij}^2}{\sum_{j=1}^{m} z_{ij}}$$

where z_{ij} is given by (1). The MPI_i^+ is a concave function of the generic z_{ik} ($k = 1, \ldots, m$), while the MPI_i^- is a convex function of z_{ik}.

Given the matrix $\mathbf{X}=\{x_{ij}\}$, the generalized index has the following properties:

i. The MPI^+ of the *i*-th unit is greater or equal than the MPI^- of the same unit, that is:

$$\mathrm{MPI}_i^+ \geq \mathrm{MPI}_i^-.$$

In particular, $\mathrm{MPI}_i^+ = \mathrm{MPI}_i^-$ iff $\mathrm{S}_{z_i} = 0$.

ii. The MPI^+ and the MPI^- of the *i*-th unit are linked by the relation:

$$\mathrm{MPI}_i^- = 2\mathrm{M}_{z_i} - \mathrm{MPI}_i^+.$$

iii. Given two units *i* and *h*, with $\mathrm{M}_{z_i} = \mathrm{M}_{z_h}$, we have:

[1] It is a generalized form since it includes “two indices in one”.

$$\mathrm{MPI}_i^- > \mathrm{MPI}_h^- \quad \text{iff} \quad \mathrm{S}_{z_h} > \mathrm{S}_{z_i} \, ;$$

$$\mathrm{MPI}_i^+ > \mathrm{MPI}_h^+ \quad \text{iff} \quad \mathrm{S}_{z_i} > \mathrm{S}_{z_h} \, .$$

iv. Given two units *i* and *h*, with $\mathrm{M}_{z_i} > \mathrm{M}_{z_h}$, we have:

$$\mathrm{MPI}_i^- > \mathrm{MPI}_h^- \quad \text{iff} \quad \mathrm{M}_{z_i} - \mathrm{M}_{z_h} > \mathrm{S}_{z_i}\mathrm{cv}_i - \mathrm{S}_{z_h}\mathrm{cv}_h \, ;$$

$$\mathrm{MPI}_i^+ > \mathrm{MPI}_h^+ \quad \text{iff} \quad \mathrm{M}_{z_i} - \mathrm{M}_{z_h} > \mathrm{S}_{z_h}\mathrm{cv}_h - \mathrm{S}_{z_i}\mathrm{cv}_i \, .$$

v. Let $\mathrm{r}_{x_j x_k}$ be the linear correlation coefficient between the *j*-th and the *k*-th indicator; if $\mathrm{r}_{x_j x_k} = 1$, for each *j* and *k* with $j \neq k$, then:

$$\mathrm{MPI}_i^+ = \mathrm{MPI}_i^- = \mathrm{M}_{z_i} \, .$$

This property derives from the fact that, for the *i*-th unit, we have $z_{ij} = z_{ik}$ for $j \neq k$.

The v. is very interesting because it shows the relation between the behaviour of the MPI and the structure of the correlations among the individual indicators.

In general, the greater is the discordance among the indicators and the higher is the horizontal variability "induced" in each unit, with consequent increasing of the penalty and hence of the difference between MPI and arithmetic mean.

4 An application to real data

In this section, an application of MPI to the variables of the Technology Achievement Index (TAI) is presented; see OECD (2008) for more details. The TAI is a composite measure of technological progress that ranks countries on a comparative global scale.

In table 1 is reported the indicators list of the TAI.

Table 1: List of individual indicators of the Technology Achievement Index (TAI)

Indicator	Definition	Unit
PATENTS	Number of patents granted to residents, to reflect the current level of invention activities (1998)	Patents granted per 1,000,000 people
ROYALTIES	Receipts of royalty and license fees from abroad per capita, so as to reflect the stock of successful innovations of the past that are still useful and hence have market value (1999)	US $ per 1,000 people
INTERNET	Diffusion of the Internet, which is indispensable to participation in the network age (2000)	Internet hosts per 1,000 people
EXPORTS	Exports of high and medium technology products as a share of total goods exports (1999)	%
TELEPHONES	Number of telephone lines (mainline and cellular), which represents old innovation needed to use newer technologies and is also pervasive input to a multitude of human activities (1999)	Telephone lines per 1,000 people (log)
ELECTRICITY	Electricity consumption, which represents old innovation needed to use newer technologies and is also pervasive input to a multitude of human activities	kWh per capita (log)
SCHOOLING	Mean years of schooling (age 15 and above), which represents the basic education needed to develop cognitive skills (2000)	Years
UNIVERSITY	Gross enrolment ratio of tertiary students enrolled in science, mathematics and engineering, which reflects the human skills needed to create and absorb innovations (1995-1997)	%

It is easy to note that the variables have all the same importance in order to represent the technological progress. So, we assume that the variables of the TAI have the property of non-substitutability, or rather it is very important that there is not a compensation among the variables; they have the same weight and a balanced distribution of the values is desirable.

The aim of this application is to compare the MPI and the arithmetic mean in order to test the action of the penalty function. One characteristic of the mean is to offset the values of the variables so if a value is high and an other is low the mean distributes the total amount and a clear consequence is the loss of statistic information. In the case of non-substitutability of the indicators, the use of the mean, like composite index, is wrong because the variables must not to be compensated. The MPI is a solution to this problem because it adjusts the mean with a penalty function that penalizes the geographical areas with an unbalanced distribution of the values of the indicators.

In table 2, a comparison between the simple arithmetic mean and the MPI^- is reported.

Table 2: TAI country rankings based on simple arithmetic mean and MPI^-

Country	Mean		MPI-		Difference of rank
	Value	Rank	Value	Rank	
Finland	112.67	1	112.00	1	0
United States	111.81	2	111.51	2	0
Sweden	111.36	3	111.09	3	0
Japan	109.52	4	107.84	4	0
Korea, Rep. of	107.80	5	106.42	5	0
Netherlands	106.25	6	105.66	6	0
United Kingdom	105.34	8	104.96	7	1
Canada	105.14	9	104.77	8	1
Norway	105.80	7	104.31	9	-2
Germany	103.75	12	103.62	10	2
Singapore	103.95	11	103.07	11	0
Ireland	103.17	13	102.93	12	1
Australia	104.13	10	102.91	13	-3
Belgium	102.39	14	102.30	14	0
Austria	101.71	16	101.60	15	1
France	101.44	17	101.27	16	1
New Zealand	101.99	15	101.01	17	-2
Israel	100.58	18	100.49	18	0
Spain	98.33	19	98.09	19	0
Italy	98.29	20	98.04	20	0
Czech Republic	97.64	21	97.48	21	0
Slovenia	97.30	22	97.13	22	0
Hungary	97.28	23	96.99	23	0
Slovakia	96.62	24	96.50	24	0
Portugal	95.42	25	95.09	25	0
Poland	94.27	26	94.09	26	0
Argentina	92.82	27	92.58	27	0
Romania	92.05	29	91.74	28	1
Mexico	92.29	28	91.62	29	-1
Brazil	88.73	30	88.49	30	0
China	87.39	31	86.90	31	0
Colombia	86.83	32	86.57	32	0
Bolivia	85.96	33	85.29	33	0
Mean absolute difference of rank					**0.48**
Spearman's rank correlation index					**0.995**

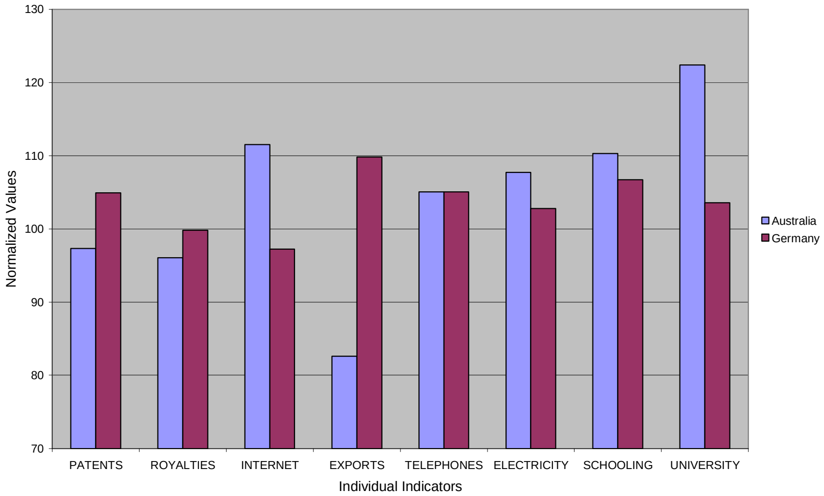
The rankings of 33 countries (ordered by MPI^-) do not present relevant surprise, from a geo-economic point of view, but this is not the aim of the work.

Let us note that the mean absolute difference of rank is equal to 0.48, i.e., each country changes, on average, less than half position. The Spearman's rank correlation index is 0.995 so that there is not a great difference between the mean and the MPI.

The divergence between the two methods is due to the penalty function, in fact the MPI is a mean adjusted on the basis of the "horizontal variability" of standardized indicators. On the whole, the greater differences among the methods are in the middle part of the ranking where, probably, the countries have a higher variability of the indicators. It is noteworthy that there are not differences of rank in the top and in the bottom of the ranking: the countries on the extremes of the table are the same in the two approaches since the penalty function does not correct the results of the countries with technologies more and less advanced.

In figure 1 is displayed a comparison between the distributions of the Australia and Germany normalized individual indicators.

Figure 1: Individual indicators of the TAI: a comparison of countries



The Germany has a balanced distribution and in fact it passes from the position 12 (mean) to the position 10 (MPI) getting ahead to countries that are penalized by the MPI. One of these is the Australia that occupies the position 10 in the mean ranking and the position 13 in the MPI ranking. The figure shows, in fact, that the Australian distribution of the normalized indicators is unbalanced; for example the variables "Internet" and "University" present high values and the variable "Exports" a very low value. In this case the penalty function penalises the country causing a decrement of three ranks.

5 Conclusions

The change from unidimensional to multidimensional measurement is without any doubt an important theoretical progress and presents many advantages for policymaking. However, there is also a flip side, because multidimensional measurement implies many theoretical, methodological and empirical problems. The international literature on composite indices offers a wide variety of aggregation methods. The most used are additive methods, but they imply requirements and properties which are often not desirable or difficult to meet, sush as a full substitutability among the components of the index.

Considering the desirable properties that a composite index should have, we proposed a new and alternative composite index, denoted as MPI (Mazziotta-Pareto Index). The MPI is based on the property of non-substitutability of indicators and can be validly applied to different scientific contexts, since it is independent from the versus and the range of the elementary indicators.

Moreover, the use of a penalty for unbalanced values of the indicators allows to distinguish different situations which otherwise would not be reflected in a composite index based on the simple arithmetic mean.

References

1. Aureli Cutillo, E.: Lezioni di statistica sociale. Parte seconda, sintesi e graduatorie. CISU, Roma (1996)
2. Delvecchio, F.: Scale di misura e indicatori sociali. Cacucci Editore, Bari (1995)
3. Mazziotta, M., Pareto, A.: Un indicatore sintetico di dotazione infrastrutturale: il metodo delle penalità per coefficiente di variazione. In: Lo sviluppo regionale nell'Unione Europea – Obiettivi, strategie, politiche. Atti della XXVIII Conferenza Italiana di Scienze Regionali. AISRe, Bolzano (2007)
4. Mazziotta, C., Mazziotta, M., Pareto, A., Vidoli, F.: La costruzione di un indicatore sintetico di dotazione infrastrutturale: metodi e applicazioni a confronto. In: Conoscenza, sviluppo umano e territorio. Atti della XXIX Conferenza Italiana di Scienze Regionali. AISRe, Bari (2008)
5. Palazzi, P.: Lo sviluppo come fenomeno multidimensionale. Confronto tra l'ISU e un indice di sviluppo sostenibile. Moneta e Credito, 227 (2004)
6. OECD: Handbook on Constructing Composite Indicators. Methodology and user guide. OECD Publications, Paris (2008)