IMG

Institut de Mathématique Gauss
9 Carré F.X. Lemieux
Lévis, Québec, Canada G6W 1H2
418 837 1139
img@globetrotter.net

Multidimensional Poverty

Theory

LOUIS-MARIE ASSELIN

This document has been produced with the financial support of the MIMAP programme, IDRC, Canada.

Composite Indicator of Multidimensional Poverty

Louis-Marie Asselin

IMG

Composite Indicator of Multidimensional Poverty[1]

LOUIS-MARIE ASSELIN, CECI, June 2002

Introduction

The technical problem we are faced to originates from the multidimensionality of the poverty concept, now universally accepted. This multidimensionality depends on the definition given to poverty, for which there is not a unique formulation, but usually a large overlapping between those given here and there. We like to share this one, which expresses our institutional views on poverty:

> *Poverty consists in any form of inequity, source of social exclusion, in living conditions essential to human dignity. These living conditions correspond to the capabilities of individuals, households and communities to meet their basic needs in the following dimensions:*
> - *nutrition*
> - *primary education*
> - *primary health care*
> - *sanitation*
> - *safe water*
> - *housing*
> - *income*
> - *community participation.*

The particular faces of poverty become particularly meaningful if we consider that, at the individual level, the different dimensions take specific forms according to gender and to age-group. From this point of view, looking at individual poverty appears as the most operational way of implementing multidimensionality. Moreover, and more importantly, looking at individual poverty seems a natural way of exploring poverty dynamics, perceived as life-cycle poverty status differentiated according to gender.

Thus, per se, multidimensional poverty is a richer concept than the traditional income approach. In addition, the technical difficulties of income measurement, especially in developing countries, have been an important incitative for looking at other poverty measures. "In the vast majority of African countries, we remain unable to make inter-temporal comparisons of poverty due the unavailability of data. And where survey data are available at more than one point in time, the determination of changes has proven problematic. First, survey designs change. It is now well established that differences in recall periods, changes in the survey instrument (e.g., the number and choice of item codes listed), and even the nature of interviewer training, can have large systematic effects on the measurement of household expenditures. Compounding this problem, intertemporal comparisons of money-metric welfare are only as precise as the deflators used. Consumer price indices are often suspect in Africa, due to weaknesses in data collection and related analytical procedures. Thus, relying on official CPIs is often precarious, at best. Alternatives such as deriving price indexes from unit values, where quantity and expenditure data are collected, also have serious drawbacks"[2]. The same comment certainly applies to most low-income countries.

---

[1] We thank the International Development Research Center (IDRC) for having generously funded this work. We thank also Lise Brochu and Anyck Dauphin who have contributed to the bibliographical research.

[2] Sahn David E. and David C. Stifel (1999), p. 1.

On the other hand, it must be recognized that the income-measure of poverty presents a great technical advantage: it is unidimensional, and thus allows for a complete ordering of households according to the poverty level. This property is very important for targeting policies and programs, for poverty mapping, data aggregation and sophisticated analysis. That's why there is a strong request for retrieving a similar property with multidimensional poverty. There are many proposals coming out of current research work on this issue. As a well-known example, there is the set of human development and human poverty indices developed and published by UNDP. The criticims addressed to these numerous proposals rely generally on the arbitrary choice of weights and functional forms used in aggregating the primary poverty indicators.

Our sole purpose here is to try to identify some basic justification for the aggregation of multiple poverty indicators into a composite one, starting first with a critical review of the literature. Elimination of the arbitrary in the definition of a composite indicator is the focus of the whole work presented here.

1. Problem definition

There is a population U of N elementary population units U_i. On each unit, K primary indicators I_k are measured, K>1. These indicators are possibly heterogenous in their nature:

- quantitative indicator, e.g. household income, number of bicycles, etc.
- qualitative or categorical
  - ordinal, e.g. level of education, etc
  - non ordinal, e.g. occupation, geographical region, etc.

The first two categories, quantitative and qualitative ordinal, having both an ordinal structure, would refer to **direct** poverty indicators, i.e. a level of deprivation in a specific life condition, while the last category, qualitative non ordinal, would refer to an **indirect** poverty indicator, usually as a cause or an effect of poverty, or more simply as a characteristic allowing to identify poverty groups (e.g. ethnicity). The relevance of mixing ordinal and non ordinal indicators into a composite indicator will be left open for the moment. Let's remark that all these indicators are or can be expressed numerically, the number being a fully significant one in the case "quantitative", and a non significant one in the case "qualitative", where it is simply a numerical code, keeping obviously its ordinal meaning with a qualitative ordinal indicator.

The problem we try to solve here is **to define a unique numerical indicator C as a composite of the K primary indicators I_k, computable for each elementary population unit U_i, and significant as generating a complete poverty ordering of the population U**.

Let's observe that the term "elementary population unit" can refer as well to individuals and households as to villages, regions, countries.

For the discussion, it's important to clarify the terminology regarding the three concepts of **poverty indicator**, **poverty measure** and **poverty index**. Let's I_{ik} be the value of indicator I_k for the elementary population unit i. I_{ik} is then properly a **poverty indicator** value. The value I_{ik} can be transformed as $g_k(I_{ik})$, with the function g_k, to better reflect a poverty concept relative to indicator I_k. This is frequently the case especially with a quantitative indicator I_k to which is associated a poverty threshold (poverty line) z_k. In that case, well known transformations are $g_k(I_{ki})=(z_k- I_{ik})^{\alpha}$. Then, $g_k(I_{ik})$ is called a **poverty measure** value, again defined on each elementary population unit. In the particular case where the function g_k is the identity function, the poverty indicator and the poverty measure are the same. Finally, poverty measure values can be aggregated over the units for the whole population U, as $W_k\{g_k(I_{ik}), i=1, N\}$. Then W_k is called a **poverty index** relative to the indicator I_k for the population U. Obviously, this index W_k can be defined on sub-populations.

It is important to keep in mind that the three concepts indicator, measure and index are **relative** to the definition given of elementary population unit, so that a household-based index, for example, can be considered as a village indicator when considering a population of villages, and so on.

A **composite poverty indicator** C takes the value $C_i(I_{ik}, k=1,K)$ for a given elementary population unit U_i. This is what we want to define.

2. Review of literature

An interesting review of literature is given in E. Maasoumi (1999). We must first distinguish between the litterature addressing the issue of computing a composite poverty index from a multidimensional distribution of poverty indicators on a given population, and the literature aiming at defining a composite poverty indicator on each unit of the given population. The first type of this literature is well represented by F. Bourguignon and S.R. Chakravarty (1999). It relies usually on an axiomatic approach to the desired properties of the composite index and on a composite poverty measure referring to a given poverty threshold for each primary indicator. The implicit context is thus a set of quantitative indicators and the resulting index is usually relevant only for that type of indicators. As follows from the objective here aimed to, this first type of literature is not immediately relevant for us and is not explored further for the moment. It should be obvious, on the other hand, that solving the problem of the composite indicator does not preclude from computing a composite poverty index based on the composite indicator, according to the univariate theory of poverty indices. It is in fact a normal sub-product of our approach to multidimensionality, which could be seen as a two-step approach according to E. Maasoumi (1999) categorization. It should also be obvious that the computation of a composite index will rely, at least implicitly, on a composite indicator, and that's why we will eventually come back on this more restricted type of literature.

In the second type of literature, we can identify basically two trends, the first one being based on the concept of **entropy**, the second one on the concept of **inertia**. The entropy approach originates from the field of dynamic mechanics, as exploited in statistical information theory applied to stochastic processes. Engineering sciences appear as the historical bakground, and continuous structures constitute the most natural environment. Different metrics thus become a possible mathematical tool. Multidimensionality is first seen as coming from a sequence of unidimensional distributions. The inertia approach originates from the field of static mechanics. It is used in statistical stuctural analysis, an area historically developed mostly by social scientists, and particularly in psychometry. Discrete structures is a familiar environment, and general metrics is a standard tool. Multidimensionality consist here simply in a finite set of simultaneous distributions on a given population.

We here review briefly the literature from this two-trend perspective, in each case looking first to some theoretical papers and second to practical applications in the area of poverty analysis.

2.1 The entropy approach

2.1.1 Theoretical review

E. Maasoumi (1986) presents a short summary of the entropy approach to the composite indicator problem. Such a summary can also be found, in a larger context, in E. Maasoumi (1993), section 7. To understand the theoretical foundation of the approach, it is important to come back to the basic concepts of information theory, and especially the entropy concept. A. Rényi (1966) is a good account of the first developments of this theory, particularly for the meaning and justification of the reference entropy concept, the Shannon's entropy, and its extension to statistical distribution analysis, essential for our problem. Let's just recall that Shannon's entropy rely on the information function $-\log_2 p$, $p \in (0,1)$, and the entropy measure corresponds to the average number of binary digits required to characterize completely a random signal known to belong to a certain set E. This entropy measure is given by

$$I_n = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k} = -\sum_{k=1}^{n} p_k \log_2 p_k \qquad (1)$$

When applied to a statistical distribution, the Shannon's entropy is a measure of the degree of uncertainty contained in the distribution, the maximal uncertainty corresponding to the uniform distribution. In addition, it can be shown that the entropy *I* is a concave function on the space of distributions, and this property of concavity will be looked for with any further generalization of the entropy measure. Closely linked to this entropy measure, to its maximal property and to its concavity, derives naturally the concept of divergence between two distributions p and q: if we average p and q, thus coming "closer" to the uniform distribution, and exploiting the concavity property, we can write [Burbea and Rao (1982a)]

$$I_n\left(\frac{p+q}{2}\right) = \frac{I_n(p)+I_n(q)}{2} + J_n(p,q)$$, where $J_n(p,q)$ is nonnegative, and thus

$$J_n(p,q) = I_n\left(\frac{p+q}{2}\right) - \frac{I_n(p)+I_n(q)}{2}$$ can be taken as as a measure of distance between the two distributions *p* and *q*. It is called the **Jenzen divergence**. This type of divergence can be generalized to the weighted mean of k distributions:

$$J_n^{\pi}(p_1,\ldots,p_k) = I_n\left(\sum_{i=1}^{k} \pi_i p_i\right) - \sum_{i=1}^{k} \pi_i I_n(p_i)$$ called the **generalized Jenzen divergence**. It is essential to remark that this divergence measure is not in general a metric, since it may not satisfy the triangular inequality. On the other hand, we must observe here that $J_n(p,q)$ is a symmetric function of *p* and *q*.

The analysis of the case of a two-dimension joint distribution (*p,q*), where *p* and *q* refer to the marginal distributions, with respective sizes *m*, *n*, generates the important concept of cross-entropy, defined by

$$CE(p,q) = I_m(p) + I_n(q) - I_{mn}(p,q)$$

Cross-entropy has been recognized as a very important measure of the association between two random variables.

Building on these concepts, there are different ways of generalizing entropy and divergence measures, usually by an axiomatic approach where desirable properties of a given measure (entropy, divergence, etc) are formulated as a small set of axioms. Then a class of functions come out as the one satisfying the axioms. We will now review quickly three classes of entropy-related functions which are the most frequently met in the literature and the most relevant in our context. All these extended classes include the Shannon's entropy as a particular case.

A) The Havrda-Charvat structural α-entropy.
J. Havrda and F. Charvat (1967) specify four entropy axioms all satisfied by the Shannon's entropy and look for the class of entropy measures satisfying these axioms. The most specific among these axioms is the "Distribution Refinement" one, also called elsewhere the "Branching/Aggregation" axiom. It is formulated in this way:

$$I(p_1,...,p_{i-1},tp_i,(1-t)p_i,p_{i+1},....,p_n;\alpha)=I(p_1,...,p_{i-1},p_i,p_{i+1},....,p_n;\alpha)+$$
$$\lambda p_i^\alpha I(t,(1-t);\alpha)$$

for *i*= 1,*n* , α>0, and 0 < t <1. The normalization axiom *I*(1;α) = 0 implies that λ = 1. We see here the meaning of the parameter α: larger is α, smaller is the <u>relative</u> increased uncertainty generated by a given refinement of the initial distribution.

It is then shown that the structural α-entropy is unambiguously determined by

$$I(p_1,...,p_n;\alpha)=\frac{2^{\alpha-1}}{2^{\alpha-1}-1}\left(1-\sum_{i=1}^{n}p_i^\alpha\right) \text{ for } \alpha > 0 \text{ , } \alpha \neq 1, \quad (2)$$
$$I(p_1,...,p_n;1)=-\sum_{i=1}^{n}p_i\log_2 p_i \text{ .}$$

We thus see that the Shannon's entropy is the structural α-entropy for α = 1.

The case α = 2 is also interesting by its direct statistical interpretation. Let's E be a random experiment whose probabilty space is given the following set of disjoint events, with their associated probabilities:

$A_1, A_2, \ldots\ldots, A_k, \ldots.., A_n$
$p_1, p_2, \ldots\ldots., p_k, \ldots..., p_n$

If E is realized once only, with A_k as output, and the observed empirical frequency f_i is taken as an estimate of the different probabilities, we have
$f_i = 0$ for $i \neq k$, and $f_k =1$.
We thus have the n following positive (absolute) errors in the estimation on the probability distribution:
p_i for $i \neq k$, and $1-p_k$, whose summation is $1-p_k+\sum_{i\neq k}p_i$ =
$1-p_k+1-p_k = 2\times(1-p_k)$

The mathematical expectation of the error is then $2\sum_{k=1}^{n}p_k(1-p_k)=2\left(1-\sum_{k=1}^{n}p_k^2\right)$,

whic is precisely the α-entropy for α = 2.

Obviously, any entropy-based divergence between distributions can be applied to the α-entropy and then generates a specific measure dependent on the parameter α. In the case of the Jenzen divergence, we then obtain the α-divergence $J_{n,\alpha}(p,q)$, whose expression is given in Burbea and Rao (1982a)[3]:

[3] We should remark that sometimes in the literature, the Havrda-Charvat structural α-entropy is defined with the scale coefficient $\frac{1}{\alpha-1}$.

$$J_{n,\alpha}(p,q) = (\alpha-1)^{-1}\sum_i \left\{ \frac{1}{2}\left[p_i^\alpha + q_i^\alpha\right] - \left[\frac{p_i+q_i}{2}\right]^\alpha \right\}, \alpha \neq 1$$

It is easy to check that in the case $\alpha = 2$, we have:

$J_{n,2}(p,q) = \frac{1}{4}\sum_i (p_i - q_i)^2$, what is simply the Euclidean metric.

B) The Rényi β-information gain measure and derived β-entropy.

Rényi (1966) develops his approach to information theory by exploring the notion of reduced entropy obtained through a conditional distribution derived from an initial distribution *p*. He thus generates the more general concept of the information gain/loss obtained when we replace a given distribution *p* by another distribution *q*, quantity notated I(q||p). The concept is first developed with the Shannon's entropy, and it is then shown that there is a close connection between the cross entropy CE(p,q) and the information gain I(q||p). The cross entropy is the mathematical expectation of the information gain relative to the conditional distributions of p relative to q.

He then generalizes the information gain concept through an appropriate axiomatics, which determines uniquely the following class of information gain measures:

$$I_\beta(q \,\|\, p) = \frac{1}{\beta-1}\log_2\left(\sum_{k=1}^n \frac{q_k^\beta}{p_k^{\beta-1}}\right) \text{ for } \beta \neq 1, \qquad (3)$$

$$I_1(q \,\|\, p) = \sum_{k=1}^n q_k \log_2 \frac{q_k}{p_k} \;.$$

The case $\beta = 1$ corresponds to the information gain related to the Shannon's entropy, and is the limiting case of (3). We observe that the information gain is a kind of measure of divergence between two distributions, and that the class β is an <u>asymmetric</u> measure of divergence.
This information gain measure can be particularized to the case where the distribution *p* is taken as the uniform distribution, and is then generated another class of entropy measures for any distribution *p*, given by:

$$I_\beta(p) = \frac{1}{1-\beta}\log_2 \sum_{k=1}^n p_k^\beta \quad \text{for } \beta \neq 1 \qquad (4)$$

with $I_1(p) = \sum_{k=1}^n p_k \log_2 \frac{1}{p_k}$ as the limiting case, that is the Shannon's entropy.

C) The Shorrocks γ-entropy based inequality measure.

H. Theil (1967) has first observed that the Shannon's entropy (1), $I_n(y)$, where y represents the income shares in a population of n units, constitutes a natural measure of income equality, taking the maximal value $\log_2 n$ when every unit has the same income. The corresponding <u>inequality</u> measure is then taken as the difference between the maximal entropy (uniform distribution) and $I_n(y)$:

$$\log_2 n - I_n(y) = \sum_{k=1}^{n} y_k \log_2\left(\frac{y_k}{1/n}\right). \qquad (5)$$

We thus observe that (5) is the Rényi divergence measure $I_1(q||p)$ where we take q = y and p = {1/n}, the uniform distribution. It is called the Theil's first inequality measure or inequality index[4].

The pioneering work of Theil on entropy-based inequality indices has generated the search for larger classes of inequality indices, on the basis of axiomatics inspired by desirable properties in respect to redistribution of income within a given population. In particular, the requirement of additively decomposable inequality indices led to important results by A. F. Shorrocks (1980). He proved that the only admissible indices satisfying, among others, the decomposable additivity axiom, belong to the following class[5]:

$$I_\gamma(y) = \frac{1}{n}\frac{1}{\gamma(\gamma-1)}\sum_{i=1}^{n}\left[\left(\frac{y_i}{1/n}\right)^{\gamma} - 1\right] \text{ for } \gamma \neq 0,1 \qquad (6)$$

$$I_\gamma^1(y) = \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{1/n}\right)$$

$$I_\gamma^0(y) = \sum_{i=1}^{n} 1/n \log\left(\frac{1/n}{y_i}\right)$$

Let's observe that (6) can be written like this:

$$I_\gamma(y) = \frac{1}{\gamma(\gamma-1)}\sum_{i=1}^{n} y_i\left[\left(\frac{y_i}{1/n}\right)^{\gamma-1} - 1\right]. \qquad (6')$$

Obviously, I_γ^1(y) is Theil's first inequality index, and I_γ^0(y) is another well-known Theil's index, his second inequality index. This γ-class of entropy-based inequality indices is called the class of **Generalized Entropy** indices.

What we want to highlight here is that this axiomatic development of inequality indices generates a class of divergence measures including as a particular case the Rényi's information gain measure $I_1(q||p)$. In fact, the case γ =1 corresponds to $I_1(q||p)$ where we take p= {1/n}, and the case γ =0 corresponds to $I_1(q||p)$ where we take q= {1/n}. The general case is quite close to the Rényi's β-class of divergence measures, as can be seen from (3), where we take p= {1/n}, and (6) or (6'). Again here, the γ-class of inequality indices is an asymmetric measure of divergence between a distribution y and the uniform distribution p= {1/n}.

E. Maasoumi (1986) relies on these developments of information theory to propose his enthropy approach to the composite indicator problem. He looks for a general interdistributional distance as a basis to derive the composite indicator C from an

[4] To be more precise, Theil and other authors use the natural logarithm in base e instead of base 2, and from now on, we will simply do not specify the base.

[5] We write the indices directly in terms of income shares instead of using the mean income μ, to keep more clearly the link with the theory of distributions.

optimisation criterion. Let's observe that the Generalized Entropy index (6') generates a divergence measure between any two distributions x and y if we subtitute a distribution x to the uniform distribution {1/n} appearing as denominator. This is precisely the divergence measure taken by Maasoumi as the distance between the composite indicator we are looking for, C, and any one of the primary indicators I_k, k= 1, K. We thus have:

$$D_\gamma(C, I_k) = \frac{1}{\gamma(\gamma-1)} \sum_{i=1}^{n} C_i \left[\left(\frac{C_i}{I_{ik}} \right)^{\gamma-1} - 1 \right] \quad \text{for } \gamma \neq 0,1 \tag{7}$$

and the Theil's first and second measures for γ = 1 and 0 respectively.

He then proposes to define the optimal indicator as the C that minimizes a weighted sum of the pairwise divergences, i.e the C that minimizes

$$D_\gamma(C, I; \delta) = \sum_{k=1}^{K} \delta_k \left\{ \frac{\sum_{i=1}^{N} C_i \left[\left(\frac{C_i}{I_{ik}} \right)^{\gamma-1} - 1 \right]}{\gamma(\gamma-1)} \right\} \tag{8}$$

where the δ_k are arbitrary weights on the divergence component relative to the indicator I_k, $\sum \delta_k = 1$[6].

By minimizing the divergence $D_\beta(C, I, \delta)$ for the function C, Maasoumi finds the following functional form for the composite indicator:

$$C_i = \left(\sum_{k=1}^{K} \delta_k I_{ik}^{-\gamma} \right)^{-1/\gamma} \quad \gamma \neq 0, -1 \tag{9}$$

We recognize here a CES function. For the two specific values γ =0, -1, the functional forms are

$$C_i = \prod_{k=1}^{K} I_{ik}^{\delta_k} \text{, for } \gamma = 0 . \tag{10}$$

$$C_i = \sum_{k=1}^{K} \delta_k I_{ik} \text{, for } \gamma = -1. \tag{11}$$

(See footnote 4 for the change in parametrization.)

Conclusion on the entropy approach

In view of our focus on the elimination of the arbitrary in the definition of a composite indicator, our short review of the entropy approach suggests some comments.

1. The whole context of entropy measures, including the associated divergence concept, refers to probability distributions, i.e. to numerical measures taking values in the interval (0,1). Thus, and as can be seen particularly from the divergence measure generated by the Generalized Entropy Index, the natural domain of application for our problem is a set of meaningful numerical indicators, i.e. of quantitative poverty indicators, expressed in terms of "shares", so that the individual value I_{ik} is in the interval (0,1). The money-metric

[6] The parametrization used by Maasoumi for the γ-class is slightly different from the Shorrocks's one, followed here until now. Maasoumi's parameter γ is Shorrocks's minus 1. From now on, we'll use Maasoumi's γ.

type of poverty indicators, once transformed in individual shares, appears as the domain of validity of a functional form like (9).

2. There is an important source of indetermination with the parametric nature of measures proposed by the entropy literature. On which basis should we chose the parameter value for the α-divergence, the β-information gain, the γ-Generalized Entropy indices? The most immediate criterion we can refer to presently seems to be: is there, in the literature, any strong argument for an appealing significant interpretation of some specific parameter values? Fortunately, yes. There is first the Shannon's information-theoretic meaning to entropy. All three classes of measures (α,β,γ) have Shannon's entropy as particular cases: $\alpha = \beta = \gamma = 1$ and $\gamma = 0$. Second, we have seen that the case $\alpha = 2$ has an immediate general statistical meaning relative to the fundamental frequency interpretation of a probability distribution. On the basis of Shannon's entropy, we are then inclined to limit our choice to the two functional forms $C_i = \prod_{k=1}^{K} I_{ik}^{\delta_k}$ (A) and $C_i = \sum_{k=1}^{K} \delta_k I_{ik}$ (B). In both cases, we are taken to linearity in the shares (see 1. above) I_{ik}. The case (A) is in fact a logarithmic linearity. It is essential here to remember that cases (A) and (B) have a very strong purely economic basis, since they correspond to the two well known Theil's inequality indices. On the basis of the Havrda-Charvat structural α-entropy for α = 2, we can again clarify our choice of a pameter value and of a functional form. In fact, since, for $\alpha = 2$, the Jenzen's divergence measure is the euclidean metric, by applying the Maasoumi's optimization criterion with this symmetric divergence measure instead of his asymmetric one, we immediately see that the optimal functional form is the centroid of the primary indicators I_k, i.e. the form (B): $C_i = \sum_{k=1}^{K} \delta_k I_{ik}$.

   There is thus a strong case for this linear form in the shares I_{ik}.

3. If the weighting approach is maintained for the optimization criterion, obviously there remains the problem of determining the weights δ_k in a non-arbitrary way. There is in fact an optimal system of weights for the functional form (B), as Maasoumi (1999) has himself observed: the principal component method. This is precisely the type of methods will be reviewed below (2.2, footnote 10).

2.1.2 Review of applications

2.2 The inertia approach

2.2.1 Theoretical review

By reviewing this part of the literature relevant to our subject, we here again keep the focus on looking for general methodologies allowing to eliminate as much as possible any kind of arbitrary in the computation of a composite indicator from multidimensional data on poverty. In this respect, Meulman (1992) presents a good unifying treatment of the set of methologies having something to do with the concept of inertia. Theoretical unity comes from developing a general equivalent, in terms of multidimensional scaling (MDS), of a large set of multivariate analysis techniques (MVA). Multidimensional scaling is precisely the problem we want to address with a composite indicator. Let's consider a multidimensional data set as a numerical matrix Z(n,m), where n is the number of observation units, or population units, also called more generally "objects", and m is the number of variables measured on each population unit or object. Each unit in the population space is then represented by a vector in

R^m. MDS techniques use a metric (distance) on the population space, by which are measured the dissimilarities between population units. Inertia in the population space is defined in reference to this metric, i.e to the between-unit distance. In contrast, the α-divergences based on the α-entropy (section 2.1.1) are semi-metrics between distributions[7], i.e. in the variate-space. Other β and γ divergence measure are not even symmetric. Said differently, the inertia approach and MDS techniques in the population space rely on within-distribution distances, while the entropy approach to the optimal composite indicator rely on between-distribution distances.

Briefly summarized, the inertia concept in the population space consists in seeing the n units as a cloud of points in the R^m space, with a mass (weight) associated to each point. The cloud has a centroid (weighted mean), and the weighted sum of distances to the centroid, according to the metric, gives the total inertia of the cloud of observation-points. The Euclidean distance is the standard one, but optimality processes can come out with a different distance and metric in R^m.

The most general MDS technique can be described as searching for an optimal low-dimension representation space, let's say a p-dimension space, $p \leq m$, where the projected cloud of population units keeps as much as possible of the inertia[8] of the source cloud, i.e., where the inertia-loss is minimized. Variants of this process respond to different analysis objectives. Inertia can here be seen as a measure of the information contained in the data set, and then a MDS technique, which can be seen as a data reduction technique, minimizes the unavoidable information-loss generated by representing the population units in a lower dimension space. As is shown in Meulman (1992), this process can be formalized and generalized as a problem of minimizing a general **STRESS** function, and to the most well known MVA techniques corresponds an objective function call a **STRAIN** function, quite similar to the STRESS function. These objective functions all refer to distances in the observation and representation spaces. Each population unit is then represented by a set of coordinates in the optimal p-dimension space called its **scores**. The score in dimension j, $j \leq p$, is a linear combination of the original observed variables, usually standardized.

According to the analysis objectives, this general theoretical framework can be applied to the most relevant MVA techniques for the problem of building a composite indicator. In this regard, we find in Meulman (1992) a review of **Principal Component Analysis (PCA)**, of **Generalized Canonical Analysis (GCA)**, and of **Multiple Correspondence Analysis (MCA)**, whose a particular case, statistically[9], is the classical **Correspondence Analysis (CA)**. MCA is itself a particular case of GCA. To differentiate between PCA and GCA, PCA looks at the m original variables as being one set of variables, while GCA distinguishes M sets of variables, the set J, J=1,M, consisting of m_J variables. The two basic general techniques, PCA and GCA, are quite old since they have been clearly defined and discussed in the years 1930. The last two, CA and MCA, appeared essentially in the seventies.

In PCA, the distance between population units is based on the usual Euclidean metric, and the representation space is built by a step-by-step process, by successive orthogonal projection on subspaces with increasing dimension. The first dimension defines the first principal component, that is "the normalized linear combination (that is, the sum of squares

[7] We recall that the special case $\alpha = 2$ generates a complete metric, the Euclidean one.
[8] Inertia defined strictly as above or in an analogous way preserving its geometrical and mechanical meaning.
[9] In fact, MCA has been developed, historically, as a generalization of CA. From the strict standpoint of computation techniques, in fact MCA can be seen as an application of the CA technique to a special kind of matrix, as will be explained farther.

of the coefficients being one) with maximum variance" [T.W. Anderson (1958), p. 272]. We could as well use here the term "inertia" in place of variance. Subsequent additional dimensions catch a decreasing portion of the total inertia, and the process can be repeated until a dimension p is reached where the total initial inertia is exhausted. "The principal components turn out to be the characteristic vectors of the covariance matrix. Thus the study of principal components can be considered as putting into statistical terms the usual developments of characteristic roots and vectors (for positive semidefinite matrices)" [Anderson (1958), p. 272]. The computation technique is thus an eigenvalue technique. The STRAIN function corresponding to the PCA optimization process is explicitly given by Meulman (1992), p. 544. With Anderson, it is essential to notice that "analysis in principal components is most suitable where all the components [original variables] … are measured in the same units. If they are not measured in the same units, the rationale of maximizing … is questionable" (loc. cit., p. 279). This condition will usually be fulfilled by standardizing the variables. Linearity of the principal components in the original variables and the requirement of same measurement units make this technique mostly appropriate with quantitative original variables. [10] H. Hotelling (1933) can been seen as the main founder of PCA.

As said above, GCA is concerned with the analysis of a data set where there are M sets (M>1) of observed variables. This type of analysis was first developed in the case M=2, by H. Hotelling (1936). In canonical correlation analysis, "we consider two sets of variates with a joint distribution, and we analyse the correlations between the variables of one set and those of the other set. We find a new coordinate system in the space of each set of variates in such a way that the new coordinates display unambiguously the system of correlation. More precisely, we find the linear combinations of variables in each set that have maximum correlation; these linear combinations are the first coordinates in the new systems. Then a second linear combination in each set is sought such that the correlation between these is the maximum of correlations between such linear combinations as are uncorrelated with the first linear combinations. The procedure is continued until the two new coordinate systems are completely specified" [Anderson (1958), p. 288]. GCA is an extension of this type of analysis to M sets, M>2. Again, this MVA technique with its general analysis motivation, can be described as an optimal inertia problem within the Meulman framework. The main interest of this framework is that it generates another metric, the Mahalanobis metric [Mahalanobis P.C. (1936)], as coming out of the inertia optimization process, and thus the arbitrary is again pushed away. With GCA, the low-dimension representation space is interpreted as the best reduced space highligting the closeness between sets of variables. Here too, the computation technique is an eigenanalysis on an appropriate matrix.

MCA (Multiple Correspondence Analysis), also called Homogeneity Analysis, can be "viewed … as a special case of generalized canonical analysis. Homogeneity analysis is applied to categorical, or nominal data, where each variable z_j is assumed to have k_j distinct categories. From the distance analysis viewpoint, the crucial part of the optimal transformation is performed beforehand: each of the m variables z_j is replaced by an $n \times k_j$ orthogonal binary matrix G_j … As in canonical distance analysis, a Mahalanobis metric is implicit … In this special case, since the *columns*[11] of G_j sum to 1, the Mahalanobis metric is equivalent to the χ^2 metric" [Meulman (1992), pp. 550-551]. Thus, here, the M sets of variables correspond to the m original categorical variables, once broken down each in their m_j constituent categories becoming as many dichotomous "variables". Thus, the problem of "same" measurement

[10] That's why it is seen as the MVA optimal technique to eliminate the arbitrary in determining the weights δ_k for the optimal functional form $C_i = \sum_{k=1}^{K} \delta_k I_{ik}$ obtained through the entropy analysis above, the quantitative variables being already homogenous, as shares.

[11] "rows" is erroneously written in the paper.

units for the variables involved in the computation disappears: all variables are binary. In fact, from the computational point of view, MCA is a CA analysis applied to the 0/1 matrix generated from the categorical indicators, what is precisely named the "indicator matrix"[12]. It is important to see that this transformation of the original categorical variables, a kind of "atomization", is in fact a process of freeing the population scores in the representation space from linearity in the original categorical variables. This fundamental property is well recognized in a good review of MCA, "Homogeneity Analysis: Exploring the Distribution of Variables and their Nonlinear Relationships", by W.J. Heiser and J.J. Meulman, in M. Greenacre and J. Blasius (1994). MCA originated in the sixties-seventies with the French school of statistics, under the leadership of J.P. Benzécri, its primary form being its application to the case M=2, under the name of CA (in French, "Analyse factorielle des correspondances"). A good account of CA is found in M. Greenacre and J. Blasius (1994), chap. 1, and also in J.-P. and F. Benzécri (1980).

In the context of these MVA techniques, we cannot avoid to mention a popular and well-known technique, Factor Analysis (FA). FA starts with a linear modeling of the set of observed variables, seen as linearly dependent variables from a small set of $p<m$ nonobservable variables called "common factors". Normality assumptions are required for optimal model estimation. The method is usually developed in two steps, the factor extraction and the rotation of axis.There are different methods for the extraction of factors, among which the principal component one. Also, there is a large variety of methods to rotate the axis. While being undoubtedly a very interesting method to generate hypothesis about the structure of the set of original variables, its parametric aspect, plus unclear principles for choosing among the rotation options (arbitrariness), in addition to the fact that its natural domain of application is certainly the case of quantitative variables, does not justify here a deeper exposition of this MVA method. Also, with the principal component method of factor extraction, the only one to provide an uniquely determined set of factor score (SPSS Applications Guide, version 10.0, p. 333), numerical results are similar to those generated by PCA, while the underlying rational is different.

Conclusion on the inertia approach

1. In contrast to the entropy analysis, we have here a non parametric approach to the composite indicator. There is thus much less space for the arbitrary in the search for a functional form to this indicator.

2. Within the large family of methods relying on the inertia concept, Multiple Correspondence Analysis (MCA), as a particular case of Generalized Canonical Analysis (GCA), has emerged as the most relevant one for our problem. That's why we will concentrate on a detailed description of MCA in section 3.

3. All these inertia methodologies are different optimal strategies to determine non arbitrary <u>weights</u> on a set of variables, the difference between each of them coming from the different analysis objectives. They thus attack directly the basic theoretical problem underlying the criticism addressed to operational proposals in view of aggregating a set of variables.

4. The <u>categorical weighting</u>, as obtained with MCA, consists in quantifying each primary qualitative indicator in a <u>non linear</u> way, thus without imposing, from the very beginning, any constraint on a functional form whose arguments are those indicators.

[12] We will see that, equivalently, MCA is a CA runned on the Burt matrix of all the 2-way contingency tables generated from the k primary indicators.

5. While with the entropy approach there is a required normalization consisting in redefining the quantitative positive indicators in terms of shares, here the normalization process consists in the basic 0-1 coding of all the primary categorical indicators, what is recognized universally as fundamental in any formalization of informational processus.

6. The inertia approach, more particularly the basic PCA technique, allows to complete the elimination of arbitrary met in the entropy approach for the weighting of the divergence measures associated to each primary indicator (see footnote 10).

2.2.2 Review of applications

3. Construction of a composite indicator from multiple categorical indicators

It has been clearly expressed in section 2 that Correspondence Analysis (CA) with its generalization to Multiple Correspondence Analysis (MCA) is a strong candidate as a non-arbitrary tool for the computation of a composite indicator, based on categorical (qualitative) indicators. Since there will always be a way of transforming a quantitative indicator into a categorical one, and since a multidimensional approach to poverty will always include a more or less large set of qualitative indicators, impossible to transform into quantitative ones, we will here describe extensively the methodology relying on the MCA approach. All the steps will be illustrated with a real data set generated by a large scale survey, the Vietnam Living Standard Survey # 1 (VLSS-1), conducted in 1992-1993 .

3.1Theoretical concepts of Multiple Correspondence Analysis

As explained in section 2.2, the basic concepts of Correspondence Analysis rely on those of factorial or principal component analysis, with a specific metric. Mains ideas are developed on the basis of the concept of **inertia**, and these ideas coming from static mechanics are profoundly geometrical in nature. These concepts and ideas are presented in Appendix I below, "Basic concepts of Correspondence Analysis and of its extension to Multiple Correspondence Analysis", with a very simple numerical example computed with the software SPSS 10.1. The main concepts are: profiles, χ^2-distance, clusters, centroid, inertia, projections, eigenvalue computation and inertia optimal disaggregation, principal axis and other factorial axis, scores, duality, weights as normalized scores on the first factorial axis. From all these concepts emerges, at the end, the technical definition of the composite indicator thus generated by an optimization process:

Definition

A composite indicator of multiple qualitative poverty indicators, defined as a set of categories, for different population units, is given by:

1. **computing the profile of the population unit relatively to these primary indicators,**
2. **applying to this profile the category-weights given by the normalized scores of these indicators on the first factorial axis coming out of the multiple correspondence analysis of the indicators.**

Specificity of the MCA case

The property of **duality**, central to correspondence analysis and to the above definition of the composite indicator, takes here a very simple and attractive simplification, due to the fact that the primary data matrix is a 0-1 matrix. With **K** indicators, the profile of a population unit is then a line-vector of numbers 1/K, and thus the value of the composite indicator is simply the average of the its category-weights, for any population unit. In the same way, with **N** population units, the profile of a category (k, j_k) is a column-vector of numbers $1/N^k_{j_k}$, where $N^k_{j_k}$ is the number of population units in the category (k, j_k), j_k = 1, J_k. Then, a category-weight is simply the average of the normalized scores of the population units belonging to this category[13].

Another important point to highlight here is that the MCA technique of quantifying categorical indicators (weighting process) frees us from the linearity hypothesis: there are unequal differences between the weights associated to the categories of a same indicator.

3.2 Implementation issues

We will go now through all the steps of computing a composite indicator and of using it with a real application, in view of facing different conceptual and technical issues arising in practice.

3.2.1 Data description

We use here a data set generated by the first Vietnam Living Standard Survey (VLSS-1), conducted in 1992-1993. There is a brief description of this survey in Annex 1, section 4. We use only the data of the Community questionnaire, administered to the 120 rural communes randomly drawn for the survey. This questionnaire, given in Appendix 2, has 24 pages and contains 142 questions. The computerized primary data come out as 23 Stata files numbered scr01 to scr23. The data dictionary is also given in Appendix 2.

From this questionnaire, we extracted a first subset of 39 indicators, described in Table 1 below. We find in this table 9 indicators in "economy and infrastructure", 10 in "education", 11 in "health" and 9 in "agriculture".

3.2.2 Data transformation and recoding

The only data transformation considered here is the transformation of quantitative variables into qualitative ones, that is the categorization of a numerical distribution. Among the 39 indicators of Table 1, some are primarily quantitative: #8, "Distance for public transportation", #10 and #11, "Fees", #12 and #17, "Enrolment rates", etc. This categorization will be done by a look at the original distribution, from which will be defined "cutting points" corresponding to significant categories. In the context of a quite large number of primary indicators, as we find here with these 39, a basic principle should be to keep the number of categories J_k, for a given indicator k, as low as possible, let's say not exceeding 5. Otherwise, the total number of categories will rapidly become excessively high, whence more complexity in the analysis. Also, looking at the distribution of primarily categorical variables, some grouping and then recoding may appear desirable, always in view of reducing the total number of categories, without loosing the basic significance of the data. As will be seen below in analysis outputs, from our 39 indicators, after these operations of categorization and recoding, we come out with a total of 111 categories.

Three basic conditions should be respected in the coding of categorical variables:

- to preserve the ordinality where it exists (e.g., #3, "Major source of water in dry season", all originally quantitative variables like #8, etc.),
- to start the coding with the value 1,
- to keep the coding sequential (no gap in the integer codes).

The first condition simplifies considerably the analysis of results coming from MCA. The second condition is a specific requirement of the program "Homals" for MCA with SPSS. The

[13] We implicitly ignore here the case of missing data (non-response).

third one reduces the volume of output coming out of the analysis. Nevertheless, none of these three conditions has any effect on the numerical results as such.

Table 1 A first set of 39 indicators

Indicator	Section	Question	Description	Stata var. name
Economy and Infrastructure (9 indicators)				
1	2	04	Does a road pass by the community?	s2q04
2	2	09	Do most hlds. have electricity?	s1q09[14]
3	2	12	Major source of water in dry season	s1q12
4	2	14	Food shop or restaurant in the community	s2q14
5	2	15	Post office in the community	s2q15
6	2	17	Public loud speaker or radio station	s2q17
7	2	20 and 21	Market infrastructure	s2q20, s2q21
8	2	24 and 25	Distance for public transportation	s2q24, s2q25
9	2	39	Large enterprise in the community	s2q39
Education (10 indicators)				
10	3	13	Fees at grade 1-3	s3q131
11	3	13	Fees at grade 4-5	s3q132
12	3	16	Primary enrolment rate	s3q161, s3q171
13	3	18	First reason not attending primary school	s3q181
14	3	19	Major problem in primary school	s3q191
15	3	20	Is there a lower secondary school?	s3q20
16	3	27	Fees at lower secondary school	s3q27
17	3	30 and 31	Enrolment rate at lower secondary school	s3q301, s3q311
18	3	32	First reason not attending lower secondary school	s3q321
19	3	33	Major problem in lower secondary school	s3q331
Health (11 indicators)				
20	4	07	Major health problem in the community	s4q071
21	4	08	Major problem with health services	s4q081
22	4	09	Where most of women give birth?	s4q09
23	4	1b	Clinic in the community	s4q01, item B
24	4	1c	Pharmacy in the community	s4q01, item C
25	4	1e	Doctor in the community	s4q01, item E
26	4	1h	Pharmacist in the community	s4q01, item H
27	4	1i	Traditional midwife in the community	s4q01, item I
28	4	1j	Trained midwife in the community	s4q01, item J
29	4	1k	Traditional healer in the community	s4q01, item K
30	4	2a	Distance (kms) to hospital	s4q02, item A
Agriculture (9 indicators)				
31-37	5	07	Proportion of land quality 1,2,3,4,5,6,7	s5q07a and b
38	5	12	Agric. extension center in the community	s5q12
39	5	15	Agricultural cooperative in the community	s5q15

3.2.3 Data exploration with a first MCA.

To try to identify some structure in a large set of categorical indicators, it is useful to run a first data reduction analysis as MCA, especially when we know that this is the one which will

[14] In the official files as in the corresponding data dictionary, there is an inconsistency in the numbering of some variables of section 2: s1 was used instead of s2.

be used at the end to produce the composite indicator. We will be particularly attentive here to the analysis of the column cluster N(J) (indicators and categories), even if the analysis of the row cluster N(I) comes out at the same time (see Appendix 1). This analysis has been runned here with the 39 indicators, with the program "Homals" as provided by SPSS 10.1. The output is given in Appendix 3.

3.2.3.1 Distributions

A quick look at the set of the K tables (K=39) giving the absolute distribution (here called "Marginal Frequencies") of each indicator provides us the only missing information relative to the specific 0-1 correspondence table here analysed with HOMALS, and not given as output since its dimensions are too large: (N, $J=\Sigma J_k$), here (120, 111). From these K tables, we in fact get the active row margins and the mass of each column, that is of each category. The mass of each category is obtained by the simple computation:

$$\text{mass} = \text{absolute frequency} \div \{(N \times J) - \text{missing values}\}$$

Here $N \times J = 120 \times 111$ = 4 680, and there are only 6 missing values, and thus the required denominator is 4 674. As an example, the mass of the category "AgricY" is: $77 \div 4674$ = 0,0165. This is the number that would appear in the bottom lene of the row profiles table.

3.2.3.2 Discriminating power of each indicator

A table called "Discrimination Measures", pp. 9-10, gives, for each of the K (39) indicators, a measure of its dispersion in each of the factorial axis. This number is the variance of the quantified indicator, that is, the variance of the factorial scores obtained by the set of categories associated to this indicator. These scores are given farther in the ouput, where they are named "Category Quantifications". It is easy to export this SPSS table as an EXCEL file and to sort it in each dimension in view of ordering the K (39) indicators according to their discrimination measure in each axis. The result is given in Appendix 3, p. 35.

Here, we easily see that the most discriminating indicators, in the first axis, are:

1. Do hlds have electricity?…………………… 0,514 (Ind #2)
2. Proportion of land quality 1. ………………. 0,492 (Ind #31)
3. First reason not attending primary school..0,388 (Ind #13)
4. Proportion of land quality 2………………... 0,371 (Ind #32)
5. Primary enrolment rate ……………………. 0,348 (Ind #12)

And the less discriminating ones are:

38. Trained midwife in community …………….. 0,009 (Ind #28)
39. Clinic in community ………………………… 0,005 (Ind #23)

Thus, the indicator of rank 1 is 100 times more discriminating, in the first axis, than the last one. The eigenvalue (inertia) of each axis, given in the table "Eigenvalues" p. 9, provides an interesting reference point for the analysis of the discrimination measures: it is <u>the mean value</u> of all the discrimination measures for this axis. Here we see that there are 21 indicators whose discrimination measure is below the average 0,186.
As can be seen p. 11, the SPSS output provides a plot of the discrimination measures, but with a large number of indicators, it is unreadable, whence the easier analysis with the EXCEL file.

Another very important information to assess the discriminating power of an indicator is to look at the spread, on each axis, of its quantified values, that is the scores of its categories, relative to this axis. These scores, called "Quantifications", are given in a set of K tables, pp. 11 to 21. Looking at this set of tables, we can identify the indicators whose categories have the largest weights on each axis. But a graphical representation of the quantified indicators is the best way to look quickly at these spreads. There is such a plot p. 21, "Quantifications", but again, it is unreadable with a large set of indicators and categories (here, respectively 39 and 111). Fortunately, there is an option in HOMALS that allows to export the set of category quantifications, as an SPSS data file. Using the "GRAPHS" tools in SPSS, we can then generate readable and significant plots for each quantified indicator. These plots are given here pp. 24-34. Looking at these plots, we observe that, for the first axis:

- the third most discriminating indicator, #13,"First reason not attending primary school ", has a larger spread than the first two, #2,"Do hlds have electricity?", and #31,"Proportion of land quality 1", which have similar spread,
- the less discriminating indicator, #23,"Clinic in community", has also the smaller spread.

These both informations, the discrimination measure and the spread of each quantified indicator, can help greatly if we intend to reduce the total number of indicators, for different reasons.

But the graphical analysis of the indicators highlights a fact that was obviously known from the beginning: we are facing here two exclusive classes of indicators. Some of them have a clear ordinal structure, and their quantified categories can thus be connected by a (polygonal) line. Its the case here of 33 indicators. Let's reserve for them the name **Poverty Indicators (PI)**. Some others, here 6 indicators, don't have this ordinal structure. They are:

- #13 "First reason not attending primary school"
- #14 "Major problem in primary school"
- #18 "First reason not attending lower secondary school"
- #19 "Major problem in lower secondary school"
- #20 "Major health problem in the community"
- #21 "Major problem with health services".

We will call these last ones **Poverty Descriptors (PD)**.

Looking specifically at the PI, we observe different patterns, in a limited number, and this takes us to the following point.

3.2.3.3 The meaning of the first factorial axis

Bearing in mind that we intend to give to the first factorial axis a central role in the computation of a composite indicator, the preceding views on the data invite us to look for the meaning of this "mysterious" dimension obtained through a data reduction technique.

Going back to the graphical representation of the indicators, pp. 24-34, but this time concentrating our analysis on the 33 PIs, we observe:

- a first dominating set of 28 PIs presenting the property that, with respect to their ordinal structure, they are consistent relatively to the first axis. We mean that from left to right, for all of them, the welfare condition expressed by the indicator is deteriorating. Some among these 28 are:
  - #1,"Does a road pass by the community?"
  - #2,"Do hlds have electricity?"
  - #4,"Food shop or restaurant in the community".

  We will say of these indicators that they have the property of *First Axis Ordering Consistency (FAOC)*.
- That a minority set of 5 PIs do not having the *FAOC* property are:
  - #3,"Major source of water in dry season"
  - #7,"Market infrastructure"
  - #8,"Distance for public transportation"
  - #33,"Proportion of land quality 3

- #34," Proportion of land quality4.

  Nevertheless, regarding carefully to the plot of these indicator, we observe that there are not far of having the *FAOC* property, except for #3, on which we come back below.

This *FAOC* property, together with the largely dominating set of PIs sharing this property (28), and the fact that the residual set is not so far from having this property, justify that the first axis be identified as a **Poverty Axis**. This is the meaning we will give to this first axis.

Having thus identify the first principal axis as a Poverty Axis, the 6 poverty descriptors (PDs) now take their own meaning: they pretend to characterize poor communes in terms of health and basic education problems. We will come back on their role below, but we already see them emerging from the analysis as potential "poverty determinants".

3.2.3.4 Reduction of the set of poverty indicators

An important conclusion of the preceding analysis is that the final determination of the set of indicators to be included in the composite indicator should logically be restricted to the subset of PIs having the *FAOC* property: they are those giving its meaning to the first principal axis. Here, they are already 28. On the other hand, for operational purposes, having a short list of primary indicators present many advantages. The preceding "browsing" of the primary analysis data set can suggest some elimination, without sacrifying what we consider as important poverty information. Let's try to illustrate a possible process in that direction, with the following steps.

- First, we must be fair with each indicator, before rejecting it. Among the 5 non-*FAOC* PIs, 3 could become *FAOC* by a simple recoding of two extreme categories: #7,"Market infrastructure",#8,"Distance for public transportation", #33,"Proportion of land quality 3. We ignore #33, since there are already 5 other indicators on land quality. But we recode #7 and #8, because they bring into the picture important domains not covered by other indicators.
- Second, we still have 5 indicators on land quality. We keep only the two extremes, #31, C_Land1 and #37, C_Land7.
- Third, the three indicators on school fees, #10, #11 and #16, have a low discriminating power (see table p. 35), and even if they are *FAOC*, their inclusion as PIs remains ambiguous. We prefer to reclassify them as PDs, and thus we don't retain them any more.

After these steps, the remaining list of PIs is reduced to 24 indicators.

The rejection of one important indicator, #3,"Major source of water in dry season", raises an interesting question. As seen from the plot p. 34, the category #2, "Deep drilled well with pump", is opposite to the category #1, "Rain water", relatively to the poverty axis. We would expect that they be close, and then we would not have the strange plot we get here. In dependently of the fact that both categories have a very low frequency in our sample, after discussing the case, we realized that one explanation could be that water supply development programs frequently install deep drilled wells by targeting the most deprived villages, those using probably rivers or lakes as sources of drinking water. More generally, any development program targeted on the poorest community and providing them with the best technology available will tend to disturb progressively the poverty ordinal structure of some indicators. This is good for the population, but makes life difficult for the statistician! This is what we will call the *Poverty Targeting Neutralization Effect*: it destroys the nice ordinal structure of some poverty indicators, good for the poor!

3.2.4 Construction of the composite indicator with a final MCA

In 3.2.3, the focus of the analysis was on the column cluster N(J), that is, on the indicators and their categories. This exploration phase led us essentially at reducing the number of indicators and categories, from 39 and 111 respectively, to 24 and 56. The final list of indicators is given in Table 2.

Table 2 A final set of 24 indicators

Indicator	Section	Question	Description	Stata var. name
Economy and Infrastructure (8 indicators)				
1	2	04	Does a road pass by the community?	s2q04
2	2	09	Do most hlds. have electricity?	s1q09[15]
4	2	14	Food shop or restaurant in the community	s2q14
5	2	15	Post office in the community	s2q15
6	2	17	Public loud speaker or radio station	s2q17
7	2	20 and 21	Market infrastructure	s2q20, s2q21
8	2	24 and 25	Distance for public transportation	s2q24, s2q25
9	2	39	Large enterprise in the community	s2q39
Education (3 indicators)				
12	3	16	Primary enrolment rate	s3q161, s3q171
15	3	20	Is there a lower secondary school?	s3q20
17	3	30 and 31	Enrolment rate at lower secondary school	s3q301, s3q311
Health (9 indicators)				
22	4	09	Where most of women give birth?	s4q09
23	4	1b	Clinic in the community	s4q01, item B
24	4	1c	Pharmacy in the community	s4q01, item C
25	4	1[e]	Doctor in the community	s4q01, item E
26	4	1h	Pharmacist in the community	s4q01, item H
27	4	1i	Traditional midwife in the community	s4q01, item I
28	4	1j	Trained midwife in the community	s4q01, item J
29	4	1k	Traditional healer in the community	s4q01, item K
30	4	2a	Distance (kms) to hospital	s4q02, item A
Agriculture (4 indicators)				
31	5	07	Proportion of land quality 1	s5q07a and b
37	5	07	Proportion of land quality 7	s5q07a and b
38	5	12	Agric. extension center in the community	s5q12
39	5	15	Agricultural cooperative in the community	s5q15

The results of the final run of Multiple Correspondence Analysis on these 24 indicators are given in Appendix 4.
Now, without ignoring N(J), we will concentrate our attention on the cluster of lines N(I), that is, the set of population units, here the 120 rural communes.

3.2.4.1 Analysis of the first factorial axis

From the outputs given in Appendix 4, we observe that:

[15] In the official files as in the corresponding data dictionary, there is an inconsistency in the numbering of some variables of section 2: s1 was used instead of s2.

- the first eigenvalue (inertia of the first axis) has increased from 0,186 to 0,216, even with a reduced total inertia caused by the elimination of 15 indicators. Thus, the average discrimination mesure of the remaining 24 indicators has increased and the first axis appears stronger (Appendix 4, pp. 6-7). But the range of the discrimination measures has been reduced: the gap between "strong" and "weak" indicators is shortened (p. 24);
- the order of the indicators, according to their discrimination measure, has changed. Indicator # 2, "Do most hlds. have electricity?", has passed from rank 1 to rank 2, and Indicator #24,"Pharmacy in the community", from rank 12, is now the first one. But as can be seen from the plot (p. 19), the spread of Indicator # 2 has not been reduced. The same situation is observed for Indicator #12,"Primary enrolment rate", whose rank has decreased from 5 to 14, but its spread (plot p. 22) has not really been reduced. In these both last cases, the central category is closer to the center of gravity (0) than with the first analysis;
- the graphical analysis (pp. 18-23) reveals that all 24 indicators have the *FAOC* property. But, interestingly, the orientation of the first axis is reversed: welfare now increases from left to right, what corresponds to the standard money-metric geometrical representation of welfare (income). The second axis is also reversed. This is easily understandable: inertia analysis is axis orientation independent;
- as in the first exploratory analysis, we don't see, from the category quantifications and their graphical representation, any obvious meaning to the second axis.

3.2.4.2 Composite indicator and the complete ordering of the population units

The table "Object Scores", pp. 15-17, gives, in the column "Dimension 1", the value of the composite poverty indicator for each population unit. As explained and defined in the theory (Appendix 1, p. 15, and above, section 3.1), the value of the composite indicator, for a given population unit, is the average of its category-weights (normalized category quantifications). To simplify the analysis of the poverty level of the population units, these object scores are automatically exported by the program HOMALS as new variables in the original data file. The very first use of this augmented file is to produce the list of the population units, here the 120 communes, according to their first axis score, that is, their multidimensional poverty level. This is precisely the basic result we are looking for since the beginning. This result is presented in the table "Communes ordered according to decreasing multidimensional poverty", with some demographic characteristics, in pages 25-27, Appendix 4.

3.2.4.3 Analysis of the second factorial axis through duality

Until now, we have not been able to extract a meaning for the second factorial axis, at least by looking at indicators and categories. But now we are in a position to look at the dual picture of poverty, the one given by the scores of the population units. We know, from the theory presented in Appendix 1 and by section 3.1, that, for any dimension, the factorial score of any group of population units is their average score in that dimension. We can then easily compute these factorial scores for different population classifications, and thus try to map multidimensional poverty in two dimensions only. This is an appealing way to build synthetic poverty profiles from a large set of indicators. As an example of this approach, let's look at the regional classification of the 120 communes.

The regional factorial scores are presented in Table 3 below.

Table 3

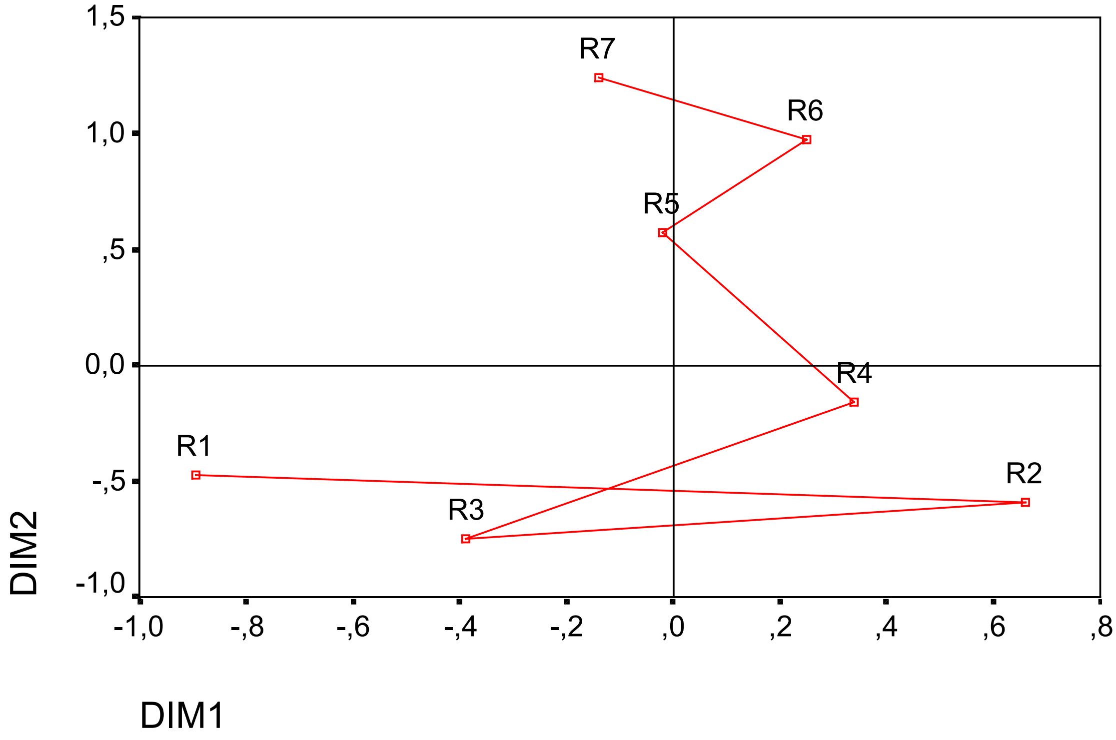
Factorial scores of the 7 regions of Vietnam

Region		Composite Poverty Indicator	Score on Second Axis
Northern Uplands	Mean	-,8950	-,4697
	N	19	19
Red River Delta	Mean	,6589	-,5876
	N	32	32
North Central	Mean	-,3880	-,7492
	N	18	18
Central Coast	Mean	,3399	-,1556
	N	12	12
Central Highlands	Mean	-,0203	,5718
	N	4	4
Southeast	Mean	,2488	,9720
	N	10	10
Mekong River Delta	Mean	-,1404	1,2429
	N	25	25
Total	Mean	,0006	,0000
	N	120	120

In Table 3, we don't see any particular regional pattern for the composite poverty indicator, but we immedialy see that for the second factorial axis, the three northern regions are opposite to the two southern ones, with the two central ones in-between.
A graphical representation of the 7 regions in the first two factorial axis seems required for a better understanding. Graph 1 gives the global picture of the regional dimension revealed in the two-dimension poverty map.

Graph 1

Regions in the Poverty Map



R1: Northern Uplands
R2: Red River Delta
R3: North Central
R4: Central Coast
R5: Central Highlands
R6: Southeast
R7: Mekong River Delta

From Graph1, we immediately confirm the regional meaning of the second factorial axis: it really discriminates between the North and South, with the Central Vietnam close to the gravity centre. The behavior of the line relatively to the first axis (quantified multidimensional poverty) reveals that this regional discrimination is not relatively to the global poverty level: the average level of R1+R2+R3 (North) is similar to the average level of R6+R7 (South). But a very important fact appears: relatively to the multidimensional poverty, there seems to be more inequality in the North (compare R1 and R2) than in the South.
There must then be different types of poverty in the North and in the South. To try to understand this qualitative difference, we have reproduced in Appendix 4, pp. 28-33, the simultaneous plot of the regions and the indicators in the factorial poverty map graphs 2 to 5. The indicators are grouped according to their thematic domain, otherwise the plots would have been unreadable. From these graphs we observe that:

- relatively to the economic infrastructure, the rural South seems less advantaged in terms of roads and electricity than the richer rural North (Red River Delta). But relatively to electricity, there is much more inequality within the Northern region itself.

- the South, especially the Mekong River Delta, seems less provided with primary and secondary schooling than the richer Red River Delta in the North.
- the availability of local clinic tends to increase in the South relatively to the North.
- local agricultural cooperatives are much less found in the South than in the North, and the availability of the better quality land seems to be concentrated in the North, but essentially in the Red River Delta.

In conclusion, we may say that the second factorial axis throws light on the different regional patterns of poverty from North to South: a Southern region generally, with some exceptions, less advantaged than the Northern Red River Delta, but apparently more equitable. Health poverty seems to be more acute in the North, particularly in the Uplands region, while education poverty dominates in the South.

3.2.4.4 A functional form to compute the composite indicator of population units not included in the analysis

Let the indice u denote any population unit not included in the sample used in the multiple correspondence analysis, and C_u its composite indicator value. Then:

$$C_u = \frac{\sum_{k=1}^{K} \sum_{j_k=1}^{J_k} W_{j_k}^k I_{j_k}^k}{K} \text{ , where}$$

$$K = \text{number of categorical indicators}$$

$$\mathrm{J_k} = \text{number of categories for indicator k}$$

$$\mathrm{W_{j_k}^k} = \text{the weight (normalized first axis score) of category } \mathrm{j_k}$$

$$I_{j_k}^k = \text{the binary variable 0/1 taking the value 1 when the unit u has the category } \mathrm{j_k}.$$

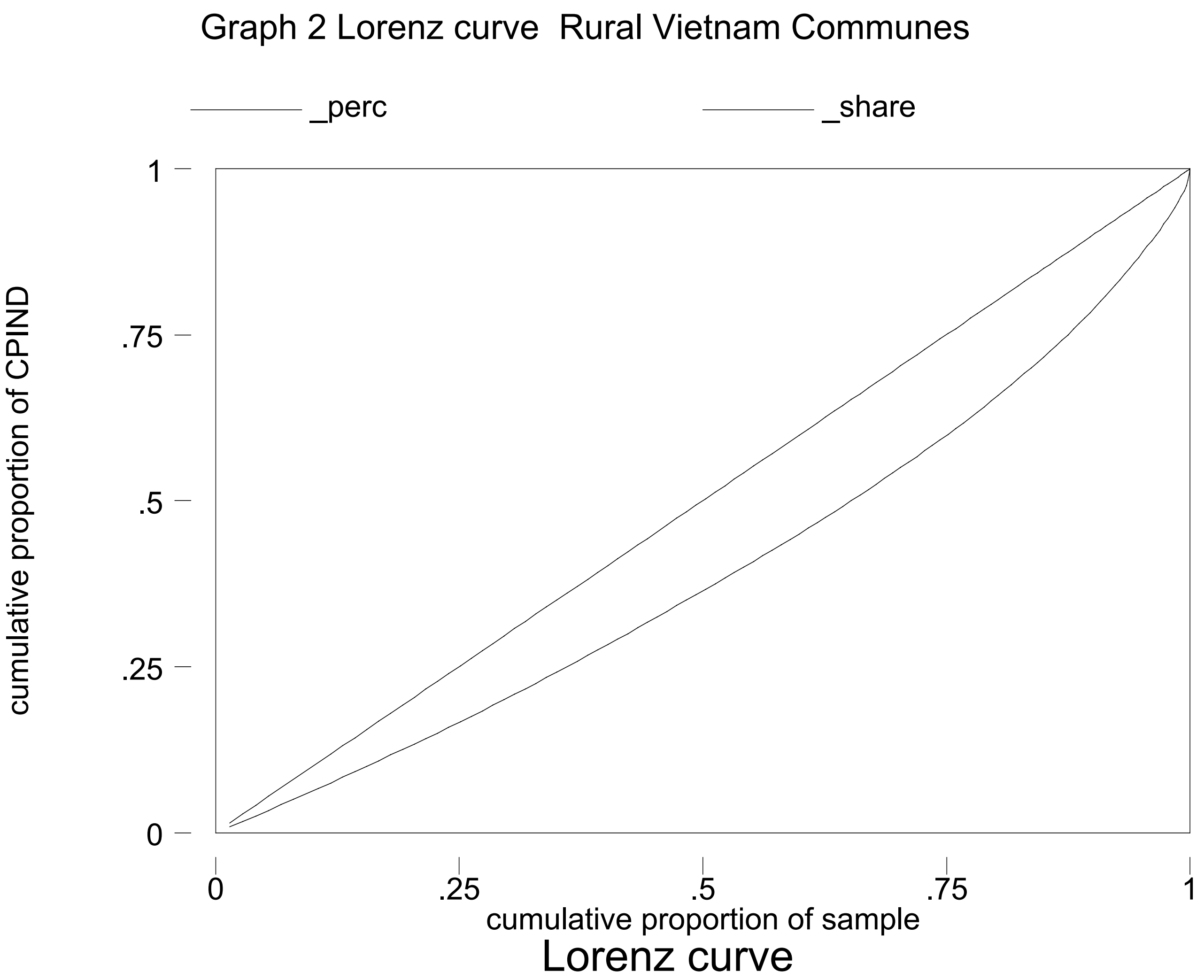
Thus, to easily compute the composite indicator, we only need a simple questionnaire that can be built from the output of the final MCA, giving the weight to apply for each category. Such a questionnaire is given in Appendix 4, pp. 33-34. The weights, obtained from the first factorial score, divided by the first eigenvalue (0,216), have been multiplied by 1 000, for simplification. The value of the composite indicator, for any commune u, is obtained by adding the weighted binary category variables and averaging.

3.3 Multidimensional inequality indices

The composite indicator C is a numerical variable measuring the level of multidimensional welfare. Like any welfare variable, it can be analysed with the most recognized analysis tools as developed essentially for money-metric indicators. But it has the property of being negative in its lowest part. Obviously, it can be made positive very easily, by a translation using the absolute value of the average C_{min} of the minimal categorical weight W^k_{min} of each primary indicator. This average is really the minimal value that can be taken by any population unit.

$$C_{\min} = \frac{\sum_{k=1}^{K} W_{\min}^k}{K}.$$

Here, C_{min} is easily seen to be -2737, from the table given pp. 33-34, Appendix 4. We thus add the absolute value of this average to each population unit score, to obtain the new positive score C^*. With C^*, standard inequality analysis can be done. We give in Appendix 5 some detailed results obtained with the package INEQUAL of STATA. There is first the Lorenz curve of C^*, reproduced below as Graph 2.



The degree of multidimensional inequality between the rural communes seems quite low, and the Lorenz curve does not change significantly if the data are weighted according to the demographic size of the commune, taken either as the total population or the number of households. As can be seen in the Appendix 5, the Lorenz curve of the Northern communes is slightly more convex than for the Southern ones.

The most usual inequality indices, computed for the composite indicator C^*, confirm these graphical results (Table 4).

Table 4
Multidimensional Inequality Indices
for rural communes of Vietnam

Region	Gini coefficient	Theil entropy	Theil mean log deviation
North	0,217	0.078	0,092
Central	0,212	0,085	0,120
South	0,185	0,059	0,068
Rural Vietnam	0,209	0,074	0,093

Thus, all these inequality indices are smaller in the South than in the North. A Gini coefficient of 0,209 reveals low inequality among the rural communes. As a comparison, the Gini coefficient among all households, computed from the same VLSS-1 survey, is 0,34[16].

It should be emphasized here that all these inequality comparisons across regions, with the composite indicator, do not rely on price adjusment that are usually an important issue in the case of a money-metric indicator. The same remark applies for inequality comparisons across time.

3.4 Multidimensional poverty indices

With the composite indicator C^*, it is possible to compute standard poverty indices, once a poverty line has been identified. There is no obvious way of establishing such a poverty line for a composite indicator. But some ideas can be explored.

3.4.1 Poverty lines

If we see a poverty line as a fixed value allowing reliable and relevant comparisons through space and time, and used as a tool for defining operational targets to poverty alleviation policies, then some poverty lines can be suggested.

First, we can adopt a *relative* approach to this issue. We can decide to fix a specific quantile of the indicator C^*, the second, or third quintile, or whatever. Here, we have taken, as an example, the third quintile, on the basis that, with the same survey data, the rural income poverty rate has been computed as being 57%[17]. The 60% quantile has been computed for the indicator C^*, either for unweighted as for weighted data. Results are given in Table 5.

Table 5
Third quintile poverty line with the composite indicator

No demographic weight	**3176**
Weight: # individuals (pop.)	**3318**
Weight: # households	**3321**

Second, we can adopt an *absolute* approach to the same issue, by fixing, for each categorical indicator, here always an ordinal variable, a specific category taken as a

[16] Viet Nam, *Poverty Assessment and Strategy*, The World Bank, January 1995, p. 4.
[17] loc. cit., Figure 1.3, p.10.

poverty level for this indicator. Let W^k_{pov} be the categorical weight of this category. There are as many poverty levels as there are primary indicators integrated in the composite indicator, here 24. A possible poverty line can then be taken as the maximum value of W^k_{pov}, over the K such weights. With such a definition, a necessary condition for a population unit to be poor is to be poor in at least one dimension, that is, in at least one primary indicator. But this condition is not sufficient. A sufficient condition is to be poor in all dimensions. But it is not a necessary one. The necessary and sufficient condition is obviously that the mean score of the population unit, over the K primary indicators, be greater than the value $Max(W^k_{pov})$.
Here, in case of a binary indicator, we have taken as W^k_{pov} the minimal weight. As can be seen from Appendix 4, pp. 33-34, there are 17 such indicators. For the seven 3-level indicators, we have taken the minimal weight in four cases:

- Where most of women give birth? (home)
- Distance kms to hospital (>10 k)
- Proportion of land quality 1 (0)
- Proportion of land quality 7 (>25).

In the remaining three cases, we have taken the central category:

- Do hlds have electricity? (a few)
- Primary enrolment rate (80-90)
- Distance public transportation (3-10k).

It comes out that $Max(W^k_{pov})$ is given by the indicator "Traditional midwife in community", with the value – 455. After the translation with value 2737, the poverty line becomes **2282**.

3.4.2 FGT poverty indices

Once a poverty line has been defined, all known poverty measures and poverty indices developed until now essentially within the money-metric literature become available with the multidimensional composite indicator C^*. This type of poverty analysis is illustrated here by computing the well-known Foster-Greer-Thorbecke (FGT) indices, for the different poverty lines defined in 3.4.1. All computations have been done with the package "POVDECO" (poverty decomposition) of STATA 7.0. The details are presented in the Appendix 6, with full decomposition of multidimensional poverty across regions. Some results are reported in Table 6.

We see from Table 6 that according to different poverty lines, the geographical decomposition of the FGT indices does not show a systematic trend. This confirms the factorial correspondence analysis, where it has been seen that it is the pattern of poverty rather than the global poverty that differentiates between the North and the South, especially when we aggregate in the Northern region the Uplands, much poorer, and the Red River Delta, much richer.

The most interesting fact revealed by Table 6 is that the multidimensional poverty analysed here is correlated to the size of the commune: small communes are poorer than large ones. Actually, the third quintile poverty line is higher when the communes are weighted with their population size: 3318 instead of 3176. In addition, with the absolute poverty line 2282, the incidence of poverty decreases with the weigted data: 26,9% instead of 34,2%. This is a reverse situation in comparison with the money-metric analysis of poverty where, frequently, it is observed that large households are poorer than small ones.

Table 6
Multidimensional FGT Poverty Indices
for rural communes of Vietnam

Regions	Unweighted data					
	pov. line = 3176			pov. line = 2282		
	P_0	P_1	P_2	P_0	P_1	P_2
North	0,569	0,203	0,099	0,373	0,107	0,044
Central	0,647	0,218	0,113	0,353	0,111	0,064
South	0,600	0,195	0,091	0,286	0,096	0,040
Rural Vietnam	0,600	0,205	0,101	0,342	0,105	0,049
Regions	**Weight = Total Population**					
	pov. line = 3318			pov. line = 2282		
	P_0	P_1	P_2	P_0	P_1	P_2
North	0,542	0,167	0,074	0,303	0,065	0,024
Central	0,652	0,155	0,069	0,211	0,059	0,030
South	0,618	0,211	0,106	0,273	0,105	0,047
Rural Vietnam	0,600	0,183	0,086	0,269	0,081	0,035

Finally, it should be recalled that this type of multidimensional poverty analysis, using the same techniques than the money-metric analysis, is not subject to the difficulties of price variability across space and time.

4. Conclusion

By exploring the fundamental ideas and the inutition on which rely the main computational techniques relevant to the composite indicator problem, we have tried to make explicit the reasons justifying the choice of a specific technique, and thus, to eliminate the arbitrary too often met in this very current debate. For a very general technique, the Multiple Correspondence Analysis (MCA), we have illustrated the theory with a detailed case study, the analysis of the Commune Questionnaire data from the first Vietnam Living Standard Survey (VLSS-1), in which 24 primary indicators have been summarized in a quantitative composite indicator.

On this basis, we propose the following solution to the composite indicator problem.

1. In the general case of a mix of qualitative and quantitative poverty indicators:

   a) to transform the quantitative indicators into ordinal qualitative ones, with a small number of categories for each of them;
   b) to run a MCA analysis on the 0-1 indicator matrix[18];
   c) to take for the composite indicator the following definition:

[18] "Indicator Matrix" is the name used in the statistical literature. We would prefer the appellation "Category Matrix", since there is a binary variable generated by each category of each primary indicator.

Definition
The composite indicator of multiple qualitative poverty indicators, defined as a set of categories, for different population units, is given by:

- **computing the profile of each population unit relatively to these primary indicators,**
- **applying to this profile the category-weights given by the normalized scores of these indicators on the first factorial axis coming out of the Multiple Correspondence Analysis (MCA) of the indicators.**

The expression of the composite indicator for the population unit i is then:

$$C_i = \frac{\sum_{k=1}^{K}\sum_{j_k=1}^{J_k} W_{j_k}^k I_{i,j_k}^k}{K}$$ **, where**

K = number of categorical indicators

$\mathrm{J_k}$ = number of categories for indicator k

$\mathrm{W_{j_k}^k}$ = the weight (normalized first axis score) of category $\mathrm{j_k}$

I_{i,j_k}^k = the binary variable 0/1 taking the value 1 when the unit i has the category $\mathrm{j_k}$.

This solution constitutes the inertia approach to the composite indicator problem.

2. In the particular case of a set of positive quantitative poverty indicators ("money-metric type" indicators), an alternative to the general solution is:

a) to transform the original poverty indicators into shares I_{ik} for the population unit i,
b) to run a Principal Component Analysis (PCA) on the K transformed indicators,
c) to compute the composite indicator for the population unit i as:

$C_i = \sum_{k=1}^{K} \delta_k I_{ik}$ where δ_k is the vector of weights provided by the first characteristic vector coming out of the PCA analysis.

This alternative solution is theoretically based, especially for the linear functional form in the shares, on the entropy literature, particular on the entropic divergence measure associated to Theil's inequality indices.

We insist on the fact that in both cases 1. and 2., the formal solution is provided by a clearly defined optimization process that contribute again to eliminate the arbitrary in the solution to the composite indicator problem.

The construction of a composite indicator can be seen as the first step towards the multidimensional analysis of inequality and poverty. For the qualitative approach to multidimensional poverty, it allows to use, in a second step, all analytical tools developed until now within the realm of money-metric analysis. This fact has been largely illustrated in sections 3.3 and 3.4.

In a particular analytic situation, nothing prevents of using a combination of both solutions, the inertia and the entropy ones. For instance, a large set of indicators could be first classified according to individual poverty analysis into age-sex specific groups; in each group, a MCA analysis could provide a composite indicator age-sex specific. These composite indicators, once transformed into "shares" relative to a given type of population unit (household, vilage, etc.), could be aggregated into only one indicator according to the PCA analysis proposed as the final step of the entropy-based approach.

Finally, we come out of this study of multidimensional poverty with a reinforced view of poverty as a relative concept, before any kind of absolute meaning that we can associate to it. This view emerges from the dominating idea of indicators as "shares" in the entropy literature relevant to economic analysis, and from the unifying treatment of multivariate analysis as multidimensional scaling (MDS) techniques, where some kind of "distance" between population units underlies the concept of inertia generating our proposed general solution to the composite indicator problem. Thus, it strengthens our feeling that poverty analysis is primarily inequality analysis; that poverty is in fact inequitable inequality which can be reduced and eliminated through appropriate policies. Poverty line concepts should then be seen essentially as tools, among others, to help fixing operational targets to poverty eradication policies, an ethical priority for any society.

References

Anderson T.W. (1958)	*An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York, London.
Benzécri J.-P. and F. (1980)	*Pratique de L'Analyse des Données, I, Analyse Des Correspondances, Exposé Élémentaire*, Dunod, Bordas, Paris.
Bourguignon F. and S.R. Chakravarty (1999)	"A Family of Multidimensional Poverty Measures", in *Advances in Econometrics, Income Distribution and Scientific Methodology*, D.J. Slottje ed.,
Burbea J. and C. R. Rao (1982a),	"On the Convexity of Some Divergence Measures Based on Entropy Functions", *IEEE Transactions on Information Theory*, vol. IT-28, no. 3, 489-495.
Burbea J. and C. R. Rao (1982b),	Entropy Differential Metric, Distance and Divergence Measures in Probability Spaces: A Unified Approach, *Journal of Multivariate Analysis* 12, pp. 575-596.
Greenacre M. and J. Blasius (1994)	*Correspondence Analysis in the Social Sciences, Recent Developments and Applications*, Academic Press, Harcourt Brace & Company Publishers.
Havrda J. and F. Charvat (1967)	"Quantification method of classification processes: Concept of structural α-entropy", *Kybernetica*, vol. 3, 30-35.
Hotelling H. (1933)	"Analysis of a complex of statistical variables into principal components", *Journal of Educational Psychometrics*, 24, 417-441, 498-520.
Hotelling H. (1936)	"Relations between two sets of variables", *Biometrika*, 28, 321-377.
Maasoumi E. (1986),	"The Measurement and Decomposition of Multi-Dimensional Inequality", *Econometrica*, vol. 54, no. 4, pp. 991-997.
Maasoumi E. (1993)	"A Compendium to Information Theory in Economics and Econometrics", *Econometric Reviews*, 12(2), 137-181.
Maasoumi E. (1999),	"Multidimensioned Approaches to Welfare Analysis", chap. 15 in J. Silber ed., *Handbook of Income Inequality Measurement*, Kluwer Academic Publishers, 437-477.
Meulman J.J (1992)	"The integration of multidimensional scaling and multivariate analysis with optimal transformations", *Psychometrika*, vol. 57, no.4, 539-565
Renyi A. (1961)	"On measures of entropy and information", Proc. 4th Berkely Symposium Statist. Probability, 1, 547-561, Univ. of California Press.
Rényi A. (1966)	"Introduction à la théorie de l'information", appendice à *Calcul des probabilités*, Dunod, Paris.
Sahn D.E. and D.C. Stifel (August 1999)	"Poverty Comparisons Over Time and Across Countries in Africa", Mimeo, Cornell Food and Nutrition Policy Program, Cornell University, Ithaca, NY.
Shorrocks A. F. (1984)	"Inequality Decomposition by Population Subgroups", *Econometrica*, vol. 52, no. 6, 1369-1385.
Shorrocks A.F. (1980)	"The class of additively decomposable inequality measures", *Econometrica*, vol. 48, no. 3, 613-625.

Appendix 1 The Basic Principles of Correspondence Analysis

Annex A SPSS Output of Correspondence Analysis

Annex B SPSS Outpiut of Multiple Correspondence Analysis (Homals)

Appendix 2 Community Questionnaire and Data Dictionary

Appendix 3 Data Exploration with Multiple Correspondence Analysis

Appendix 4 Final Data Analysis with Multiple Correspondence Analysis

Appendix 5 Multidimensional Inequality Indices

Appendix 6 Multidimensional Poverty Indices

Appendix 1

The Basic Principles of Correspondence Analysis

and its extension to

Multiple Correspondence Analysis

A composite indicator from multidimensional qualitative data

Louis-Marie Asselin
Canadian Centre for International Studies and Cooperation

December 30, 2002

Contents

8 Multiple Correspondence Analysis 15

Abstract

Data reduction techniques, more specifically factorial correspondence analysis, is used to build a composite numerical variable from a set of qualitative (categorical) variables.

1 Foreword

The approach will be here to present a statistical technique resorting to the set of **data reduction techniques** in view of "attacking" systematically and rationally the problem of aggregating multidimensional qualitative variables. The presentation is illustrated in reference to poverty data and analysis. It is also, as much as possible, oriented on habilitating the reader to become operational with a specific statistical software offering that type of technique among its routines: we refer to the SPSS 10.1 program **Correspondence Analysis,** and its extension to Multiple Correspondence Analysis, runned with the program **Homogeneity Analysis**.

2 Problem description

- We have in hands a database consisting of a set of qualitative poverty indicators (categorical variables) measured on statistical (population) units who can be *individuals, households, communities, regions, countries,* etc. These variables generate J categories on I population units.

- Motivation: income/expenditure variables not only can be viewed as reflecting just one dimension of poverty, but are also heavy and costly to measure and may be more or less reliable due to non sampling errors (particularly recall errors). For all these reasons, light and more reliable qualitative indicators are frequently used to describe different dimensions of poverty.

- From the J categories, we would like to construct an unique indicator synthetizing the information contained in the multiple indicators.

- One of the main objectives, not necessarily the only one, is to classify the I population units according to their relative poverty level.

3 Case study: data from Vietnam survey VLSS-1

- The first Vietnam Living Standard Survey (VLSS-1) was conducted in 1992-1993. The sample consists of 4 800 households randomly selected within 150 communes, themselves randomly selected among the about 10 000 communes in Vietnam. Among the 150 selected communes, 120 are rural and 30 are urban.

- Three questionnaires were administered: a household questionnaire, a community questionnaire and a price questionnaire. The community questionnaire was administered only in the 120 rural communes. It contains

142 questions, distributed in 5 sections: demography, economy and infrastructure, education, health and agriculture.

- For our case study, we use only the community questionnaire. In view of illustrating as simply as possible the CA approach to the computation of a multidimensional poverty composite indicator, we retain only two poverty indicators, generated by the two following questions:

  section 3 (education), question #16: how many children aged 6 to 11 are enrolled?

  section 5 (agriculture), question #7: what is the proportion of each type of quality land in the land fund of this commune?

  From the education question, since the total number of children in the age-group 6-11 is available, it is possible to compute the primary enrolment rate. This rate has been transformed into a categorical indicator with the three following categories:

  category 1: rate $< 80\%$

  category 2: 80% □ rate □ 90%

  category 3: rate $> 90\%$.

  The quality land question considers seven levels of quality. We have retained only the first level, which is the best quality. The 3 categories are the following:

  category 1: percentage $= 0\%$ (no land of quality 1)

  category 2: $0\% <$ percentage □ 25%

  category 3: percentage $> 25\%$.

- The 120 communities considered here are the 120 rural communes distributed in the 7 regions:

  1. Northern Uplands : 19
  2. Red River Delta : 32
  3. North Central : 18
  4. Central Coast : 12
  5. Central Highlands : 04
  6. Southeast : 10
  7. Mekong River Delta: 25

4 Data description

- Data consists of a table $I \times J$ of positive numbers, in many cases only 0 or 1.

- Notation

  $k(i,j)$: number in cell (i,j)

  $k(i) = \sum_{j=1}^{J} k(i,j)$: total of line i

  $k(j) = \sum_{i=1}^{I} k(i,j)$: total of column j

  $k = \sum_{i=1}^{I} \sum_{j=1}^{J} k(i,j)$: the general total

- $k(i,j)$ is usually interpreted as the frequency of occurrence of category j for the unit i.

- The values of indicators for our case study are given in Annex A, Correspondence Table, pages 1-3. It is seen that the values are 1 or 0, and there are only two 1 in each line, according to the fact that for each indicator, a given commune belongs to exactly one category.

5 Data transformation

Looking at the Correspondance Table , it is difficult to "see" a poverty structure and to identify poorer and richer communes, on a rational basis. A statistical analysis is needed to try to see better the poverty content of this table, and it begins by elementary transformations of data. At the same time, some terminology and notation relative to correspondence analysis is introduced.

- relative frequency of category j for unit i : $f_j^i = \frac{k(i,j)}{k(i)}$
- relative frequency of unit i for category j : $f_i^j = \frac{k(i,j)}{k(j)}$
- **profile** of unit i : $f_J^i = \left\{ f_j^i \mid j \in J \right\}$
- **profile** of category j : $f_I = \left\{ f_i^j \mid i \in I \right\}$
- **mass** (relative weight, marginal frequency) of unit i : $f_i = \frac{k(i)}{k}$

  $f_I = \{ f_i \mid i \in I \}$
- **mass** (relative weight, marginal frequency) of category j : $f_j = \frac{k(j)}{k}$

  $f_J = \{ f_j \mid j \in J \}$

Remark 1 *The notion of* ***profile*** *of a population unit i allows to show the categorical structure of the unit, independently of its size. By comparing a given unit-profile f_J^i with the mean profile f_J , we can begin to view to which extent a population unit differs, structurally, from the general population structure, in regard to the observed indicators. Mutatis mutandis, the notion of* ***profile*** *of a category j allows to show the population structure of the category, independently of its importance as a social phenomena. By comparing a given category-profile f_I^j with the mean profile f_I , we can begin to view to which extent a category differs, in its demographic structure, from the general population structure, in regard to the observed population units.*

- **cluster** $N(I)$ in dim-J space

  $N(I) = \left\{ f_J^i \mid i \in I \right\}$. So, $N(I)$ is the set of the I unit-profiles in dim-J space.
- **cluster** $N(J)$ in dim-I space

  $N(J) = \left\{ f_I^j \mid j \in J \right\}$. So, $N(J)$ is the set of the J category-profiles in dim-I space.
- **centre of gravity (centroid)** of cluster $N(I)$

It's the weighted mean g_J of the I unit-profiles belonging to the cluster $N(I)$.

$g_J = \sum_{i=1}^{I} f_i f_J^i$. It's easy to see that $g_J = f_J$: the centroid of $N(I)$ is simply the mean unit-profile.

- **centre of gravity (centroid)** of cluster $N(J)$

It's the weighted mean g_I of the J category-profiles belonging to the cluster $N(J)$.

$g_I = \sum_{j=1}^{J} f_j f_I^j$. It's easy to see that $g_I = f_I$: the centroid of $N(J)$ is simply the mean category-profile.

Remark 2 *With the two clusters $N(I)$ and $N(J)$, we have now two different standpoints from which to look at the original data, corresponding to two $I \times J$ tables. We introduce so the notion of **duality** extremely important in correspondence analysis. Having now two tables instead of one, are we really on the way of simplifying our looking at the data? It must be observed that for any unit-profile f_J^i we have $\sum_{j=1}^{J} f_j^i = 1$. Thus, all unit-profiles, when represented in the J-dim euclidian space, belong, by their end-point to the (J-1)-dim unit simplex, the same for their centroid. Then the analysis of the cluster $N(I)$ can in fact be done in a (J-1)-dim subspace. Mutatis mutandis, the cluster $N(J)$ can be analyzed in a (I-1)-dim subspace. Thus, with a very simple data transformation, we have already reduced the number of dimensions relevant for data analysis.*

- Case study: the population unit profiles, the cluster $N(I)$, the centroid of $N(I)$, are given in Annex A, table "Row Profiles", pp. 4-6. The analogous elements for the category profiles are given in the table "Column Profiles", pp. 6-9.

6 Data analysis

We will now proceed to the statistical analysis of the data transformed in profiles. The analysis will be done and presented for the cluster of population units $N(I)$, but it is immediately transposable for the cluster of categories $N(J)$. At the end, the close link between both analysis, due to the duality, will clearly appear.

6.1 χ^2-Distance between profiles

The intuitive comparaison we can make between two unit-profiles i and i' needs to be formalized in a numerical measure. The distance uses in correspondence analysis goes back to the great statistician Pearson, who invented it sixty years ago to compare a sampling distribution with a theoretical probability distribution: the **chi-2 distance**, also called the **distributional distance**:

$$d^2\left(f_J^i, f_J^{i'}\right) = \sum_{j=1}^{J}\left(\frac{1}{f_j}\right)\left(f_j^i - f_j^{i'}\right)^2$$

The χ^2-distance is thus the usual distance in the I-dim euclidean space, but with a weight on axes (categories), inversely proportional to its mass. We still have a metric space.

- invariance property

If two columns (categories) are proportional, i.e have the same structure, if we replace them by a unique one, sum of both, then the distance between two lines (population units) remains unchanged.

6.2 Inertia of the cluster $N(I)$ to its centre of gravity g_J

We need also to summarize the whole variability observed in the cluster of population units $N(I)$. This is done with the concept of **inertia**, built with the χ^2-distance of each profile to the centre of gravity.

$$I_G\left[N(I)\right] = \sum_{i=1}^{I} f_i d^2\left(g_J, f_J^i\right)$$

Thus, the inertia of the cluster $N(I)$ to its centre of gravity g_J is the weighted mean of the individual profiles distances to g_J, the weight being the mass of each profile.

- Case study

The total inertia for our cluster $N(I)$ of 120 commune-profiles is 2,000, as given in the table "Summary", Annex A, p. 9.

6.3 Additive disaggregation of the total inertia

6.3.1 Normal subspaces through the centroid g_J

- In the (J-1)-simplex where lies the cluster $N(I)$, let's take any straight line $\triangle$ through the centroid g_J. In the same simplex, the (J-2)-dim subspace normal (perpendicular) to $\triangle$ is denoted $\triangle_\perp$ and called the complementary space to $\triangle$.
- $\triangle$ and $\triangle_\perp$ are thus two normal subspaces allowing to cover completely the (J-1) simplex.

6.3.2 Projections of a profile f_J^i on $\triangle$ and $\triangle_\perp$

Any centred profile $\left(f_J^i - g_J\right)$ can be projected

- perpendicularly to $\triangle$. This point determined by this projection is notated: $pr_\triangle\left(f_J^i\right)$.
- perpendicularly to $\triangle_\perp$. This point determined by this projection is notated: $pr_{\triangle_\perp}\left(f_J^i\right)$.

It is then obvious by the Pythagoras theorem that

$$d^2\left(g_J, f_J^i\right) = d^2\left(g_J, pr_\triangle\left(f_J^i\right)\right) + d^2\left(g_J, pr_{\triangle_\perp}\left(f_J^i\right)\right) \tag{1}$$

6.3.3 Total inertia disaggregation

From equation 1, by the weighted sum on all the unit-profiles, it follows that the total inertia can be disaggregated in two terms:

$$I_G\left[N\left(I\right)\right] = I_\triangle\left[N\left(I\right)\right] + I_{\triangle_\perp}\left[N\left(I\right)\right] \tag{2}$$

So, the inertia relative to the centre of gravity is the sum of the inertia relative to $\triangle$ and of the inertia relative to $\triangle_\perp$.

6.4 The first principal axis

The disaggregation process of the preceding section suggests to look for a straigth line $\triangle$ which could maximize the inertia component $I_\triangle\left[N\left(I\right)\right]$. By rotating the line $\triangle$ through the centre of gravity in the (J-1)-simplex, the value of inertia relative to that line, $I_\triangle\left[N\left(I\right)\right]$,varies. We need a computable process to find the rotation which maximizes $I_\triangle\left[N\left(I\right)\right]$. This computable process exists since a long time in statistics. It is called **principal component analysis**. Numerically, it implies the computation of the eigenvalues of a specific numerical matrix which we will not give explicitly here.

- By using principal component analysis, the line $\triangle$ catching by itself the maximal inertia from the cluster $N\left(I\right)$ is called the **first principal (or factorial) axis.** This optimal line is then denoted $\triangle_1$. Let's denote λ_1 the square root of the eigenvalue associated to the first principal axis. The value λ_1 is usually referred to as the **singular value** relative to the first principal axis.
- An important result from statistical theory is that

$$I_{\triangle_1}\left[N\left(I\right)\right] = \lambda_1^2 \tag{3}$$

Then, the inertia relative to the first principal axis is given by the associated eigenvalue λ_1^2.

- A result from correspondence analysis with the χ^2-distance is that

$$\lambda_1 \square 1 \tag{4}$$

- Case study

In our case study, we find in the table "Summary", Annex A, p. 9, that:

the first principal axis has a singular value $\lambda_1 = 0,834$.

the eigenvalue, and thus the inertia, associated to the first principal axis is $\lambda_1^2 = 0,696$.

the proportion of the total inertia 2,000 accounted for by the first principal axis is then 0,348.

We usually say that the first principal axis **explains** 34,8% of the variability observed among the 120 population units, relatively to the 2 indicators.

6.5 The r principal (factorial) axis: complete disaggregation of inertia

Once the first principal axis $\triangle_1$ has been found, a similar process can be applied in its complementary normal subspace $\triangle_{\perp_1}$ to find the second axis $\triangle_2$, and so on repetitively until there is no more inertia to explain. Since, according to 2, the cluster $N(I)$ lies in (J-1) dimensions, the number of principal axis, denoted here by **r**, cannot exceed (J-1). But it can be much less than (J-1). In fact, it can be shown that

$$r \square \min(I-1, J-1) \tag{5}$$

- The process of finding, by repetition, all the principal axis of the cluster $N(I)$ generates the **factorial disaggregation of the total inertia**. We then have for each axis (factor) the different statistics seen for the first axis.

- As one among the numerous results of the disaggregation we have:

$$I_G[N(I)] = \sum_{\alpha=1}^{r} \lambda_\alpha^2 \tag{6}$$

- We also have

$$\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_r \tag{7}$$

- Case study

  In our case study, we find in the table "Summary", Annex A, p. 9, that:

  There are 4 factorial axis.

  The second axis accounts for 25,6% of the total inertia, so that the first two axis explain 60,4% of the variability found in the 2 indicators.

6.6 Scores in dimensions: discriminating between population units

The r factorial axis are perpendicular by construction and then constitute a cartesian axis system where each profile f_J^i has new coordinates: its r projections

$$pr_{\triangle_1}\left(f_J^i\right), pr_{\triangle_2}\left(f_J^i\right), ..., pr_{\triangle_r}\left(f_J^i\right)$$

These projections are called the "scores" of the population unit in the different dimensions.

Notation

The score of the population unit i on the factorial axis α is notated: $F_\alpha(i)$. Thus,

$$F_\alpha(i) = pr_{\triangle_\alpha}\left(f_J^i\right) \tag{8}$$

It can be shown that

$$\sum_{i=1}^{I} f_i F_\alpha(i) = 0 \tag{9}$$

Thus, the weighted distribution of the scores $\{F_\alpha(i)\}$ is centered on 0.

It can also be shown that

$$\sum_{i=1}^{I} f_i F_\alpha^2(i) = \lambda_\alpha^2 \tag{10}$$

Thus, the variance of the same distribution is given by the contribution λ_α^2 of the factorial axis α to the total inertia.

6.6.1 First dimension scores: numerical analysis

Since the first dimension accounts for the highest proportion of the total inertia (equation (7)), just looking at this first score can be considered as giving a good information on the differences between the population units. Here, we precisely see how a **data reduction technique** like **factorial correspondence analysis** facilitates the classification of population units represented in multidimensional data.

- Case study

In our case study, we find in the table "Overview Row points", Annex A, p. 10-12, that: The score in dimension 1 takes a small number, more precisely 8, different values. This is normal in this simple case since, with only 2 indicators of 3 categories each, the maximum number of different commune profiles is 9, but only 8 of these profiles are in fact found in the sample. The table 1 below presents the ranking of these profiles.

Table 1 Ranking of communes according to score in dimension 1

Score in dim. 1	% land quality 1	primary enrolment rate	# communes	communes id
1,824	0%	< 80%	19	1,3,12,15 etc.
0,818	0%	80% - 90%	8	7,13,84,85 etc.
0,763	□ 25%	< 80%	3	50,66,83
0,262	0%	> 90%	30	4,5,8,9,11 etc.
-0,242	□ 25%	80% - 90%	5	6,56,70,103,104
-0,577	> 25%	80% - 90%	2	102,109
-0,798	□ 25%	> 90%	33	2,10,16,20 etc.
-1,133	> 25%	> 90%	20	19,21,27,32 etc.

We clearly see that, according to the first axis, poverty is decreasing from the highest score (1,824) to the lowest score (-1,133). The only profile not represented in the sample is a commune having > 25% of land quality 1 and a primary enrolment rate < 80%.

The same table of Annex A, in column "Inertia", displays the contribution of each commune to the total inertia of 2: commune #1 contributes 0,023 while commune #109 contributes 0,048.

The same table displays the proportion of the inertia of the different axis which is contributed by each commune. Here, it has been requested only for the first two axis. So, commune #1 contributes 2,8% of the inertia of axis 1, while commune #109 contributes only 0,3%.

6.6.2 First and second dimension scores: graphical analysis

Instead of looking only at the scores on the first factorial axis, we can look at the two first dimensions. Then, the most useful and significant analysis is the one obtained by a graphical representation of the population units in a cartesian plane with the first factorial score reported on the x-axis, and the second factorial score reported on the y-axis. This graphical representation is not given here with the Correspondance Analysis program, since with the 120 communes, the graph is unreadable. But we will see below an interesting graphical capacity with the Multiple Correspondence Analysis (Homals) of the same data.

7 Duality in correspondence analysis: the key to composite indicators

With the preceding analysis, we can certainly begin to discriminate more clearly between the population units, but we have no explicit numerical relation between a population unit score and its profile on the set of the basic qualitative indicators. This relation is needed if we want to discriminate between a much larger set of population units which where not included in this specific factorial analysis, without having to recompute that type of analysis. Here the duality properties of correspondence analysis provide the required tools.

7.1 Analysis of the cluster $N(J)$

The preceding analysis of the cluster of population units (wards) $N(I)$ can be done for the cluster of categories (indicators) $N(J)$. The cluster $N(J)$ having been first transformed in category-profiles, the χ^2-distance between these profiles is defined the same way. From this follows the total inertia $I_G[N(J)]$, the calculation of the principal axis and the associated singular values, and the disaggregation of the total inertia as the sum of the principal axis inertia (eigenvalues). The beauty of the theory, due to the χ^2-distance definition, is that:

- the total inertia is the same: $I_G[N(J)] = I_G[N(I)]$
- the r singular values λ_α are the same,
- the disaggregation of the total inertia is the same $I_G[N(J)] = \sum_{\alpha=1}^{r} \lambda_\alpha^2$.

The only new element is that instead of having population unit scores, we now have category scores relative to the r factorial axis of the cluster $N(J)$. For the category j, these scores are notated $G_\alpha(j)$ and we have $G_\alpha(j) = pr_{\triangle_\alpha}\left(f_I^j\right)$. By comparing these scores, especially first one, we can see the "proximity" of different categories. Two categories having similar scores can be considered as closely correlated, and then, by this type of analysis, we have a means of eliminating redundant categories and indicators, and thus to reduce the number of indicators needed to describe our population units.

- Case study

  In our case study, we find in the table "Overview Column Points", Annex A, page 16, the value and the analysis of the score values of the 6 categories corresponding to the 2 primary indicators, for the first two factorial axis.

- the interpretation of the factorial axis from the graphical presentation

  Graphical analysis of the categories and corresponding indicators, in the two first factorial axis, is essential for understanding the meaning of these axis. More precisely, is there any poverty meaning to these axis? The relative position of the categories in such a graphical presentation reveals the underlying meaning of the axis, if there is any.

- Case study

  From the two dimensions graph given in Annex A, p. 17, we see obviously that the first axis discriminates between poorest and richest communes, according to both indicators here retained.

7.2 Linkage between both analysis: the basic duality equation

Between the only two different components of the two analysis of clusters $N(I)$ and $N(J)$, the factorial scores of population units and of primary indicators, it is shown that the following relation holds:

$$F_\alpha\left(i\right)=\sum_{j=1}^{J}f_j^i\times\frac{G_\alpha\left(j\right)}{\lambda_\alpha^2} \tag{11}$$

The equation 11 says that the factor-α score of unit i is given by multiplying its category-profile by the factor-α scores of all the ctegories, divided by the inertia (eigen) value λ_α^2 relative to this factorial axis. This is the nicest dual relation in factorial correspondence analysis: it really opens the way to build the synthetic indicator we are looking for, on a scientific basis. More than that, this relation gives us, by the relative values of the scores $G_\alpha\left(j\right)$ obtained by the categories, the "poverty dimension" represented by the factorial axis α.

7.3 Normalization and the composite indicator

From equation 11, we see that the relation between the factor-α score of unit i and the set $\{G_\alpha\left(j\right)\}$ of the indicators scores on axis α requires that these scores be deflated by the inertia (eigen) value λ_α^2,which is also the variance of the distribution of $\{G_\alpha\left(j\right)\}$ according to equation 10. It appears immediately that if we normalize the scores of the categories generated by the primary indicators, i.e. if we divide each score $G_\alpha\left(j\right)$ by λ_α^2, the relation between these catetories normalized scores and the population unit scores will be direct. So, let's define the **normalized scores** of indicator j as

$$G^*_\alpha(j) = \frac{G_\alpha(j)}{\lambda^2_\alpha} \tag{12}$$

We then have

$$F_\alpha(i) = \sum_{j=1}^{J} f^i_j \times G^*_\alpha(j) \tag{13}$$

7.3.1 Statistical definition of the composite indicator

On the basis of the objective approach recognized in the factorial correspondence analysis and of its capacity to effectively generate a simple composite indicator structure from multidimensional qualitative data, we suggest as a serious composite indicator candidate the one defined by equation 13, **for the first factorial axis**. So, the **weight** to be given to any category of a primary qualitative indicator would be its **normalized score on the first factorial axis**, as given in equation 12.

Definition 3 *A composite indicator of multiple qualitative poverty indicators, each defined as a finite set of categories, for different population units, is given by*

1. computing the profile of the population unit relatively to these primary indicators

2. applying to this profile the category-weights given by the normalized scores of these indicators on the first factorial axis coming out of correspondence analysis.

8 Multiple Correspondence Analysis

- Correspondence Analysis is a general data reduction technique applicable to the analysis of any matrix of non negative numbers. We have used it, as an example, to the analysis of two categorical variables, for us taken as poverty indicators. From this point of view, it is simply a particular case of the Multiple Correspondence Analysis (MCA), which allows to consider simultaneously any number of categorical variables. From the computational side, MCA is obtained by running a CA analysis of 0-1 indicator matrix associated to the set of categorical indicators[1].

[1] Equivalently, MCA is a CA applied to the Burt matrix of all contingency tables built from the indicators. See [2], chapter 7.

- To illustrate the specificity and the interest of MCA, we have run here, on the same data, the SPSS program HOMALS, precisely the one that computes a MCA. The output is presented in Annex B.

- We first observe the complete convergence regarding the eigenvalue (inertia) relative to each axis.

- There are important differences between the respective outputs of CA and MCA:

MCA does not produce the Correspondence Table neither the Row and Column Profiles, here taken as too simple, due to the fact that we have a 0-1 initial matrix.

MCA names "quantifications" the column (category) scores in the different dimensions, and "object scores" the row scores. Again here, we see the convergence in the scores provided by both analysis. But it does not present the inertia relative to each point (row or column).

On the other hand, MCA keeps the individuality of each indicator as a subset of the whole set of categories (columns), and presents the marginal frequencies observed for each indicator. It gives also an additional information for each indicator, its "Discrimination Measure" in each dimension, which is the variance of the quantified indicator (its different scores) in each dimension. In that sense, it is really a measure of the discrimination power of each indicator, in each dimension.

This individuality of each indicator allows to produce a graph as the one given in Annex B, p. 4, where the categories of each indicator can be linked with a line, making here quite evident the poverty meaning of the axis 1.

We notice here, Annex B, p. 7, the possibility of having a graph representing the 8 profiles taken by the 120 communes, profiles described in Table 1 above.

References

[1] Benzécri, J.P and F., *L'analyse des données, Analyse des correspondances, Exposé élémentaire,* Dunod 1980, 424 p.

[2] Greenacre, M. J., *Theory and Applications of Correspondence Analysis,* Academic Press 1984, 364 p.

Annex A

SPSS Output of Correspondence Analysis

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Two indicators: proportion of land of quality 1, primary school enrolment rate

Credit

CORRESPONDENCE
Version 1.0
by
Data Theory Scaling System Group (DTSS)
Faculty of Social and Behavioral Sciences
Leiden University, The Netherlands

Correspondence Table

	Column						
Row	Land1 >25	Land1 <=25	Land1 =0	Rate >90	Rate 80-90	Rate <80	Active Margin
1	0	0	1	0	0	1	2
2	0	1	0	1	0	0	2
3	0	0	1	0	0	1	2
4	0	0	1	1	0	0	2
5	0	0	1	1	0	0	2
6	0	1	0	0	1	0	2
7	0	0	1	0	1	0	2
8	0	0	1	1	0	0	2
9	0	0	1	1	0	0	2
10	0	1	0	1	0	0	2
11	0	0	1	1	0	0	2
12	0	0	1	0	0	1	2
13	0	0	1	0	1	0	2
14	0	0	1	1	0	0	2
15	0	0	1	0	0	1	2
16	0	1	0	1	0	0	2
17	0	0	1	1	0	0	2
18	0	0	1	1	0	0	2
19	1	0	0	1	0	0	2
20	0	1	0	1	0	0	2
21	1	0	0	1	0	0	2
22	0	1	0	1	0	0	2
23	0	0	1	0	0	1	2
24	0	1	0	1	0	0	2
25	0	0	1	1	0	0	2
26	0	1	0	1	0	0	2
27	1	0	0	1	0	0	2
28	0	0	1	1	0	0	2
29	0	1	0	1	0	0	2
30	0	0	1	1	0	0	2
31	0	1	0	1	0	0	2

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Correspondence Table

Row	Column						
	Land1 >25	Land1 <=25	Land1 =0	Rate >90	Rate 80-90	Rate <80	Active Margin
32	1	0	0	1	0	0	2
33	1	0	0	1	0	0	2
34	1	0	0	1	0	0	2
35	0	1	0	1	0	0	2
36	1	0	0	1	0	0	2
37	1	0	0	1	0	0	2
38	1	0	0	1	0	0	2
39	0	1	0	1	0	0	2
40	0	1	0	1	0	0	2
41	0	1	0	1	0	0	2
42	1	0	0	1	0	0	2
43	0	0	1	1	0	0	2
44	1	0	0	1	0	0	2
45	1	0	0	1	0	0	2
46	0	0	1	1	0	0	2
47	0	1	0	1	0	0	2
48	0	1	0	1	0	0	2
49	0	1	0	1	0	0	2
50	0	1	0	0	0	1	2
51	0	0	1	1	0	0	2
52	1	0	0	1	0	0	2
53	0	1	0	1	0	0	2
54	0	0	1	1	0	0	2
55	0	1	0	1	0	0	2
56	0	1	0	0	1	0	2
57	0	1	0	1	0	0	2
58	0	1	0	1	0	0	2
59	0	1	0	1	0	0	2
60	0	1	0	1	0	0	2
61	0	0	1	1	0	0	2
62	0	0	1	1	0	0	2
63	1	0	0	1	0	0	2
64	0	0	1	1	0	0	2
65	0	0	1	1	0	0	2
66	0	1	0	0	0	1	2
67	0	0	1	1	0	0	2
68	0	1	0	1	0	0	2
69	0	0	1	0	0	1	2
70	0	1	0	0	1	0	2
71	1	0	0	1	0	0	2
72	0	1	0	1	0	0	2
73	0	0	1	0	0	1	2
74	0	1	0	1	0	0	2
75	0	0	1	0	0	1	2
76	0	1	0	1	0	0	2
77	0	1	0	1	0	0	2

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Correspondence Table

	Column						
Row	Land1 >25	Land1 <=25	Land1 =0	Rate >90	Rate 80-90	Rate <80	Active Margin
78	0	1	0	1	0	0	2
79	1	0	0	1	0	0	2
80	0	1	0	1	0	0	2
81	1	0	0	1	0	0	2
82	1	0	0	1	0	0	2
83	0	1	0	0	0	1	2
84	0	0	1	0	1	0	2
85	0	0	1	0	1	0	2
86	0	0	1	1	0	0	2
87	0	0	1	1	0	0	2
88	0	0	1	0	1	0	2
89	0	0	1	1	0	0	2
90	0	0	1	0	0	1	2
91	1	0	0	1	0	0	2
92	0	0	1	0	0	1	2
93	0	1	0	1	0	0	2
94	0	0	1	1	0	0	2
95	0	1	0	1	0	0	2
96	0	0	1	0	1	0	2
97	0	0	1	0	0	1	2
98	1	0	0	1	0	0	2
99	0	0	1	1	0	0	2
100	0	0	1	0	0	1	2
101	0	0	1	1	0	0	2
102	1	0	0	0	1	0	2
103	0	1	0	0	1	0	2
104	0	1	0	0	1	0	2
105	0	0	1	1	0	0	2
106	0	0	1	0	0	1	2
107	0	0	1	1	0	0	2
108	0	0	1	1	0	0	2
109	1	0	0	0	1	0	2
110	0	0	1	0	1	0	2
111	0	0	1	0	1	0	2
112	0	0	1	0	0	1	2
113	0	0	1	0	0	1	2
114	0	1	0	1	0	0	2
115	0	1	0	1	0	0	2
116	0	0	1	0	0	1	2
117	0	0	1	0	0	1	2
118	0	0	1	1	0	0	2
119	0	0	1	0	0	1	2
120	0	0	1	0	0	1	2
Active Margin	22	41	57	83	15	22	240

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Row Profiles

	Column						
Row	Land1 >25	Land1 <=25	Land1 =0	Rate >90	Rate 80-90	Rate <80	Active Margin
1	,000	,000	,500	,000	,000	,500	1,000
2	,000	,500	,000	,500	,000	,000	1,000
3	,000	,000	,500	,000	,000	,500	1,000
4	,000	,000	,500	,500	,000	,000	1,000
5	,000	,000	,500	,500	,000	,000	1,000
6	,000	,500	,000	,000	,500	,000	1,000
7	,000	,000	,500	,000	,500	,000	1,000
8	,000	,000	,500	,500	,000	,000	1,000
9	,000	,000	,500	,500	,000	,000	1,000
10	,000	,500	,000	,500	,000	,000	1,000
11	,000	,000	,500	,500	,000	,000	1,000
12	,000	,000	,500	,000	,000	,500	1,000
13	,000	,000	,500	,000	,500	,000	1,000
14	,000	,000	,500	,500	,000	,000	1,000
15	,000	,000	,500	,000	,000	,500	1,000
16	,000	,500	,000	,500	,000	,000	1,000
17	,000	,000	,500	,500	,000	,000	1,000
18	,000	,000	,500	,500	,000	,000	1,000
19	,500	,000	,000	,500	,000	,000	1,000
20	,000	,500	,000	,500	,000	,000	1,000
21	,500	,000	,000	,500	,000	,000	1,000
22	,000	,500	,000	,500	,000	,000	1,000
23	,000	,000	,500	,000	,000	,500	1,000
24	,000	,500	,000	,500	,000	,000	1,000
25	,000	,000	,500	,500	,000	,000	1,000
26	,000	,500	,000	,500	,000	,000	1,000
27	,500	,000	,000	,500	,000	,000	1,000
28	,000	,000	,500	,500	,000	,000	1,000
29	,000	,500	,000	,500	,000	,000	1,000
30	,000	,000	,500	,500	,000	,000	1,000
31	,000	,500	,000	,500	,000	,000	1,000
32	,500	,000	,000	,500	,000	,000	1,000
33	,500	,000	,000	,500	,000	,000	1,000
34	,500	,000	,000	,500	,000	,000	1,000
35	,000	,500	,000	,500	,000	,000	1,000
36	,500	,000	,000	,500	,000	,000	1,000
37	,500	,000	,000	,500	,000	,000	1,000
38	,500	,000	,000	,500	,000	,000	1,000
39	,000	,500	,000	,500	,000	,000	1,000
40	,000	,500	,000	,500	,000	,000	1,000
41	,000	,500	,000	,500	,000	,000	1,000
42	,500	,000	,000	,500	,000	,000	1,000
43	,000	,000	,500	,500	,000	,000	1,000
44	,500	,000	,000	,500	,000	,000	1,000
45	,500	,000	,000	,500	,000	,000	1,000
46	,000	,000	,500	,500	,000	,000	1,000

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Row Profiles

Row	Column						
	Land1 >25	Land1 <=25	Land1 =0	Rate >90	Rate 80-90	Rate <80	Active Margin
47	,000	,500	,000	,500	,000	,000	1,000
48	,000	,500	,000	,500	,000	,000	1,000
49	,000	,500	,000	,500	,000	,000	1,000
50	,000	,500	,000	,000	,000	,500	1,000
51	,000	,000	,500	,500	,000	,000	1,000
52	,500	,000	,000	,500	,000	,000	1,000
53	,000	,500	,000	,500	,000	,000	1,000
54	,000	,000	,500	,500	,000	,000	1,000
55	,000	,500	,000	,500	,000	,000	1,000
56	,000	,500	,000	,000	,500	,000	1,000
57	,000	,500	,000	,500	,000	,000	1,000
58	,000	,500	,000	,500	,000	,000	1,000
59	,000	,500	,000	,500	,000	,000	1,000
60	,000	,500	,000	,500	,000	,000	1,000
61	,000	,000	,500	,500	,000	,000	1,000
62	,000	,000	,500	,500	,000	,000	1,000
63	,500	,000	,000	,500	,000	,000	1,000
64	,000	,000	,500	,500	,000	,000	1,000
65	,000	,000	,500	,500	,000	,000	1,000
66	,000	,500	,000	,000	,000	,500	1,000
67	,000	,000	,500	,500	,000	,000	1,000
68	,000	,500	,000	,500	,000	,000	1,000
69	,000	,000	,500	,000	,000	,500	1,000
70	,000	,500	,000	,000	,500	,000	1,000
71	,500	,000	,000	,500	,000	,000	1,000
72	,000	,500	,000	,500	,000	,000	1,000
73	,000	,000	,500	,000	,000	,500	1,000
74	,000	,500	,000	,500	,000	,000	1,000
75	,000	,000	,500	,000	,000	,500	1,000
76	,000	,500	,000	,500	,000	,000	1,000
77	,000	,500	,000	,500	,000	,000	1,000
78	,000	,500	,000	,500	,000	,000	1,000
79	,500	,000	,000	,500	,000	,000	1,000
80	,000	,500	,000	,500	,000	,000	1,000
81	,500	,000	,000	,500	,000	,000	1,000
82	,500	,000	,000	,500	,000	,000	1,000
83	,000	,500	,000	,000	,000	,500	1,000
84	,000	,000	,500	,000	,500	,000	1,000
85	,000	,000	,500	,000	,500	,000	1,000
86	,000	,000	,500	,500	,000	,000	1,000
87	,000	,000	,500	,500	,000	,000	1,000
88	,000	,000	,500	,000	,500	,000	1,000
89	,000	,000	,500	,500	,000	,000	1,000
90	,000	,000	,500	,000	,000	,500	1,000
91	,500	,000	,000	,500	,000	,000	1,000
92	,000	,000	,500	,000	,000	,500	1,000

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Row Profiles

	Column						
Row	Land1 >25	Land1 <=25	Land1 =0	Rate >90	Rate 80-90	Rate <80	Active Margin
93	,000	,500	,000	,500	,000	,000	1,000
94	,000	,000	,500	,500	,000	,000	1,000
95	,000	,500	,000	,500	,000	,000	1,000
96	,000	,000	,500	,000	,500	,000	1,000
97	,000	,000	,500	,000	,000	,500	1,000
98	,500	,000	,000	,500	,000	,000	1,000
99	,000	,000	,500	,500	,000	,000	1,000
100	,000	,000	,500	,000	,000	,500	1,000
101	,000	,000	,500	,500	,000	,000	1,000
102	,500	,000	,000	,000	,500	,000	1,000
103	,000	,500	,000	,000	,500	,000	1,000
104	,000	,500	,000	,000	,500	,000	1,000
105	,000	,000	,500	,500	,000	,000	1,000
106	,000	,000	,500	,000	,000	,500	1,000
107	,000	,000	,500	,500	,000	,000	1,000
108	,000	,000	,500	,500	,000	,000	1,000
109	,500	,000	,000	,000	,500	,000	1,000
110	,000	,000	,500	,000	,500	,000	1,000
111	,000	,000	,500	,000	,500	,000	1,000
112	,000	,000	,500	,000	,000	,500	1,000
113	,000	,000	,500	,000	,000	,500	1,000
114	,000	,500	,000	,500	,000	,000	1,000
115	,000	,500	,000	,500	,000	,000	1,000
116	,000	,000	,500	,000	,000	,500	1,000
117	,000	,000	,500	,000	,000	,500	1,000
118	,000	,000	,500	,500	,000	,000	1,000
119	,000	,000	,500	,000	,000	,500	1,000
120	,000	,000	,500	,000	,000	,500	1,000
Mass	,092	,171	,238	,346	,063	,092	

Column Profiles

	Column						
Row	Land1 >25	Land1 <=25	Land1 =0	Rate >90	Rate 80-90	Rate <80	Mass
1	,000	,000	,018	,000	,000	,045	,008
2	,000	,024	,000	,012	,000	,000	,008
3	,000	,000	,018	,000	,000	,045	,008
4	,000	,000	,018	,012	,000	,000	,008
5	,000	,000	,018	,012	,000	,000	,008
6	,000	,024	,000	,000	,067	,000	,008
7	,000	,000	,018	,000	,067	,000	,008
8	,000	,000	,018	,012	,000	,000	,008
9	,000	,000	,018	,012	,000	,000	,008
10	,000	,024	,000	,012	,000	,000	,008
11	,000	,000	,018	,012	,000	,000	,008

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Column Profiles

	Column						
Row	Land1 >25	Land1 <=25	Land1 =0	Rate >90	Rate 80-90	Rate <80	Mass
12	,000	,000	,018	,000	,000	,045	,008
13	,000	,000	,018	,000	,067	,000	,008
14	,000	,000	,018	,012	,000	,000	,008
15	,000	,000	,018	,000	,000	,045	,008
16	,000	,024	,000	,012	,000	,000	,008
17	,000	,000	,018	,012	,000	,000	,008
18	,000	,000	,018	,012	,000	,000	,008
19	,045	,000	,000	,012	,000	,000	,008
20	,000	,024	,000	,012	,000	,000	,008
21	,045	,000	,000	,012	,000	,000	,008
22	,000	,024	,000	,012	,000	,000	,008
23	,000	,000	,018	,000	,000	,045	,008
24	,000	,024	,000	,012	,000	,000	,008
25	,000	,000	,018	,012	,000	,000	,008
26	,000	,024	,000	,012	,000	,000	,008
27	,045	,000	,000	,012	,000	,000	,008
28	,000	,000	,018	,012	,000	,000	,008
29	,000	,024	,000	,012	,000	,000	,008
30	,000	,000	,018	,012	,000	,000	,008
31	,000	,024	,000	,012	,000	,000	,008
32	,045	,000	,000	,012	,000	,000	,008
33	,045	,000	,000	,012	,000	,000	,008
34	,045	,000	,000	,012	,000	,000	,008
35	,000	,024	,000	,012	,000	,000	,008
36	,045	,000	,000	,012	,000	,000	,008
37	,045	,000	,000	,012	,000	,000	,008
38	,045	,000	,000	,012	,000	,000	,008
39	,000	,024	,000	,012	,000	,000	,008
40	,000	,024	,000	,012	,000	,000	,008
41	,000	,024	,000	,012	,000	,000	,008
42	,045	,000	,000	,012	,000	,000	,008
43	,000	,000	,018	,012	,000	,000	,008
44	,045	,000	,000	,012	,000	,000	,008
45	,045	,000	,000	,012	,000	,000	,008
46	,000	,000	,018	,012	,000	,000	,008
47	,000	,024	,000	,012	,000	,000	,008
48	,000	,024	,000	,012	,000	,000	,008
49	,000	,024	,000	,012	,000	,000	,008
50	,000	,024	,000	,000	,000	,045	,008
51	,000	,000	,018	,012	,000	,000	,008
52	,045	,000	,000	,012	,000	,000	,008
53	,000	,024	,000	,012	,000	,000	,008
54	,000	,000	,018	,012	,000	,000	,008
55	,000	,024	,000	,012	,000	,000	,008
56	,000	,024	,000	,000	,067	,000	,008
57	,000	,024	,000	,012	,000	,000	,008

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Column Profiles

Row	Column: Land1 >25	Land1 <=25	Land1 =0	Rate >90	Rate 80-90	Rate <80	Mass
58	,000	,024	,000	,012	,000	,000	,008
59	,000	,024	,000	,012	,000	,000	,008
60	,000	,024	,000	,012	,000	,000	,008
61	,000	,000	,018	,012	,000	,000	,008
62	,000	,000	,018	,012	,000	,000	,008
63	,045	,000	,000	,012	,000	,000	,008
64	,000	,000	,018	,012	,000	,000	,008
65	,000	,000	,018	,012	,000	,000	,008
66	,000	,024	,000	,000	,000	,045	,008
67	,000	,000	,018	,012	,000	,000	,008
68	,000	,024	,000	,012	,000	,000	,008
69	,000	,000	,018	,000	,000	,045	,008
70	,000	,024	,000	,000	,067	,000	,008
71	,045	,000	,000	,012	,000	,000	,008
72	,000	,024	,000	,012	,000	,000	,008
73	,000	,000	,018	,000	,000	,045	,008
74	,000	,024	,000	,012	,000	,000	,008
75	,000	,000	,018	,000	,000	,045	,008
76	,000	,024	,000	,012	,000	,000	,008
77	,000	,024	,000	,012	,000	,000	,008
78	,000	,024	,000	,012	,000	,000	,008
79	,045	,000	,000	,012	,000	,000	,008
80	,000	,024	,000	,012	,000	,000	,008
81	,045	,000	,000	,012	,000	,000	,008
82	,045	,000	,000	,012	,000	,000	,008
83	,000	,024	,000	,000	,000	,045	,008
84	,000	,000	,018	,000	,067	,000	,008
85	,000	,000	,018	,000	,067	,000	,008
86	,000	,000	,018	,012	,000	,000	,008
87	,000	,000	,018	,012	,000	,000	,008
88	,000	,000	,018	,000	,067	,000	,008
89	,000	,000	,018	,012	,000	,000	,008
90	,000	,000	,018	,000	,000	,045	,008
91	,045	,000	,000	,012	,000	,000	,008
92	,000	,000	,018	,000	,000	,045	,008
93	,000	,024	,000	,012	,000	,000	,008
94	,000	,000	,018	,012	,000	,000	,008
95	,000	,024	,000	,012	,000	,000	,008
96	,000	,000	,018	,000	,067	,000	,008
97	,000	,000	,018	,000	,000	,045	,008
98	,045	,000	,000	,012	,000	,000	,008
99	,000	,000	,018	,012	,000	,000	,008
100	,000	,000	,018	,000	,000	,045	,008
101	,000	,000	,018	,012	,000	,000	,008
102	,045	,000	,000	,000	,067	,000	,008
103	,000	,024	,000	,000	,067	,000	,008

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Column Profiles

Row	Column Land1 >25	Land1 <=25	Land1 =0	Rate >90	Rate 80-90	Rate <80	Mass
104	,000	,024	,000	,000	,067	,000	,008
105	,000	,000	,018	,012	,000	,000	,008
106	,000	,000	,018	,000	,000	,045	,008
107	,000	,000	,018	,012	,000	,000	,008
108	,000	,000	,018	,012	,000	,000	,008
109	,045	,000	,000	,000	,067	,000	,008
110	,000	,000	,018	,000	,067	,000	,008
111	,000	,000	,018	,000	,067	,000	,008
112	,000	,000	,018	,000	,000	,045	,008
113	,000	,000	,018	,000	,000	,045	,008
114	,000	,024	,000	,012	,000	,000	,008
115	,000	,024	,000	,012	,000	,000	,008
116	,000	,000	,018	,000	,000	,045	,008
117	,000	,000	,018	,000	,000	,045	,008
118	,000	,000	,018	,012	,000	,000	,008
119	,000	,000	,018	,000	,000	,045	,008
120	,000	,000	,018	,000	,000	,045	,008
Active Margin	1,000	1,000	1,000	1,000	1,000	1,000	

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia Accounted for	Proportion of Inertia Cumulative
1	,834	,696			,348	,348
2	,715	,512			,256	,604
3	,699	,488			,244	,848
4	,551	,304			,152	1,000
Total		2,000	480,000	1,000[a]	1,000	1,000

Summary

Dimension	Confidence Singular Value Standard Deviation	Confidence Singular Value Correlation 2
1	,017	,057
2	,035	
3		
4		
Total		

a. 595 degrees of freedom

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Overview Row Points[a]

Row	Mass	Score in Dimension		Inertia	Contribution: Of Point to Inertia of Dimension	
		1	2		1	2
1	,008	1,824	-,587	,023	,028	,003
2	,008	-,798	,608	,010	,005	,003
3	,008	1,824	-,587	,023	,028	,003
4	,008	,262	-,355	,006	,001	,001
5	,008	,262	-,355	,006	,001	,001
6	,008	-,242	2,656	,037	,000	,059
7	,008	,818	1,693	,034	,006	,024
8	,008	,262	-,355	,006	,001	,001
9	,008	,262	-,355	,006	,001	,001
10	,008	-,798	,608	,010	,005	,003
11	,008	,262	-,355	,006	,001	,001
12	,008	1,824	-,587	,023	,028	,003
13	,008	,818	1,693	,034	,006	,024
14	,008	,262	-,355	,006	,001	,001
15	,008	1,824	-,587	,023	,028	,003
16	,008	-,798	,608	,010	,005	,003
17	,008	,262	-,355	,006	,001	,001
18	,008	,262	-,355	,006	,001	,001
19	,008	-1,133	-1,377	,020	,011	,016
20	,008	-,798	,608	,010	,005	,003
21	,008	-1,133	-1,377	,020	,011	,016
22	,008	-,798	,608	,010	,005	,003
23	,008	1,824	-,587	,023	,028	,003
24	,008	-,798	,608	,010	,005	,003
25	,008	,262	-,355	,006	,001	,001
26	,008	-,798	,608	,010	,005	,003
27	,008	-1,133	-1,377	,020	,011	,016
28	,008	,262	-,355	,006	,001	,001
29	,008	-,798	,608	,010	,005	,003
30	,008	,262	-,355	,006	,001	,001
31	,008	-,798	,608	,010	,005	,003
32	,008	-1,133	-1,377	,020	,011	,016
33	,008	-1,133	-1,377	,020	,011	,016
34	,008	-1,133	-1,377	,020	,011	,016
35	,008	-,798	,608	,010	,005	,003
36	,008	-1,133	-1,377	,020	,011	,016
37	,008	-1,133	-1,377	,020	,011	,016
38	,008	-1,133	-1,377	,020	,011	,016
39	,008	-,798	,608	,010	,005	,003
40	,008	-,798	,608	,010	,005	,003
41	,008	-,798	,608	,010	,005	,003
42	,008	-1,133	-1,377	,020	,011	,016
43	,008	,262	-,355	,006	,001	,001
44	,008	-1,133	-1,377	,020	,011	,016

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Overview Row Points[a]

Row	Mass	Score in Dimension		Inertia	Contribution: Of Point to Inertia of Dimension	
		1	2		1	2
45	,008	-1,133	-1,377	,020	,011	,016
46	,008	,262	-,355	,006	,001	,001
47	,008	-,798	,608	,010	,005	,003
48	,008	-,798	,608	,010	,005	,003
49	,008	-,798	,608	,010	,005	,003
50	,008	,763	,376	,027	,005	,001
51	,008	,262	-,355	,006	,001	,001
52	,008	-1,133	-1,377	,020	,011	,016
53	,008	-,798	,608	,010	,005	,003
54	,008	,262	-,355	,006	,001	,001
55	,008	-,798	,608	,010	,005	,003
56	,008	-,242	2,656	,037	,000	,059
57	,008	-,798	,608	,010	,005	,003
58	,008	-,798	,608	,010	,005	,003
59	,008	-,798	,608	,010	,005	,003
60	,008	-,798	,608	,010	,005	,003
61	,008	,262	-,355	,006	,001	,001
62	,008	,262	-,355	,006	,001	,001
63	,008	-1,133	-1,377	,020	,011	,016
64	,008	,262	-,355	,006	,001	,001
65	,008	,262	-,355	,006	,001	,001
66	,008	,763	,376	,027	,005	,001
67	,008	,262	-,355	,006	,001	,001
68	,008	-,798	,608	,010	,005	,003
69	,008	1,824	-,587	,023	,028	,003
70	,008	-,242	2,656	,037	,000	,059
71	,008	-1,133	-1,377	,020	,011	,016
72	,008	-,798	,608	,010	,005	,003
73	,008	1,824	-,587	,023	,028	,003
74	,008	-,798	,608	,010	,005	,003
75	,008	1,824	-,587	,023	,028	,003
76	,008	-,798	,608	,010	,005	,003
77	,008	-,798	,608	,010	,005	,003
78	,008	-,798	,608	,010	,005	,003
79	,008	-1,133	-1,377	,020	,011	,016
80	,008	-,798	,608	,010	,005	,003
81	,008	-1,133	-1,377	,020	,011	,016
82	,008	-1,133	-1,377	,020	,011	,016
83	,008	,763	,376	,027	,005	,001
84	,008	,818	1,693	,034	,006	,024
85	,008	,818	1,693	,034	,006	,024
86	,008	,262	-,355	,006	,001	,001
87	,008	,262	-,355	,006	,001	,001
88	,008	,818	1,693	,034	,006	,024

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Overview Row Points[a]

Row	Mass	Score in Dimension		Inertia	Contribution	
					Of Point to Inertia of Dimension	
		1	2		1	2
89	,008	,262	-,355	,006	,001	,001
90	,008	1,824	-,587	,023	,028	,003
91	,008	-1,133	-1,377	,020	,011	,016
92	,008	1,824	-,587	,023	,028	,003
93	,008	-,798	,608	,010	,005	,003
94	,008	,262	-,355	,006	,001	,001
95	,008	-,798	,608	,010	,005	,003
96	,008	,818	1,693	,034	,006	,024
97	,008	1,824	-,587	,023	,028	,003
98	,008	-1,133	-1,377	,020	,011	,016
99	,008	,262	-,355	,006	,001	,001
100	,008	1,824	-,587	,023	,028	,003
101	,008	,262	-,355	,006	,001	,001
102	,008	-,577	,671	,048	,003	,004
103	,008	-,242	2,656	,037	,000	,059
104	,008	-,242	2,656	,037	,000	,059
105	,008	,262	-,355	,006	,001	,001
106	,008	1,824	-,587	,023	,028	,003
107	,008	,262	-,355	,006	,001	,001
108	,008	,262	-,355	,006	,001	,001
109	,008	-,577	,671	,048	,003	,004
110	,008	,818	1,693	,034	,006	,024
111	,008	,818	1,693	,034	,006	,024
112	,008	1,824	-,587	,023	,028	,003
113	,008	1,824	-,587	,023	,028	,003
114	,008	-,798	,608	,010	,005	,003
115	,008	-,798	,608	,010	,005	,003
116	,008	1,824	-,587	,023	,028	,003
117	,008	1,824	-,587	,023	,028	,003
118	,008	,262	-,355	,006	,001	,001
119	,008	1,824	-,587	,023	,028	,003
120	,008	1,824	-,587	,023	,028	,003
Active Total	1,000			2,000	1,000	1,000

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Overview Row Points[a]

	Contribution		
	Of Dimension to Inertia of Point		
Row	1	2	Total
1	,833	,063	,897
2	,374	,160	,533
3	,833	,063	,897
4	,062	,083	,145
5	,062	,083	,145
6	,009	,809	,818
7	,115	,362	,477
8	,062	,083	,145
9	,062	,083	,145
10	,374	,160	,533
11	,062	,083	,145
12	,833	,063	,897
13	,115	,362	,477
14	,062	,083	,145
15	,833	,063	,897
16	,374	,160	,533
17	,062	,083	,145
18	,062	,083	,145
19	,365	,396	,761
20	,374	,160	,533
21	,365	,396	,761
22	,374	,160	,533
23	,833	,063	,897
24	,374	,160	,533
25	,062	,083	,145
26	,374	,160	,533
27	,365	,396	,761
28	,062	,083	,145
29	,374	,160	,533
30	,062	,083	,145
31	,374	,160	,533
32	,365	,396	,761
33	,365	,396	,761
34	,365	,396	,761
35	,374	,160	,533
36	,365	,396	,761
37	,365	,396	,761
38	,365	,396	,761
39	,374	,160	,533
40	,374	,160	,533
41	,374	,160	,533
42	,365	,396	,761
43	,062	,083	,145
44	,365	,396	,761

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Overview Row Points[a]

Row	Contribution		
	Of Dimension to Inertia of Point		
	1	2	Total
45	,365	,396	,761
46	,062	,083	,145
47	,374	,160	,533
48	,374	,160	,533
49	,374	,160	,533
50	,127	,023	,150
51	,062	,083	,145
52	,365	,396	,761
53	,374	,160	,533
54	,062	,083	,145
55	,374	,160	,533
56	,009	,809	,818
57	,374	,160	,533
58	,374	,160	,533
59	,374	,160	,533
60	,374	,160	,533
61	,062	,083	,145
62	,062	,083	,145
63	,365	,396	,761
64	,062	,083	,145
65	,062	,083	,145
66	,127	,023	,150
67	,062	,083	,145
68	,374	,160	,533
69	,833	,063	,897
70	,009	,809	,818
71	,365	,396	,761
72	,374	,160	,533
73	,833	,063	,897
74	,374	,160	,533
75	,833	,063	,897
76	,374	,160	,533
77	,374	,160	,533
78	,374	,160	,533
79	,365	,396	,761
80	,374	,160	,533
81	,365	,396	,761
82	,365	,396	,761
83	,127	,023	,150
84	,115	,362	,477
85	,115	,362	,477
86	,062	,083	,145
87	,062	,083	,145
88	,115	,362	,477

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Overview Row Points[a]

	Contribution		
	Of Dimension to Inertia of Point		
Row	1	2	Total
89	,062	,083	,145
90	,833	,063	,897
91	,365	,396	,761
92	,833	,063	,897
93	,374	,160	,533
94	,062	,083	,145
95	,374	,160	,533
96	,115	,362	,477
97	,833	,063	,897
98	,365	,396	,761
99	,062	,083	,145
100	,833	,063	,897
101	,062	,083	,145
102	,040	,040	,081
103	,009	,809	,818
104	,009	,809	,818
105	,062	,083	,145
106	,833	,063	,897
107	,062	,083	,145
108	,062	,083	,145
109	,040	,040	,081
110	,115	,362	,477
111	,115	,362	,477
112	,833	,063	,897
113	,833	,063	,897
114	,374	,160	,533
115	,374	,160	,533
116	,833	,063	,897
117	,833	,063	,897
118	,062	,083	,145
119	,833	,063	,897
120	,833	,063	,897
Active Total			

a. Column Principal normalization

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

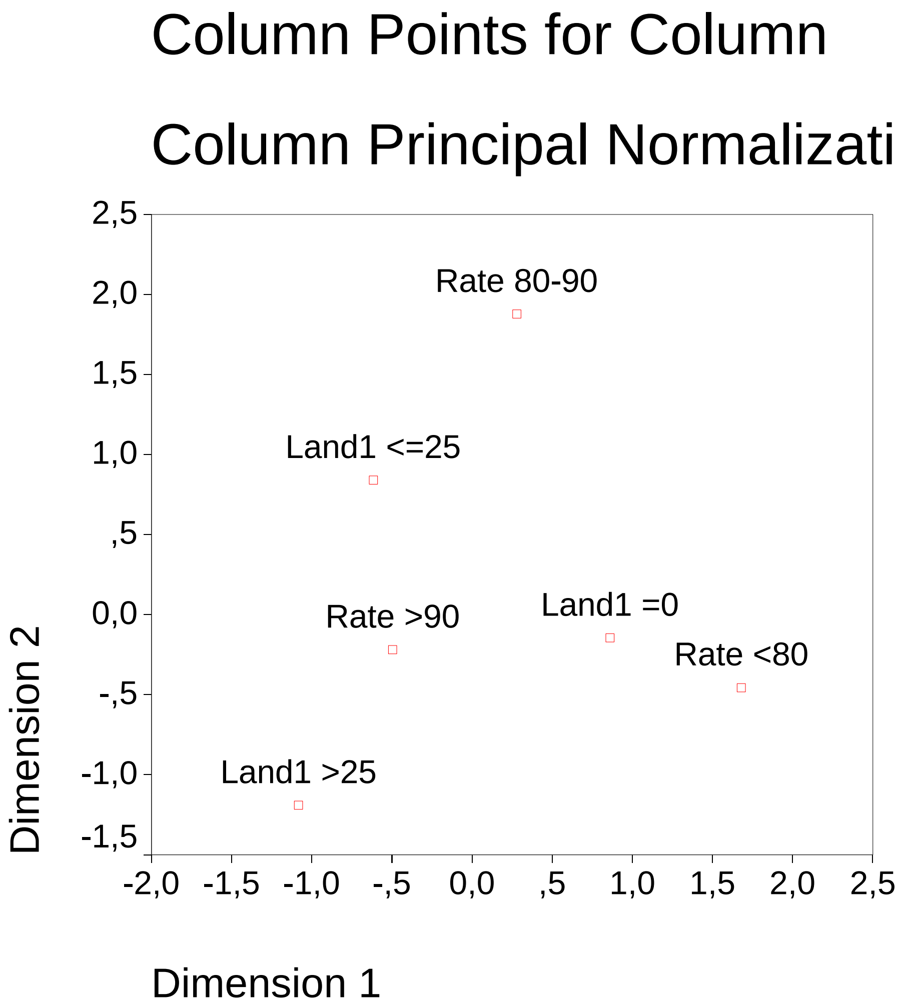
Overview Column Points[a]

		Score in Dimension			Contribution	
					Of Point to Inertia of Dimension	
Column	Mass	1	2	Inertia	1	2
Land1 >25	,092	-1,082	-1,191	,408	,154	,254
Land1 <=25	,171	-,616	,841	,329	,093	,236
Land1 =0	,238	,861	-,145	,263	,253	,010
Rate >90	,346	-,495	-,219	,154	,122	,032
Rate 80-90	,063	,279	1,878	,438	,007	,431
Rate <80	,092	1,679	-,456	,408	,371	,037
Active Total	1,000			2,000	1,000	1,000

Overview Column Points[a]

	Contribution		
	Of Dimension to Inertia of Point		
Column	1	2	Total
Land1 >25	,263	,318	,581
Land1 <=25	,197	,367	,564
Land1 =0	,670	,019	,689
Rate >90	,551	,107	,658
Rate 80-90	,011	,504	,515
Rate <80	,633	,047	,680
Active Total			

a. Column Principal normalization


Column Points for Column
Column Principal Normalization
Rate 80-90
Land1 <=25
Rate >90
Land1 =0
Rate <80
Land1 >25
2,5
2,0
1,5
1,0
,5
0,0
-,5
-1,0
-1,5
-2,0
-1,5
-1,0
-,5
0,0
,5
1,0
1,5
2,0
2,5
Dimension 2
Dimension 1

Annex B

SPSS Output of Multiple Correspondence Analysis (Homals)

Two indicators: proportion of land of quality 1, prim. school enrolment rate

Credit

HOMALS Version 1.0 by Data Theory Scaling System Group (DTSS) Faculty of Social and Behavioral Sciences Leiden University, The Netherlands

Case Processing Summary

Cases Used in Analysis	120

Marginal Frequencies

C-Proportion of land quality 1

	Marginal Frequency
Land1=0	57
Land1<=25	41
Land1>25	22
Missing	0

C-Primary enrolment rate

	Marginal Frequency
primr<80%	22
primr80-90%	15
primr>90%	83
Missing	0

Iteration History

Iteration	Fit	Difference from the Previous Iteration
1	,032030	,032030
2	,940961	,908931
3	1,059457	,118497
4	1,147416	,087958
5	1,185088	,037673
6	1,198812	,013724
7	1,204069	,005257
8	1,206265	,002196
9	1,207241	,000977
10	1,207692	,000451
11	1,207905	,000213
12	1,208007	,000102
13	1,208056	,000049
14	1,208080	,000024
15	1,208092	,000012
16[a]	1,208097	,000006

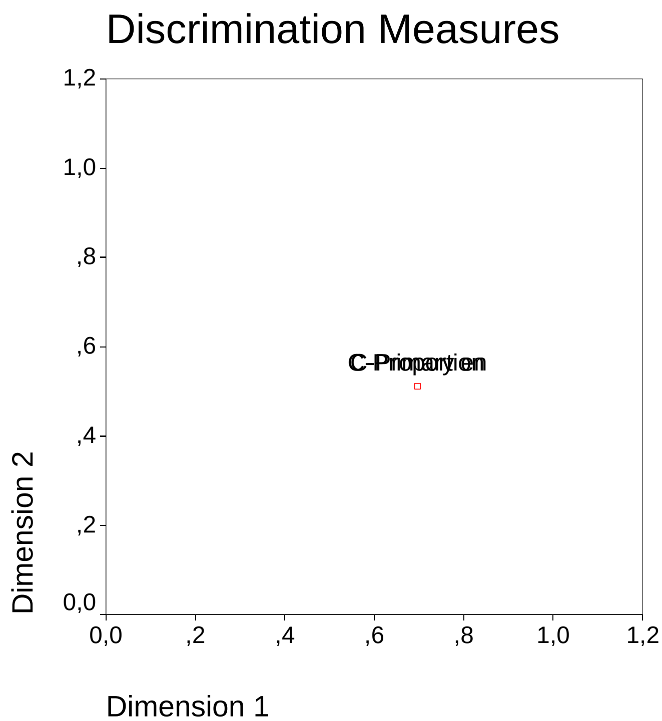
a. The iteration was terminated because convergence criteria are satisified.

Eigenvalues

Dimension	Eigenvalue
1	,696
2	,512

Discrimination Measures

	Dimension	
	1	2
C-Proportion of land quality 1	,696	,512
C-Primary enrolment rate	,696	,511



Quantifications

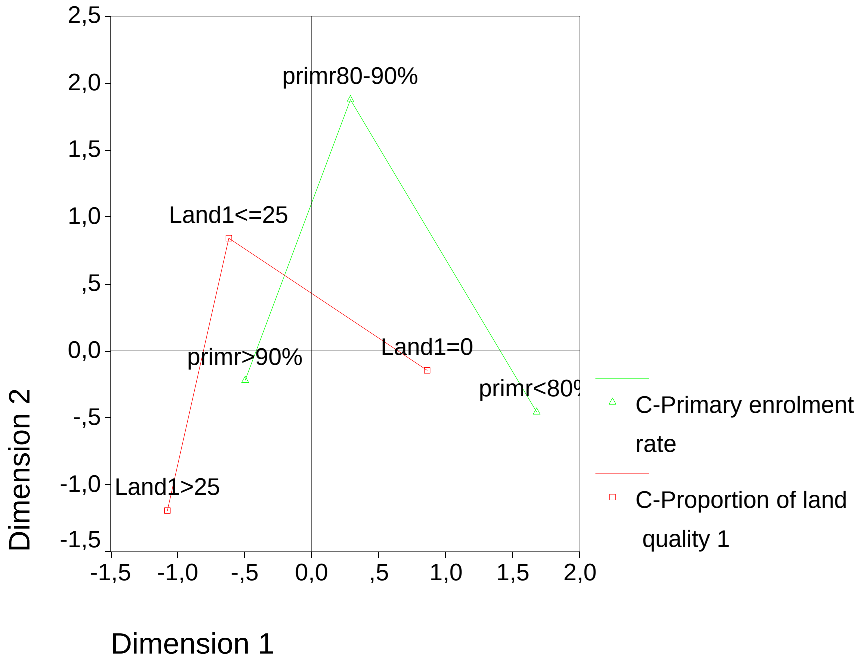
C-Proportion of land quality 1

	Marginal Frequency	Category Quantifications	
		Dimension	
		1	2
Land1=0	57	,862	-,146
Land1<=25	41	-,620	,841
Land1>25	22	-1,076	-1,191
Missing	0		

C-Primary enrolment rate

	Marginal Frequency	Category Quantifications	
		Dimension	
		1	2
primr<80%	22	1,677	-,455
primr80-90%	15	,288	1,877
primr>90%	83	-,497	-,219
Missing	0		

Quantifications



Object Scores

	Dimension	
	1	2
1	1,822	-,587
2	-,803	,609
3	1,822	-,587
4	,262	-,356
5	,262	-,356
6	-,237	2,656
7	,828	1,692
8	,262	-,356
9	,262	-,356
10	-,803	,609
11	,262	-,356
12	1,822	-,587
13	,828	1,692
14	,262	-,356
15	1,822	-,587
16	-,803	,609
17	,262	-,356
18	,262	-,356
19	-1,128	-1,377
20	-,803	,609
21	-1,128	-1,377
22	-,803	,609

Object Scores

	Dimension	
	1	2
23	1,822	-,587
24	-,803	,609
25	,262	-,356
26	-,803	,609
27	-1,128	-1,377
28	,262	-,356
29	-,803	,609
30	,262	-,356
31	-,803	,609
32	-1,128	-1,377
33	-1,128	-1,377
34	-1,128	-1,377
35	-,803	,609
36	-1,128	-1,377
37	-1,128	-1,377
38	-1,128	-1,377
39	-,803	,609
40	-,803	,609
41	-,803	,609
42	-1,128	-1,377
43	,262	-,356
44	-1,128	-1,377
45	-1,128	-1,377
46	,262	-,356
47	-,803	,609
48	-,803	,609
49	-,803	,609
50	,757	,378
51	,262	-,356
52	-1,128	-1,377
53	-,803	,609
54	,262	-,356
55	-,803	,609
56	-,237	2,656
57	-,803	,609
58	-,803	,609
59	-,803	,609
60	-,803	,609
61	,262	-,356
62	,262	-,356
63	-1,128	-1,377
64	,262	-,356
65	,262	-,356
66	,757	,378
67	,262	-,356

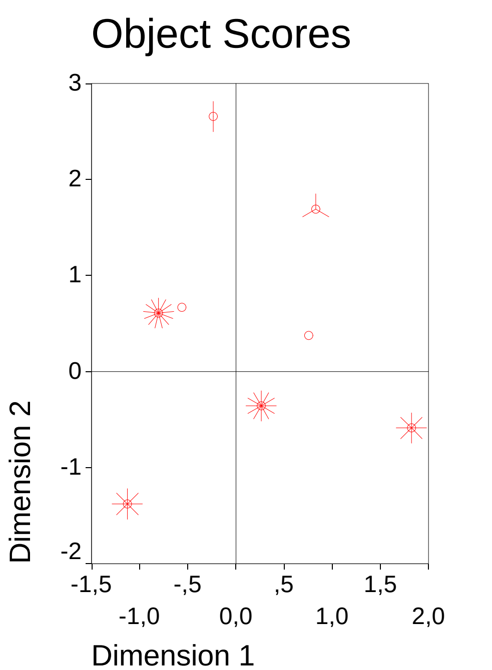
COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Object Scores

	Dimension	
	1	2
68	-,803	,609
69	1,822	-,587
70	-,237	2,656
71	-1,128	-1,377
72	-,803	,609
73	1,822	-,587
74	-,803	,609
75	1,822	-,587
76	-,803	,609
77	-,803	,609
78	-,803	,609
79	-1,128	-1,377
80	-,803	,609
81	-1,128	-1,377
82	-1,128	-1,377
83	,757	,378
84	,828	1,692
85	,828	1,692
86	,262	-,356
87	,262	-,356
88	,828	1,692
89	,262	-,356
90	1,822	-,587
91	-1,128	-1,377
92	1,822	-,587
93	-,803	,609
94	,262	-,356
95	-,803	,609
96	,828	1,692
97	1,822	-,587
98	-1,128	-1,377
99	,262	-,356
100	1,822	-,587
101	,262	-,356
102	-,561	,670
103	-,237	2,656
104	-,237	2,656
105	,262	-,356
106	1,822	-,587
107	,262	-,356
108	,262	-,356
109	-,561	,670
110	,828	1,692
111	,828	1,692
112	1,822	-,587

Object Scores

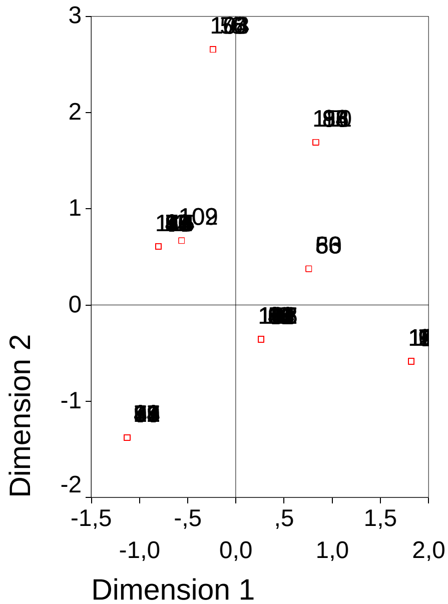
	Dimension	
	1	2
113	1,822	-,587
114	-,803	,609
115	-,803	,609
116	1,822	-,587
117	1,822	-,587
118	,262	-,356
119	1,822	-,587
120	1,822	-,587



Cases weighted by number of objects.

COMPOSITE POVERTY INDICATOR
VLSS-1 Communes Data
Example with 2 indicators, 6 categories

Object Scores Labeled by Code



Cases weighted by number of objects.