# A Multiple Correspondence Analysis to Organize Data Cubes

Riadh BEN MESSAOUD, Omar BOUSSAID and Sabine LOUDCHER RABASÉDA
*Laboratory ERIC – University of Lyon 2*
*5 avenue Pierre Mendès-France, 69676, Bron Cedex – France*

**Abstract.** On Line Analytical Processing (OLAP) is a technology basically created to provide users with tools in order to explore and navigate into data cubes. Unfortunately, in huge and sparse data, exploration becomes a tedious task and the simple user's intuition or experience does not lead to efficient results. In this paper, we propose to exploit the results of the Multiple Correspondence Analysis (MCA) in order to enhance data cube representations and make them more suitable for visualization and thus, easier to analyze. Our approach addresses the issues of organizing data in an interesting way and detects relevant facts. Our purpose is to help the interpretation of multidimensional data by efficient and simple visual effects. To validate our approach, we compute its efficiency by measuring the quality of resulting multidimensional data representations. In order to do so, we propose an homogeneity criterion to measure the visual relevance of data representations. This criterion is based on the concept of geometric neighborhood and similarity between cells. Experimental results on real data have shown the interest of using our approach on sparse data cubes.

**Keywords.** OLAP, Data cubes, data representation, MCA, test-values, arrangement of attributes, characteristic attributes, homogeneity criterion

## Introduction

On-Line Analytical Processing (OLAP) is a technology supported by most data warehousing systems [8,11]. It provides a platform for analyzing data according to multiple dimensions and multiple hierarchical levels. Data are presented in multidimensional views, commonly called data cubes [3]. A data cube can be considered as a space representation composed by a set of cells. A cell is associated with one or more measures and identified by coordinates represented by one attribute from each dimension. Each cell in a cube represents a precise fact. For example, if dimensions are *products*, *stores* and *months*, the measure of a particular cell can be the *sales* of one *product* in a particular *store* on a given *month*. OLAP provides the user with visual based tools to summarize, explore and navigate into data cubes in order to detect interesting and relevant information. However, exploring a data cube is not always an easy task to perform. Obviously, in large cubes containing sparse data, the whole analysis process becomes tedious and complex. In such a case, an intuitive exploration based on the user's experience does not quickly lead to efficient results. More generally, in the case of a data cube with more than three dimensions, a user is naturally faced to a hard task of navigation and explo-

ration in order to detect relevant information. Current OLAP provides query-driven and visual tools to browse data cubes, but does not deeply assist the user and help him/her to investigate interesting patterns.

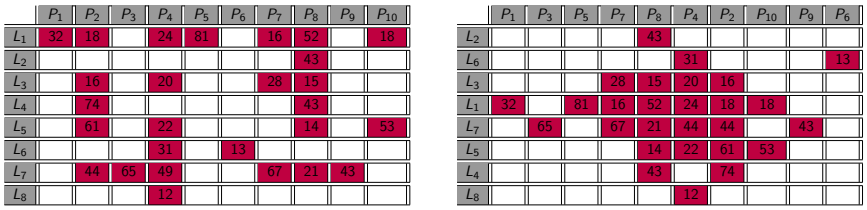


**Figure 1.** Example of two representations of a 2-dimensional data cube.

For example, consider the cube of Figure 1. On the one hand, representation 1(a) displays sales of products ($P_1, \ldots, P_{10}$) crossed by geographic locations of stores ($L_1, \ldots, L_8$). In this representation, full cells (gray cells) are displayed randomly according to the lexical ordering of the *attributes* – also called *members* – in each dimension. The way the cube is displayed does not provide an attractive representation that visually helps a user to easily interpret data.

On the other hand, Figure 1(b) contains the same information as Figure 1(a). However, it displays a data represen-tation which is visually easier to analyze. In fact, the cube of Figure 1(b) expresses important relationships by providing a visual representation that gathers full cells together and separates them from empty ones. In a natural way, such a representation is more comfortable to the user and allows to drive easy and efficient analysis.

Nevertheless, note that the representation (b) of Figure 1 can be interactively constructed by the user from representation (a) via some classic OLAP operators. This suppose that the user intuitively knows how to arrange the attributes of each dimension. Hence, we propose to provide the user with an automatic assistance to identify interesting facts and arrange them in a suitable visual representation [19,18]. As shown in Figure 1, we propose an approach that allows the user to get relevant facts expressing relationships and displays them in an appropriate way that enhances the exploration process independently of the cube's size. Thus, we suggest to carry out a Multiple Correspondence Analysis [9] (MCA) on a data cube as a preprocessing step. Basically, MCA is a powerful describing method even for huge volumes of data. It factors categorical variables and displays data in a factorial space constructed by orthogonal system of axes that provides relevant views of data. These elements motivate us to exploit the results of the MCA in order to better explore large data cubes by identifying and arranging its interesting facts. The first constructed factorial axis summarizes the maximum of information contained in the cube. We focus on relevant OLAP facts associated with characteristic attributes (variables) given by the factorial axes. These facts are interesting since they reflect relationships and concentrate a significant information. For a better visualization of these facts, we highlight them and arrange their attributes in the data space representation by using the *test-values* [14].

In order to evaluate the visual relevance of multidimensional data representations, we also propose in this paper a novel criterion to measure the homogeneity of cells distribution in the space representation of a data cube [17]. This criterion is based on geometric neighborhood of data cube cells, and also takes into account the similarity of cells' measures and provides a scalar quantification for the homogeneity of a given data cube representation. It also allows to evaluate the performance of our approach by comparing the quality of the initial data representation and the arranged one.

This paper is organized as follows. In section 1, we present related work to our approach. We provide in section 2 the problem formalization and present the general context of this work. The section 4 introduces the *test-values* and details the steps of our approach. We define in the next section our quality representation criterion. The section 6 presents a real world case study on a huge and sparse data cube. We propose experimental results in the section 7. Finally, we conclude and propose some future researches directions.

## 1. Related Work

Several works have already treated the issue of enhancing the space representation of data cubes. These works were undertaken following different motivations and adopted different ways to address the problem. We note that while many efforts are interested to computational aspects of data cubes (optimization of storage space, compression strategies, queries response time, etc.), a small number of studies have focused on OLAP aspects. Our present work fits into the second category. In our work, we focus on assisting OLAP users in order to improve and help the analysis process on large and sparse data cubes. We use a factorial approach to highlight relevant facts and provide an interesting visual data representations. Nevertheless, we dress an overview of main studies as well in the first as in the second category.

Some studies approximate computation of compressed data cube. In [23], Vitter *et al.* proposed to build compact data cubes by using approximation through wavelets. Quasi-Cube [1] compresses data representation by materializing only sufficient parts of a data cube. In [21] approximation is performed by estimating the density function of data. Other efforts address the issue of computing data cubes with index structure. For instance, Dwarf [22] uses indexes to reduce the storage space of a cube by identifying and factoring redundant tuples. Wang *et al.* propose to factorize data redundancies with BST [24] (*Base Single Tuple*). In [6], Feng *et al.* introduce PrefixCube, a data structure based on only one BST. The Quotient Cube [12] summarizes the semantic contents of a data cube and partitions it into similar cells. In [13], QC-Tree is directly constructed from the base table in order to maintain it under updates. Some other studies optimize storage spaces by partitioning the initial cube. Range CUBE [7] identifies correlations between attributes and compresses the data cube. Partitioned-Cube [20] partitions large relations into fragments. Operations on the cube are, therefore, performed in memory-sized fragments independently. In [15], high dimensional data are transformed into small local cubes and used to for online queries.

Finally, our approach shares already the same motivation of Choong *et al* [4]. The authors address the problem of high dimensionality of data cubes. They try to enhance analysis processes by preparing the dataset into appropriate representation. Thus, the

user can explore it in a more effective manner. The authors use an approach that combines association rules algorithm and a fuzzy subsets method. Their approach consists in identifying blocks of similar measures in the data cube. However, this approach does not take into account the problem of data sparsity. Furthermore, it does not provide a quality evaluation of the resulting multidimensional representations.

We emphasize that our approach does not deal with the issues of data cube compression, reduction of dimensionality or optimization of storage space. Through this study, we try to act on sparsity in huge multidimensional representations. Not to reduce it, but to reduce its negative effects on the interpretations and OLAP analysis of data [19,18]. Thus, we use the MCA to arrange differently the facts and highlight their relevant relationships in a data cube within a visual effect that gathers them as well as possible in the space representation.

## 2. Problem Formalization

Let $\mathcal{C}$ denote a data cube. Note that, our approach can be applied directly on $\mathcal{C}$ or on a data view (a sub-cube) extracted from $\mathcal{C}$. It is up to the user to select dimensions, fix one hierarchical level per dimension and select measures in order to create a particular data view (s)he wishes to visualize. Thus, to enhance the data representation of the constructed view, the user can apply on it our proposed approach. In order to lighten the formalization, in the followings of the paper, we assume that a user has selected a data cube $\mathcal{C}$, with $d$ dimensions ($D_1, \ldots, D_t, \ldots, D_d$), $m$ measures ($M_1, \ldots, M_q, \ldots, M_m$) and $n$ facts. We also assume that the user has fixed one hierarchical level with $p_t$ categorical attributes per dimension. Let $a_j^t$ the $j^{th}$ attribute of the dimension $D_t$ and $p = \sum_{t=1}^{d} p_t$ the total number of attributes in $\mathcal{C}$. For each dimension $D_t$, we note $\{a_1^t, \ldots, a_j^t, \ldots, a_{p_t}^t\}$ the set of its attributes.

In a first step, the aim of our approach is to organize the space representation of a given data cube $\mathcal{C}$ by arranging the attributes of its dimensions. For each dimension $D_t$, our approach establishes a new arrangement of its attributes $a_j^t$ in the data space (see subsection 4.2). This arrangement provides a data representation visually easier to interpret and displays multidimensional information in a more suitable way for analysis. In a second step, our approach detects from the resulted representation relevant facts expressing interesting relationships. To do that, we select from each dimension $D_t$ a subset $\Phi_t$ of significant attributes, also called characteristic attributes (see subsection 4.3). The crossing of these particular attributes allows to identify relevant cells in the cube.

Our approach is based on the MCA [9,14]. The MCA is a factorial method that displays categorical variables in a property space which maps their associations in two or more dimensions. From a table of $n$ observations and $p$ categorical variables, describing a $p$-dimensional cloud of individuals ($p < n$), the MCA provides orthogonal axes to describe the most variance of the whole data cloud. The fundamental idea is to reduce the dimensionality of the original data thanks to a reduced number of variables (factors) which are a combination of the original ones. The MCA is generally used as an exploratory approach to unearth empirical regularities of a dataset.

In our case, we assume the cube's facts as the individuals of the MCA, the cube's dimensions as its variables, and the attributes of a dimension as values of their corresponding variables. We apply the MCA on the $n$ facts of the cube $\mathcal{C}$ and use its results

| Id | $D_1$ | $D_2$ | $D_3$ | $M_1$ |
|---|---|---|---|---|
| 1 | $L1$ | $T2$ | $P1$ | 9 |
| 2 | $L2$ | $T2$ | $P3$ | 5 |
| 3 | $L2$ | $T1$ | $P2$ | 6 |
| 4 | $L1$ | $T1$ | $P3$ | 7 |

(a)

| | $Z$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | $Z_1$ | | $Z_2$ | | $Z_3$ | | |
| Id | $L1$ | $L2$ | $T1$ | $T2$ | $P1$ | $P2$ | $P3$ |
| 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 3 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 4 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |

(b)

**Figure 2.** Example of a conversion of a data cube to a complete disjunctive table.

to build *test-values* (see subsection 4.1) for the attributes $a_j^t$ of the dimensions $D_t$. We exploit these *test-values* to arrange attributes and detect characteristic ones in their corresponding dimensions.

## 3. Applying the MCA on a Data Cube

Like all statistical methods, MCA needs a tabular representation of data as input. Therefore, we can not apply it directly on multidimensional representations like data cubes. Therefore, we need to convert $\mathcal{C}$ to a *complete disjunctive table*. For each dimension $D_t$, we generate a binary matrix $Z_t$ with $n$ rows and $p_t$ columns. Rows represent facts, and columns represent dimension's attributes. The $i^{th}$ row of $Z_t$ contains $(p_t - 1)$ times the value 0 and one time the value 1 in the column that fits with the attribute taken by the fact $i$. The general term of $Z_t$ is:

$$z_{ij}^t = \begin{cases} 1 \textit{ if the fact } i \textit{ takes the attribute } a_j^t \\ 0 \textit{ else} \end{cases} \tag{1}$$

By merging the $d$ matrices $Z_t$, we obtain a complete disjunctive table $Z = [Z_1, Z_2, \ldots, Z_t, \ldots, Z_d]$ with $n$ rows and $p$ columns. It describes the $d$ positions of the $n$ facts of $\mathcal{C}$ through a binary coding. For instance, Figure 2 shows an simple example of a data cube (a), with 3 dimensions $D_1 : \{L_1, L_2\}$, $D_2 : \{T_1, T_2\}$, and $D_3 : \{P_1, P_2, P_3\}$. This cube is converted to a complete disjunctive table $Z$ in Figure 2(b). In the case of a large data cube, we naturally obtain a very huge matrix $Z$. Recall that MCA is a factorial method perfectly suited to huge input dataset with high numbers of rows and columns.

Once the complete disjunctive table $Z$ is built, MCA starts by constructing a matrix $B = Z'Z$ – called *Burt* table –, where $Z'$ is the transposed matrix of $Z$. *Burt* table $B$ is a $(p, p)$ symmetric matrix which contains all the category marginal on the main diagonal and all possible cross-tables of the $d$ dimensions of $\mathcal{C}$ in the off-diagonal. Let $X$ be a $(p, p)$ diagonal matrix which has the same diagonal elements of $B$ and zeros otherwise. We construct from $Z$ and $X$ a new matrix $S$ according to the formula:

$$S = \frac{1}{d} Z'ZX^{-1} = \frac{1}{d} BX^{-1} \tag{2}$$

By diagonalizing $S$, we obtain $(p - d)$ diagonal elements, called *eigenvalues* and denoted $\lambda_\alpha$. Each eigenvalue $\lambda_\alpha$ is associated to a directory vector $u_\alpha$ and corresponds to a factorial axis $F_\alpha$, where $Su_\alpha = \lambda_\alpha u_\alpha$.

An eigenvalue represents the amount of inertia (variance) that reflects the relative importance of its axis. The first axis always explains the most inertia and has the largest eigenvalue. Usually, in a factorial analysis process, researchers keep only the first, two or three axes of inertia. Other researchers give complex mathematical criterion [2,10,16,5] to determine the number of axes to keep. In [9], Benzecri suggests that this limit should be fixed by user's capacity to give a meaningful interpretation to the axes he keeps. It is not because an axis has a relatively small eigenvalue that we should discard it. It can often help to make a fine point about the data. It is up to the user to choose the number $k$ of axis to keep by checking eigenvalues and the general meaning of axes.

## 4. Organizing Data Cubes and Detecting Relevant Facts

Usually in a factorial analysis, relative contributions of variables are used to give sense to the axes. A relative contribution shows the percentage of inertia of a particular axis which is explained by an attribute. The largest relative contribution of a variable to an axis is, the more it gives sense to this axis. In our approach, we interpret a factorial axis by characteristic attributes detected through the use of the *test-values* proposed by Lebart *et al.* in [14]. In the followings, we present the theoretical principle of test-values applied to the context of our approach.

### *4.1. Test-Values*

Let $I(a_j^t)$ denotes the set of facts having $a_j^t$ as attribute in the dimension $D_t$. We also note $n_j^t = \text{Card}(I(a_j^t)) = \sum_{i=1}^{n} z_{ij}^t$ the number of elements in $I(a_j^t)$. It corresponds to the number of facts in $\mathcal{C}$ having $a_j^t$ as attribute (weight of $a_j^t$ in the cube). $\varphi_{\alpha j}^t = \frac{1}{n_j^t \sqrt{\lambda_\alpha}} \sum_{i \in I(a_j^t)} \psi_{\alpha i}$ is the coordinate of $a_j^t$ on the factorial axis $F_\alpha$, where $\psi_{\alpha i}$ is the coordinate of the facts $i$ on $F_\alpha$. Suppose that, under a null hypothesis $H_0$, the $n_j^t$ facts are selected randomly in the set of the $n$ facts, the mean of their coordinates in $F_\alpha$ can be represented by a random variable $Y_{\alpha j}^t = \frac{1}{n_j^t} \sum_{i \in I(a_j^t)} \psi_{\alpha i}$, where $\text{E}(Y_{\alpha j}^t) = 0$ and $\text{VAR}_{H_0}(Y_{\alpha j}^t) = \frac{n - n_j^t}{n-1} \frac{\lambda_\alpha}{n_j^t}$.

Remark that $\varphi_{\alpha j}^t = \frac{1}{\sqrt{\lambda_\alpha}} Y_{\alpha j}^t$. Thus, $\text{E}(\varphi_{\alpha j}^t) = 0,$ and $\text{VAR}_{H_0}(\varphi_{\alpha j}^t) = \frac{n - n_j^t}{n-1} \frac{1}{n_j^t}$. Therefore, the test-value of the attribute $a_j^t$ is:

$$V_{\alpha j}^t = \sqrt{n_j^t \frac{n-1}{n - n_j^t}} \varphi_{\alpha j}^t \tag{3}$$

$V_{\alpha j}^t$ measures the number of standard deviations between the attribute $a_j^t$ (the gravity center of the $n_j^t$ facts) and the center of the factorial axis $F_\alpha$. The position of an attribute is interesting for a given axis $F_\alpha$ if its cloud of facts is located in a narrow zone in the direction $\alpha$. This zone should also be as far as possible from the center of the axis. The test-value is a criterion that quickly provides an appreciation if an attribute has a *significant* position on a given factorial axis or not.

### *4.2. Arrangement of Attributes*

In a classic OLAP representation of data cubes, attributes are usually organized according to a lexical order such as alphabetic order for *geographic* dimensions or chronological order for *times* dimensions. In our approach, we propose to exploit the test-values of attributes in order to organize differently the data cube's facts. The new organization will display a relevant data representation easier to analyze and to interpret, especially in the case of large and sparse cubes. For each dimension, we sort its attributes according to the increasing order of their test-values. Actually, a test-value indicates the position of an attribute on a given axis. The relative geometric position of an attribute is more significant to factorial axes when these axes are important (have the greatest eigenvalues). For this, we propose to sort attributes according to the $k$ first axes selected by the user. We sort the $p_t$ test-values $V^t_{\alpha j}$ of the attributes $a^t_j$ on the axis $F_\alpha$. This will provide a new order of indices $j$. According to this order, we arrange attributes $a^t_j$ in the dimension $D_t$.

In general, we assume that all attributes of a dimension $D_t$ are geometrically ordered in the data cube space representation according to the order of indices $j_t$. i.e, the attribute $a^t_{j_t-1}$ precedes $a^t_{j_t}$ and $a^t_{j_t}$ precedes $a^t_{j_t+1}$ (see the example of Figure 3). Indices $j_t$ are ordered according to the arrangement of the attributes in the space representation of the dimension $D_t$.

### *4.3. Characteristic Attributes*

In general, an attribute is considered significant for an axis if the absolute value of its test-value is higher than $\tau = 2$. This roughly corresponds to an error threshold of $5\%$. We note that, the lower error threshold is, the greater $\tau$ is. In our case, for one attribute, the test of the hypothesis $H_0$ can induce a possible error. This error will inevitably be increased when we perform the test $p$ times for all the attributes of the cube. To minimize this accumulation of errors, we propose to fix for each test an error threshold of $1\%$ which correspond to $\tau = 3$. We also note that, when a given axis can be characterized by too much attributes according to their test-values, instead of taking them all, we can restrict the selection by considering only a percentage of the most characteristic ones. i.e, those having the highest absolute test-values. Finally to detect interesting facts in a data cube, for each dimension $D_t$, we select the following set of characteristic attributes.

$$\Phi_t = \left\{ \begin{array}{l} a^t_j, \text{ where } \forall\, j \in \{1, \ldots, p_t\}, \\ \exists\, \alpha \in \{1, \ldots, k\} \text{ such as } |V^t_{\alpha j}| \geq 3 \end{array} \right\} \tag{4}$$

## 5. Quality of a Data Representation

We provide a quality criterion of data cube representations [17]. It measures the homogeneity of the geometric distribution of cells in a data cube. One cell contains one or more measures of an OLAP fact. The attributes of a cell are coordinates of a fact according to dimensions in the data space representation. Let $A = (a^1_{j_1}, \ldots, a^t_{j_t}, \ldots, a^d_{j_d})$ be a cell in $\mathcal{C}$, where $t \in \{1, \ldots, d\}$ and $j_t \in \{1, \ldots, p_t\}$. $j_t$ is the index of the attribute that takes the cell $A$ according to $D_t$. We denote $|A|$ the value of a measure contained in $A$ which is equal to NULL if $A$ is empty. For example, in Figure 3, $|A| = 5.7$ whereas

$|Y| = \mathsf{NULL}$. We say that a cell $B = (b^1_{j_1}, \ldots, b^t_{j_t}, \ldots, b^d_{j_d})$ is neighbor of $A$, denoted $B \dashv A$, if $\forall t \in \{1, \ldots, d\}$, the coordinates of $B$ satisfy:

$$b^t_{j_t} = a^t_{j_t-1} \text{ or } b^t_{j_t} = a^t_{j_t} \text{ or } b^t_{j_t} = a^t_{j_t+1}$$

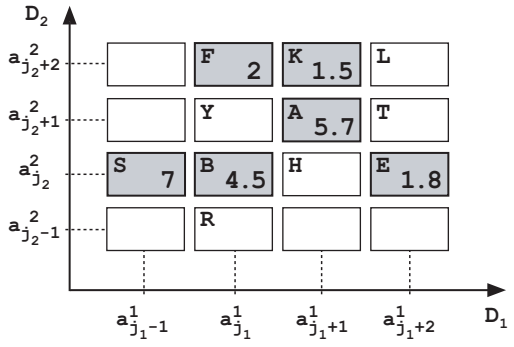


**Figure 3.** A 2-dimensional example of a data cube.

In Figure 3, cell $B$ is neighbor of $A$ ($B \dashv A$). $Y$ is also neighbor of $A$ ($Y \dashv A$). Whereas cells $S$ and $R$ are not neighbors of $A$. For a cell $A$ of a cube $\mathcal{C}$, we define the neighborhood of $A$, denoted $\mathcal{N}(A)$, by the set of all cells $B$ of $\mathcal{C}$ neighbors of $A$.

$$\mathcal{N}(A) = \{B \in \mathcal{C} \text{ where } B \dashv A\}$$

For example, in Figure 3, the neighborhood of $A$ corresponds to the set $\mathcal{N}(A) = \{F, K, L, T, E, H, B, Y\}$. To evaluate similarities between neighbor cells, we define a similarity function $\delta$.

**Definition** The similarity $\delta$ of two cells $A$ and $B$ from a cube $\mathcal{C}$ is defined as follows:

$$\delta : \mathcal{C} \times \mathcal{C} \longrightarrow \mathbb{R}$$
$$\delta(A, B) \longmapsto \begin{cases} 1 - (\frac{||A|-|B||}{\max(\mathcal{C})-\min(\mathcal{C})}) & \text{if } A \text{ and } B \text{ are full} \\ 0 & \text{else} \end{cases}$$

Where $||A| - |B||$ is the absolute difference of measures contained in cells $A$ and $B$, and $\max(\mathcal{C})$ (respectively, $\min(\mathcal{C})$) is the maximum (respectively, the minimum) measure value in cube $\mathcal{C}$.

In the cube of Figure 3, where grayed cells are full and white ones are empty, $\max(\mathcal{C}) = 7$, which matches with the cell $S$ and $\min(\mathcal{C}) = 1.5$, which matches with the cell $K$. For instance, $\delta(A, B) = 1 - (\frac{|5.7-4.5|}{7-1.5}) \simeq 0.78$ and $\delta(A, Y) = 0$.

Now, let consider a function $\Delta$ from $\mathcal{C}$ to $\mathbb{R}$ such as $\forall A \in \mathcal{C}$, $\Delta(A) = \sum_{B \in \mathcal{N}(A)} \delta(A, B)$. It corresponds to the sum of the similarities of $A$ with all its full neighbor cells. For instance, according to Figure 3, $\Delta(A)$ is computed as follows: $\Delta(A) = \delta(A, F) + \delta(A, K) + \delta(A, L) + \delta(A, T) + \delta(A, E) + \delta(A, H) + \delta(A, B) + \delta(A, Y) \simeq 1.64$.

We introduce the crude homogeneity criterion of a data cube $\mathcal{C}$ according to:

$$chc(\mathcal{C}) = \sum_{\substack{A \in \mathcal{C} \\ |A| \neq \text{NULL}}} \sum_{B \in \mathcal{N}(A)} \delta(A, B) = \sum_{\substack{A \in \mathcal{C} \\ |A| \neq \text{NULL}}} \Delta(A)$$

The crude homogeneity criterion computes the sum of similarities of every couple of full and neighbor cells in a data cube. For instance, in Figure 3, the crude homogeneity criterion is computed as $chc(\mathcal{C}) = \Delta(F) + \Delta(K) + \Delta(A) + \Delta(S) + \Delta(B) + \Delta(E) \simeq 6.67$. Note that, the crude homogeneity criterion of a data cube touches its maximum when all the cells of the cube are full and have equal measures. We denote $chc_{max}(\mathcal{C}) = \sum_{A \in \mathcal{C}} \sum_{B \in \mathcal{N}(A)} 1$.

**Definition** The homogeneity criterion of a data cube is defined as:

$$hc(\mathcal{C}) = \frac{chc(\mathcal{C})}{chc_{max}(\mathcal{C})} = \frac{\sum\limits_{\substack{A \in \mathcal{C} \\ |A| \neq \text{NULL}}} \Delta(A)}{\sum\limits_{A \in \mathcal{C}} \sum\limits_{B \in \mathcal{N}(A)} 1}$$

The homogeneity criterion evaluates the quality of a multidimensional data representation. This quality is rather better when full and similar cells are neighbors. Indeed, when similar cells are gathered in specific regions of the space representation of a data cube, this cube is easier to visualize and so, a user can directly focus his/her data interpretation on these regions.

For example, in Figure 3, $chc_{max}(\mathcal{C}) = 84$. So, the homogeneity criterion of this representation is: $hc(\mathcal{C}) = \frac{6.67}{84} \simeq 0.08$. Nevertheless, such a criterion can not make real sense for a single situation of a data representation. In all cases, we should rather compare it to other data representations of the same cube. In fact, recall that the aim of our method is to organize the facts of an initial data cube representation by arranging attributes in each dimensions according to the order of test-values. Let us denote the initial cube $\mathcal{C}_{ini}$ and the organized one $\mathcal{C}_{org}$. To measure the relevance of the organization provided by our method, we compute the gain $g = \frac{hc(\mathcal{C}_{org}) - hc(\mathcal{C}_{ini})}{hc(\mathcal{C}_{ini})}$ realized by the homogeneity criterion.

We also note that, for the same cube, its organized representation does not depend on the initial representation because the results of the MCA are insensitive to the order of input variables.

## 6. A Case Study

To test and validate our approach, we apply it on a 5-dimensional cube ($d = 5$) that we have constructed from the *Census-Income Database*[1] of the *UCI Knowledge Discovery in Databases Archive*[2]. This data set contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the *U.S. Census Bureau*. The data contains demographic and employment related variables. The constructed cube contains $199\,523$ facts and one fact represents a particular profile of a sub population measured by the *Wage per hour*. The dimensions of the cube are : *Education level* ($D_1$,

[1] http://kdd.ics.uci.edu/databases/census-income/census-income.html
[2] http://kdd.ics.uci.edu/

$p_1 = 17$), *Professional category* ($D_2$, $p_2 = 22$), *State of residence* ($D_3$, $p_3 = 51$), *Household situation* ($D_4$, $p_4 = 38$), and *Country of birth* ($D_5$, $p_5 = 42$).

We generate a complete disjunctive table $Z = [Z_1, Z_2, Z_3, Z_4, Z_5]$ according to a binary coding of the cube dimensions. $Z$ contains $199523$ rows and $p = \sum_{t=1}^{5} p_t = 170$ columns. By applying the MCA on $Z$ we obtain $p - d = 165$ factorial axes $F_\alpha$. Each axis is associated to an eigenvalue $\lambda_\alpha$. Suppose that, according to the histogram of eigenvalues, a user chooses the three first axes ($k = 3$). These axes explain $15.35\%$ of the total inertia of the facts cloud. This contribution does not seem very important at a first sight. But we should note that in a case of a uniform distribution of eigenvalues, we get normally a contribution of $\frac{1}{p-d} = 0.6\%$ per axis, i.e. the three first axes represent an inertia already $25$ times more important than a uniform distribution.

| | | Test-values | | |
|---|---|---|---|---|
| $j$ | Attributes | $V^1_{1j}$ | $V^1_{2j}$ | $V^1_{3j}$ |
| 9 | *Hospital services* | -99.90 | -99.90 | -99.90 |
| 14 | *Other professional services* | -99.90 | -99.90 | 99.90 |
| 17 | *Public administration* | -99.90 | -99.90 | 99.90 |
| 12 | *Medical except hospital* | -99.90 | 99.90 | -99.90 |
| 5 | *Education* | -99.90 | 99.90 | 99.90 |
| 7 | *Finance insurance* | -99.90 | 99.90 | 99.90 |
| 19 | *Social services* | -99.90 | 99.90 | 99.90 |
| 8 | *Forestry and fisheries* | -35.43 | -8.11 | 83.57 |
| 3 | *Communications* | -34.05 | -99.90 | 99.90 |
| 15 | *Personal services except private* | -21.92 | -5.50 | 10.28 |
| 13 | *Mining* | -6.59 | -99.64 | -5.25 |
| 16 | *Private household services* | 7.77 | 51.45 | 11.68 |
| 6 | *Entertainment* | 40.04 | 99.90 | 96.23 |
| 1 | *Agriculture* | 68.66 | 3.39 | -27.38 |
| 4 | *Construction* | 99.90 | -99.90 | -99.90 |
| 10 | *Manufact. durable goods* | 99.90 | -99.90 | -99.90 |
| 11 | *Manufact. nondurable goods* | 99.90 | -99.90 | -99.90 |
| 21 | *Utilities and sanitary services* | 99.90 | -99.90 | -99.90 |
| 22 | *Wholesale trade* | 99.90 | -99.90 | -24.37 |
| 20 | *Transportation* | 99.90 | -99.90 | 99.90 |
| 18 | *Retail trade* | 99.90 | 99.90 | -99.90 |
| 2 | *Business and repair* | 99.90 | 99.90 | 99.90 |

**Table 1.** Attribute's test-values of *Professional category* dimension.

The organized *Census-Income* data cube is obtained by sorting the attributes of its dimensions. For each dimension $D_t$ its attributes are sorted by the increasing values of $V^t_{1j}$, then by $V^t_{2j}$, and then by $V^t_{3j}$. Table 1 shows the new attributes' order of the *Professional category* dimension ($D_2$). Note that $j$ is the index of the original alphabetic order of the attributes. This order is replaced by a new one according to the sort of test-values. In the Figures 4(a) and 4(b), we can clearly see the visual effect of this arrangement of attributes. These figures display views of data by crossing the *Professional category* dimension on columns ($D_2$) and the *Country of birth* dimension on rows ($D_5$). The representation 4(a) displays the initial view according to the alphabetic order of attributes, whereas representation 4(b) displays the same view where attributes are rather sorted according to their test-values.

Remember that the aim of our current approach is not to compress or reduce the dimensions of a data cube. We do not also reduce sparsity of a data representation. Nevertheless, we act on this sparsity and reduce its negative effect on OLAP interpretation. Thus, we arrange differently original facts within a visual effect that gathers them as

well as possible in the space representation of the data cube. At a first sight, the visual representation 4(b) is more suitable to interpretation than 4(a). We clearly distinguish in Figure 4(b) four dense regions of full cells. In this regions, the homogeneity is higher than the rest of the space representation of the data cube.

This is confirmed by the measure of homogeneity criterion. Indeed, for a sparsity ratio of 63.42%, the homogeneity criterion for the organized cube of representation 4(b) is $hc(\mathcal{C}_{org}) = 0.17$; whereas it measures $hc(\mathcal{C}_{ini}) = 0.14$ for the initial cube of representation 4(a), i.e, we release a gain $g = 17.19\%$ of homogeneity when arranging the attributes of the cube according to test-values.

According to the test of the Equation (4), for each $t \in \{1, \ldots, 5\}$, we select from $D_t$ the set of characteristic attributes for the three selected factorial axes. These characteristic attributes give the best semantic interpretation of factorial axes and express strong relationships for their corresponding facts. To avoid great number of possible characteristic attributes per axis, we can consider, for each axis, only the first $50\%$ of attributes having the highest absolute test-values. For instance, in the *Professional category* dimension $D_2$, the set $\Phi_2$ of characteristic attributes correspond to grayed rows in table 1.

In the same way, we apply the test of the Equation (4) on the other dimensions of the cube. In the representation of Figure 4(b), we clearly see that the zones of facts corresponding to characteristic attributes of the dimensions $D_2$ and $D_5$ seem to be more interesting and denser than other regions of the data space representation. These zones contains relevant information and reflect interesting association between facts. For instance, we can easily note that industrial and physical jobs, like construction, agriculture and manufacturing are highly performed by *Native Latin Americans* from Ecuador, Peru, Nicaragua and Mexico for example. At the opposite, *Asians* people from India, Iran, Japan and China are rather concentrated in commerce and trade.

## 7. Experimental Results

We have realized some experiments on the *Census-Income* data cube presented in section 6. The aim of these experiments is to appreciate the efficiency of our approach by measuring the homogeneity gain realized by our MCA-based organization on data representations with different sparsity ratios. To vary sparsity we proceeded by a random sampling on the initial dataset of the $199\,523$ facts from the considered cube.

According to Figure 5, the homogeneity gain has an increasing general trend. Nevertheless, we should note that for low sparsity ratios, the curve is rather oscillating around the null value of the homogeneity gain. In fact, when sparsity is less then 60%, the gain does not have a constant variation. It sometimes drops to negative values. This means that our method does not bring a value added to the quality of data representation. For dense data cubes, the employment of our method is not always significant. This is naturally due to the construction of the homogeneity criterion which closely depends on the number of empty and full cells. It can also be due to the structure of the random data samples that can generate data representations already having good qualities and high homogeneity values.

Our MCA-based organization method is rather interesting for data representations with high sparsity. In Figure 5, we clearly see that curve is rapidly increasing to high positive values of gain when sparsity is greater than 60%. Actually, with high relative

| | Agriculture | Business and repair services | Communications | Construction | Education | Entertainment | Finance insurance | Forestry and fisheries | Hospital services | Manufact. durable goods | Manufact. nondurable goods | Medical except hospital | Mining | Other professional services | Personal services except private | Private household services | Public administration | Retail trade | Social services | Transportation | Utilities and sanitary services | Wholesale trade |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cambodia | | | | | | | | | | 125.0 | | | | | | | | | | 750.0 | | |
| Canada | | 35.0 | | 93.1 | 54.1 | | | 112.5 | 253.1 | 182.3 | | 373.4 | | 22.2 | | | 169.2 | 94.0 | | 267.6 | 11.1 | 350.0 |
| China | 622.0 | | | 40.7 | 50.1 | | 105.0 | 566.7 | 336.8 | 46.7 | 64.2 | 60.7 | | | | | 833.8 | 21.6 | | | 329.0 | 206.3 |
| Columbia | | | | | | | 79.0 | | | 46.6 | | 80.3 | | | | | 175.0 | | | | | |
| Cuba | | | 501.5 | | | | | | 31.8 | | | 19.0 | | | | | | 28.9 | | | | |
| Dominican-Republic | | | 375.0 | | 116.7 | | | | 146.0 | 92.7 | 38.1 | | | | | | | 35.1 | 75.0 | | | |
| Ecuador | 107.2 | 109.1 | 250.0 | 205.6 | 515.0 | | 206.7 | 68.8 | | | 128.1 | 265.6 | 100.0 | | | | 300.0 | 41.9 | 175.0 | | 333.3 | 212.5 |
| El-Salvador | 55.6 | 46.1 | | 36.1 | 81.0 | 950.8 | 344.0 | | | 184.7 | 19.4 | 120.0 | | | 79.5 | | | 20.7 | 400.0 | 36.9 | | 365.6 |
| England | | 77.9 | 222.7 | 418.1 | 90.2 | 50.0 | 46.9 | | 383.0 | 257.1 | 365.0 | | | 194.7 | | | 136.4 | 26.3 | 198.9 | | | |
| France | 450.0 | | | | | | | | | | 394.8 | | | | | | | 229.0 | | | | |
| Germany | | 115.0 | 200.0 | 157.1 | | | 97.9 | | 417.2 | 152.3 | 31.7 | 128.6 | | 22.2 | | 218.9 | 108.7 | 77.9 | | 253.1 | 428.2 | |
| Greece | | | | | 257.1 | | | | | 300.0 | 150.0 | | | 241.7 | | | 400.0 | 52.4 | 400.0 | | | 63.6 |
| Guatemala | | | | 121.8 | | | | | | 47.5 | 39.8 | | | | 136.2 | 25.8 | | | | | | |
| Haiti | | | | | | | | | 90.0 | | 80.6 | | | | 178.7 | | | | | | | ### |
| Holand-Netherlands | | | | | | | | | | | 21.4 | | | | | | | | | | | |
| Honduras | | | | | | | | | | | | | | | 151.7 | | 945.0 | | | | | |
| Hong Kong | 125.4 | | | 190.5 | | | 590.4 | 183.3 | | | 100.0 | | | 225.0 | ### | | | 150.0 | | 566.7 | 55.1 | 484.3 |
| Hungary | | | | | | | | | | | | | | | | 400.0 | | | | | | |
| India | | 94.2 | | | 101.2 | | 17.9 | 228.1 | 157.2 | 145.9 | | | | | | | | 100.0 | | | 167.1 | 81.3 |
| Iran | | 95.8 | | 225.0 | | | 66.7 | | | 160.7 | | | | 311.1 | 100.0 | | 316.7 | 90.0 | | 159.0 | | |
| Ireland | | | 500.0 | 100.0 | | | | | | | 533.3 | | | | | | | | | | | |
| Italy | | | | | 80.3 | | | | | | | | | | 27.8 | | | 32.9 | | | | |
| Jamaica | 250.0 | 158.8 | ### | 100.0 | 147.0 | | 79.2 | | 343.1 | | 571.4 | 106.0 | | 55.6 | 91.7 | 100.0 | 803.8 | | 604.7 | 533.3 | 19.4 | |
| Japan | | 107.1 | | | 63.5 | 425.0 | | 192.1 | 678.9 | 50.9 | 164.6 | | | | | | | 26.4 | 150.0 | 273.3 | 107.5 | |
| Laos | | | | | | 500.0 | | | | 116.6 | | | | 350.0 | | | | 71.4 | | | | |
| Mexico | 34.5 | 89.6 | 75.0 | 95.0 | 155.2 | 46.5 | 67.6 | | 122.2 | 61.9 | 59.8 | 89.7 | | 159.1 | 59.9 | 17.1 | | 52.9 | 40.3 | 140.3 | 121.7 | 82.1 |
| Nicaragua | 159.5 | | 83.3 | | 140.0 | | | 47.6 | | 340.0 | 76.5 | 65.6 | | 74.1 | | 160.0 | 178.3 | 81.0 | | | 85.7 | |
| Outlying-U S | | | | | | | | ### | | | | | | | | | | 93.8 | | | | 200.0 |
| Panama | | | | | | | | | | | | | | 452.5 | | | | | | | | |
| Peru | 225.0 | 699.6 | 69.7 | | 106.3 | | 47.0 | 450.0 | 166.7 | 215.4 | 76.2 | | | 134.5 | | | | 127.3 | 124.2 | 86.4 | 20.0 | 32.0 |
| Philippines | 200.0 | 122.7 | 265.0 | 270.0 | 317.8 | 62.5 | 165.0 | | 331.1 | 66.7 | 166.1 | 95.6 | | | | | 77.8 | 134.7 | | 197.3 | | 322.7 |
| Poland | | 252.9 | 175.6 | | | | 105.0 | | 325.0 | 185.5 | 92.6 | 175.2 | | | | | 180.0 | 196.2 | | 187.5 | | 212.5 |
| Portugal | | | | 166.7 | 155.6 | | 107.1 | | 141.1 | | | | | | | | | | | | | 236.7 |
| Puerto-Rico | | 87.8 | 250.0 | 54.2 | | 66.7 | 80.7 | 250.0 | 37.5 | 122.3 | 48.3 | 420.7 | | 40.0 | | | 110.1 | 23.9 | 43.5 | 163.8 | 142.9 | 33.6 |
| Scotland | | | | 87.5 | | 725.0 | 300.0 | | 785.0 | 95.2 | 14.0 | | 23.9 | | | | | 131.3 | 350.0 | 173.6 | 700.0 | 36.5 |
| South Korea | | | | | | | | | | | | | | | | | | | | | 870.0 | |
| Taiwan | | | | | | | | | | | | | | | | | | | | 46.2 | | |
| Thailand | | | | | | | | | | | | 150.0 | | | | | | 43.8 | | | | |
| Trinadad&Tobago | 66.3 | 243.8 | | 63.8 | | 920.0 | 333.3 | 89.3 | | 466.7 | | 175.0 | | | | | | 453.0 | 200.0 | | | 250.0 |
| United-States | 37.8 | 92.6 | 153.4 | 130.6 | 75.4 | 117.9 | 71.1 | 84.3 | 214.4 | 165.4 | 146.9 | 141.7 | | 76.0 | | | 142.1 | 99.3 | 96.0 | 157.0 | 199.9 | 84.4 |
| Vietnam | | | ### | | | | 75.0 | | 327.5 | 173.8 | | | | 250.0 | 32.1 | | | | | | | |
| Yugoslavia | | 42.1 | | | | | | | | | | | | | | | | | | | | |

**(a)**

| | Hospital services | Other professional services | Public administration | Medical except hospital | Education | Finance insurance | Social services | Forestry and fisheries | Communications | Personal services except private | Mining | Private household services | Entertainment | Agriculture | Construction | Manufact. durable goods | Manufact. nondurable goods | Utilities and sanitary services | Wholesale trade | Transportation | Retail trade | Business and repair services |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Philippines | 331.1 | | 77.8 | 95.6 | 317.8 | 165.0 | | | 265.0 | | | | 62.5 | 200.0 | 270.0 | 66.7 | 166.1 | | 322.7 | 197.3 | 134.7 | 122.7 |
| India | 157.2 | | | | 101.2 | 17.9 | | 228.1 | | | | | | | | 145.9 | | 167.1 | 81.3 | | 100.0 | 94.2 |
| Canada | 253.1 | 22.2 | 169.2 | 373.4 | 54.1 | | | 112.5 | | | | | | | 93.1 | 182.3 | | 11.1 | 350.0 | 267.6 | 94.0 | 35.0 |
| Jamaica | 343.1 | 55.6 | 803.8 | 106.0 | 147.0 | 79.2 | 604.7 | | ### | 91.7 | | 100.0 | | 250.0 | 100.0 | | 571.4 | 19.4 | | 533.3 | | 158.8 |
| Iran | | 311.1 | 316.7 | | | 66.7 | | | | 100.0 | | | | | 225.0 | 160.7 | | | | 159.0 | 90.0 | 95.8 |
| Japan | 678.9 | | | 63.5 | | | 150.0 | 192.1 | | | | | 425.0 | | | 50.9 | 164.6 | 107.5 | | 273.3 | 26.4 | 107.1 |
| China | 336.8 | | 833.8 | 60.7 | 50.1 | 105.0 | | 566.7 | | | | | | 622.0 | 40.7 | 46.7 | 64.2 | 329.0 | 206.3 | | 21.6 | |
| Hong Kong | | 225.0 | | | | 590.4 | | 183.3 | | ### | | | | 125.4 | 190.5 | | 100.0 | 55.1 | 484.3 | 566.7 | 150.0 | |
| Greece | | 241.7 | 400.0 | | 257.1 | | 400.0 | | | | | | | | | 300.0 | 150.0 | | 63.6 | | 52.4 | |
| Germany | 417.2 | 22.2 | 108.7 | 128.6 | | 97.9 | | | 200.0 | | | 218.9 | | | 157.1 | 152.3 | 31.7 | 428.2 | | 253.1 | 77.9 | 115.0 |
| Scotland | 785.0 | | | | | 300.0 | 350.0 | | | 23.9 | | | 725.0 | | 87.5 | 95.2 | 14.0 | 700.0 | 36.5 | 173.6 | 131.3 | |
| Poland | 325.0 | | 180.0 | 175.2 | | 105.0 | | | 175.6 | | | | | | | 185.5 | 92.6 | | 212.5 | 187.5 | 196.2 | 252.9 |
| England | 383.0 | 194.7 | 136.4 | | 90.2 | 46.9 | 198.9 | | 222.7 | | | | 50.0 | | 418.1 | 257.1 | 365.0 | | | | 26.3 | 77.9 |
| Haiti | 90.0 | | | | | | | | | 178.7 | | | | | | | 80.6 | | ### | | | |
| Taiwan | | | | | | | | | | | | | | | | | | | | 46.2 | | |
| Panama | | 452.5 | | | | | | | | | | | | | | | | | | | | |
| Outlying-U S | | | | | | | | ### | | | | | | | | | | | 200.0 | | 93.8 | |
| Thailand | | | | 150.0 | | | | | | | | | | | | | | | | | 43.8 | |
| Italy | | | | | 80.3 | | | | | 27.8 | | | | | | | | | | | 32.9 | |
| Hungary | | | | | | | | | | | | 400.0 | | | | | | | | | | |
| Vietnam | 327.5 | 250.0 | | | | 75.0 | | | ### | 32.1 | | | | | | 173.8 | | | | | | |
| Holand-Netherlands | | | | | | | | | | | | | | | | | 21.4 | | | | | |
| Portugal | 141.1 | | | | 155.6 | 107.1 | | | | | | | | | 166.7 | | | | 236.7 | | | |
| Yugoslavia | | | | | | | | | | | | | | | | | | | | | | 42.1 |
| South Korea | | | | | | | | | | | | | | | | | | 870.0 | | | | |
| Honduras | | | 945.0 | | | | | | | 151.7 | | | | | | | | | | | | |
| Cuba | 31.8 | | | 19.0 | | | | | 501.5 | | | | | | | | | | | | 28.9 | |
| France | | | | | | | | | | | | | | 450.0 | | | 394.8 | | | | 229.0 | |
| Cambodia | | | | | | | | | | | | | | | | 125.0 | | | | 750.0 | | |
| Dominican-Republic | 146.0 | | | | 116.7 | | 75.0 | | 375.0 | | | | | | | 92.7 | 38.1 | | | | 35.1 | |
| Laos | | 350.0 | | | | | | | | | | | 500.0 | | | 116.6 | | | | | 71.4 | |
| Guatemala | | | | | | | | | | 136.2 | | 25.8 | | | 121.8 | 47.5 | 39.8 | | | | | |
| Columbia | | | 175.0 | 80.3 | | 79.0 | | | | | | | | | | 46.6 | | | | | | |
| Ireland | | | | | | | | | 500.0 | | | | | | 100.0 | | 533.3 | | | | | |
| Trinadad&Tobago | | | | 175.0 | | 333.3 | 200.0 | 89.3 | | | | | 920.0 | 66.3 | 63.8 | 466.7 | | 250.0 | | 453.0 | 243.8 | |
| Puerto-Rico | 37.5 | 40.0 | 110.1 | 420.7 | | 80.7 | 43.5 | 250.0 | 250.0 | | | | 66.7 | | 54.2 | 122.3 | 48.3 | 142.9 | 33.6 | 163.8 | 23.9 | 87.8 |
| Ecuador | | | 300.0 | 265.6 | 515.0 | 206.7 | 175.0 | 68.8 | 250.0 | | 100.0 | | | 107.2 | 205.6 | | 128.1 | 333.3 | 212.5 | | 41.9 | 109.1 |
| Peru | 166.7 | 134.5 | | | 106.3 | 47.0 | 124.2 | 450.0 | 69.7 | | | | | 225.0 | | 215.4 | 76.2 | 20.0 | 32.0 | 86.4 | 127.3 | 699.6 |
| Nicaragua | | 74.1 | 178.3 | 65.6 | 140.0 | | | 47.6 | 83.3 | | | 160.0 | | 159.5 | | 340.0 | 76.5 | 85.7 | | | 81.0 | |
| Mexico | 122.2 | 159.1 | | 89.7 | 155.2 | 67.6 | 40.3 | | 75.0 | 59.9 | | 17.1 | 46.5 | 34.5 | 95.0 | 61.9 | 59.8 | 121.7 | 82.1 | 140.3 | 52.9 | 89.6 |
| El-Salvador | | | | 120.0 | 81.0 | 344.0 | 400.0 | | | 79.5 | | | 950.8 | 55.6 | 36.1 | 184.7 | 19.4 | | 365.6 | 36.9 | 20.7 | 46.1 |
| United-States | 214.4 | 76.0 | 142.1 | 141.7 | 75.4 | 71.1 | 96.0 | 84.3 | 153.4 | | | | 117.9 | 37.8 | 130.6 | 165.4 | 146.9 | 199.9 | 84.4 | 157.0 | 99.3 | 92.6 |

**(b)**

**Figure 4.** (a) Initial and (b) organized data representations of the *Census-Income*'s data cube.

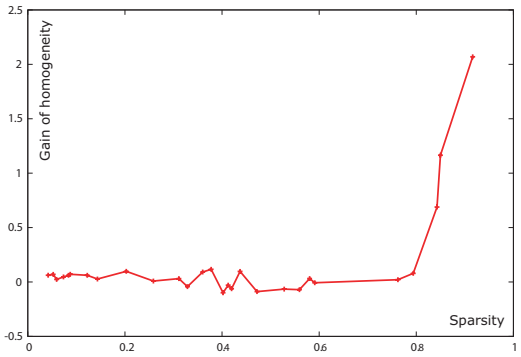


**Figure 5.** Evolution of the homogeneity gain according to sparsity.

number of empty cells in a data cube, we have a large manoeuvre margin for concentrating similar full cells and gathering them in the space representation. This shows the vocation of using our approach in order to enhance the visual quality representation, and thus the analysis of huge and sparse data cubes.

## 8. Conclusion and Future Work

In this paper, we introduced a MCA-based approach to enhance the space representation of large and sparse data cubes. This approach aims to provide an assistance to the OLAP user and helps him/her to easily explore huge volumes of data. For a given data cube, we compute the test-values of its attributes. According to these test-values, we arrange attributes of each dimension and so display in an appropriate way the space representation of facts. This representation provides better property for data visualization since it gather full cells expression interesting relationships of data. We also identify relevant regions of facts in this data representation by detecting characteristic attributes of factorial axes. This solve the problem of high dimensionality and sparsity of data and allows the user to directly focus his exploration and data interpretation on these regions. We have also proposed an homogeneity criterion to measure the quality of data representations. This criterion is based on the notion of geometric neighborhood of cells and their measures' similarities. Through experiments we led on real world data, our criterion proved the efficiency of our approach for huge and sparse data cubes.

Currently, we are studying some possible extensions for this work. We consider the problem of optimizing complexity of our approach. We also try to involve our approach in order to take into account the issue of data updates. Finally, we project to implement this approach under a Web environment that offers an interesting on-line aspect and a good user interaction context.

## References

[1] D. Barbará and M. Sullivan. Quasi-Cubes: Exploiting Approximations in Multidimensional Databases. *SIGMOD Record*, 26(3):12–17, 1997.

[2] R. Cattell. The Scree Test for the Number of Fators. *Multivariate Behavioral Research*, 1:245–276, 1966.

[3] S. Chaudhuri and U. Dayal. An Overview of Data Warehousing and OLAP Technology. *SIGMOD Record*, 26(1):65–74, 1997.

[4] Y. W. Choong, D. Laurent, and P. Marcel. Computing Appropriate Representations for Multidimensional Data. *Data & knowledge Engineering Journal*, 45(2):181–203, 2003.

[5] B. Escofier and B. Leroux. Etude de trois problèmes de stabilité en analyse factorielle. *Publication de l'Institut Statistique de l'Université de Paris*, 11:1–48, 1972.

[6] J. Feng, Q. Fang, and H. Ding. PrefixCube: Prefix-sharing Condensed Data Cube. In $7^{th}$ *ACM International Workshop on Data warehousing and OLAP (DOLAP 2004)*, pages 38–47, Washington D.C., U.S.A., November 2004.

[7] Y. Feng, D. Agrawal, A. E. Abbadi, and A. Metwally. Range CUBE: Efficient Cube Computation by Exploiting Data Correlation. In $20^{th}$ *International Conference on Data Engineering (ICDE 2004)*, pages 658–670, Boston, Massachusetts, U.S.A., March–April 2004.

[8] W. H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, 1996.

[9] J.P. Benzecri. *Correspondence Analysis Handbook*. Marcel Dekker, hardcover edition, January 1992.

[10] H. Kaiser. A note on Guttman's lower bound for the number of common factors. *Brit. J. Statist. Psychol.*, 14:1–2, 1961.

[11] R. Kimball. *The Data Warehouse toolkit*. John Wiley & Sons, 1996.

[12] L. V. Lakashmanan, J. Pei, and J. Han. Quotient Cube: How to Summarize the Semantics of a Data Cube. In $28^{th}$ *International Conference of Very Large Data Bases (VLDB 2002)*, Hong Kong, China, August 2002.

[13] L. V. Lakshmanan, J. Pei, and Y. Zhao. QC-Trees: An Efficient Summary Structure for Semantic OLAP. In A. Press, editor, *ACM SIGMOD International Conference on Management of Data (SIGMOD 2003)*, pages 64–75, San Diego, California, U.S.A., 2003.

[14] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunold, Paris, $3^e$ édition edition, 2000.

[15] X. Li, J. Han, and H. Gonzalez. High-Dimensional OLAP: A Minimal Cubing Approach. In $30^{th}$ *International Conference on Very Large Data Bases (VLDB 2004)*, pages 528–539, Toronto, Canada, August 2004.

[16] E. Malinvaud. Data Analysis in Applied Socio-Economic Statistics with Special Consideration of Correspondence Analysis. In *Marketing Science Conference*, Jouy en Josas, France, 1987.

[17] R. B. Messaoud, O. Boussaid, and S. L. Rabaséda. Evaluation of a MCA-Based Approach to Organize Data Cubes. In *Proceedings of the* $14^{th}$ *ACM International Conference on Information and Knowledge Management (CIKM'2005)*, pages 341–342, Bremen, Germany, October – November 2005. ACM Press.

[18] R. B. Messaoud, O. Boussaid, and S. L. Rabaséda. Efficient Multidimensional Data Representation Based on Multiple Correspondence Analysis. In *Proceedings of the* $12^{th}$ *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2006)*, pages 662–667, Philadelphia, PA, USA, August 2006. ACM Press.

[19] R. B. Messaoud, O. Boussaid, and S. L. Rabaséda. Using a Factorial Approach for Efficient Representation of Relevant OLAP Facts. In *Proceedings of the* $7^{th}$ *International Baltic Conference on Databases and Information Systems (DB&IS'2006)*, pages 98–105, Vilnius, Lithuania, July 2006. IEEE Communications Society.

[20] K. A. Ross and D. Srivastava. Fast Computation of Sparse Datacubes. In $23^{rd}$ *International Conference on Very Large Data Bases (VLDB 1997)*, pages 116–125, Athens, Greece, August 1997.

[21] J. Shanmugasundaram, U. M. Fayyad, and P. S. Bradley. Compressed Data Cubes for OLAP Aggregate Query Approximation on Continuous Dimensions. In $5^{th}$ *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 1999)*, pages 223–232, San Diego, California, U.S.A., August 1999.

[22] Y. Sismanis, A. Deligiannakis, N. Roussopoulos, and Y. Kotidis. Dwarf: Shrinking the PetaCube. In *ACM SIGMOD International Conference on Management of Data (SIGMOD 2002)*, pages 464–475, Madison, Wisconsin, U.S.A., 2002.

[23] J. S. Vitter and M. Wang. Approximate Computation of Multidimensional Aggregates of Sparse Data Using Wavelets. In *ACM SIGMOD International Conference on Management of Data (SIGMOD 1999)*, pages 193–204, Philadelphia, Pennsylvania, U.S.A., June 1999. ACM Press.

[24] W. Wang, H. Lu, J. Feng, and J. X. Yu. Condensed Cube: An Effective Approach to Reducing Data Cube Size. In $18^{th}$ *IEEE International Conference on Data Engineering (ICDE 2002)*, pages 155–165, San Jose, California, U.S.A., February–March 2002.