# Using a Factorial Approach for Efficient Representation of Relevant OLAP Facts

Riadh Ben Messaoud[♭], Omar Boussaid[‡] and Sabine Loudcher Rabaséda[‡]
[♭] rbenmessaoud@eric.univ-lyon2.fr, [‡]{ omar.boussaid | sabine.loudcher }@univ-lyon2.fr
Laboratory ERIC - University of Lyon 2
5 avenue Pierre Mendès-France
69676, Bron Cedex – France
http://eric.univ-lyon2.fr

***Abstract*— On Line Analytical Processing (OLAP) is a technology basically created to explore data cubes and detect relevant information. Unfortunately, in huge and sparse data volumes, exploration becomes a tedious task. In such a case, simple user's intuition or experience does not always lead to efficient results. In this paper, we propose to exploit the Multiple Correspondence Analysis (MCA) in order to assist exploration of cubes by enhancing their space representations. MCA is a factorial method that maps associations of huge number of categorical variables and displays them within an appropriate space representation. Our approach uses *test-values* provided by MCA in order to detect and arrange OLAP facts in a large and sparse data cube within an interesting visual effect which gathers full cells in relevant regions and separates them from empty cells. Thus, it is possible to focus analysis on interesting facts by browsing directly the provided regions in the data cube.**

## I. Introduction

On-Line Analytical Processing (OLAP) is a technology supported by most data warehousing systems [1], [2]. 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 users with visual 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 user's experience does not quickly lead to efficient results. Current OLAP provides query-driven and visual tools to browse data cubes, but does not deeply assist users and help them to investigate interesting patterns.

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 lexical order of members – also called *attributes* – of each dimension. The way the cube is displayed does not provide an attractive representation that visually helps a user to interpret and analyze data. On the other hand, Figure 1(b) contains the same information as Figure 1(a). However, it displays a data representation visually easier to analyze. Figure 1(b) expresses important relationships by providing a visual representation that gathers full cells together and separates them from empty cells. In a natural way, such a representation is more comfortable and allows to drive easy and efficient analysis. Nevertheless, note that representation (b) of Figure 1 may be interactively constructed from representation (a) via classic OLAP operators. However, this suppose that we intuitively know how to arrange attributes of each dimension. Hence, we propose an automatic assistance to identify interesting facts and arrange them in a suitable visual representation.

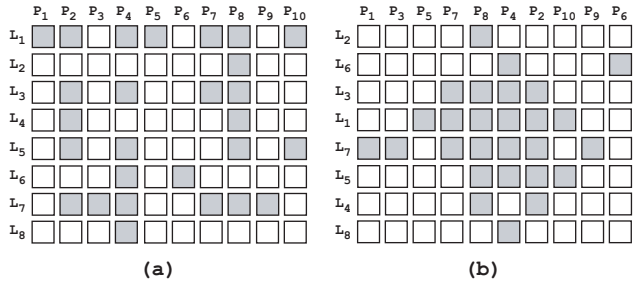


Fig. 1. Two representations of a data cube.

As shown in Figure 1, we propose an approach that enables relevant facts expressing interesting relationships and displays them in an appropriate way which enhances the exploration process independently of the cube's size. We suggest to carry out a Multiple Correspondence Analysis [4] (MCA) on data cubes 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 MCA in order to enhance exploration of large data cubes by identifying and arranging their interesting facts. 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. We highlight these facts and arrange their attributes in the data space representation by using *test-values* [5].

This paper is organized as follows. In section II, we provide a formalization and a general framework to define notations and the goal of our approach. We detail, in section III, the steps we follow to apply MCA on a multidimensional structure (data cube). Section IV introduces test-values and details how we use them to detect relevant facts in a data cube. We propose a case study of our approach on a real world data cube in section V. In section VI, we present some related work. Finally, we conclude and propose some future works.

## II. Problem Formalization

Let $\mathcal{C}$ denotes a data cube. We emphasize 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 users to select dimensions, fix one hierarchical level per dimension and select measures in order to create a particular data view to organize. Our approach can be applied on the constructed sub-cube. In the followings, in order to facilitate our formalization, 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 one hierarchical level, with $p_t$ categorical attributes, is fixed per dimension. $p = \sum_{t=1}^{d} p_t$ is the the total number of attributes in $\mathcal{C}$. We suppose that $a_j^t$ is the $j^{th}$ attribute of dimension $D_t$, and we assume that for each dimension $D_t$, $\{a_1^t, \ldots, a_j^t, \ldots, a_{p_t}^t\}$ is the set of its attributes.

The objective of our proposal is to detect from the initial cube $\mathcal{C}$ relevant facts expressing interesting relationships. In order to do so, we propose to select from each dimension $D_t$ subsets of characteristic attributes $\Phi_t$. These attributes give a specific interpretation for factorial axes of a MCA [4], [5] built over the whole set of cube's facts. The crossing of these attributes enables the identification of relevant cells. Indeed, 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 (rows) on $p$ categorical variables (columns), describing a $p$-dimensional cloud of individuals ($p < n$), 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 – commonly called *factors* – which are a combination of the original variables. MCA is generally used as an exploratory approach to unearth empirical regularities of a dataset. In our case, OLAP facts represent individuals of MCA, cube's dimensions represent variables of MCA, and the attributes of a dimension represent values of variables.

| 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)

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

## III. Applying 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:

$$z_{ij}^t = \begin{cases} 1 & \text{if the fact } i \text{ takes the attribute } a_j^t \\ 0 & \text{else} \end{cases} \quad (1)$$

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 a 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 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. According to the $Z$ table of Figure 2(b), the matrix $B$ and $X$ are written as follow:

$$B = Z'Z = \begin{pmatrix} 2 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 2 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 2 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 2 \end{pmatrix}$$

$$X = \begin{pmatrix} 2 & 0 & & \dots & & & 0 \\ 0 & 2 & . & & & & . \\ . & . & 2 & . & & & . \\ . & & . & 2 & . & & . \\ . & & & . & 1 & . & . \\ . & & & & . & 1 & 0 \\ 0 & & & \dots & & 0 & 2 \end{pmatrix}$$

We construct from $Z$ and $X$ a new matrix $S$ according to the formula:

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

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 $Su_\alpha = \lambda_\alpha u_\alpha$. The algorithm `CubeToMCA` of Figure 3 illustrates the previous process. This algorithm creates a complete disjunctive table from an input cube $\mathcal{C}$, applies MCA on $\mathcal{C}$, and returns eigenvalues in output.

**Algorithm** CubeToMCA($\mathcal{C}$)
**Input:**
  $\mathcal{C}$: data cube
**Begin**
  **for** $(t = 1; t \leq p; t++)$ **do**
    $Z_t \leftarrow 0$;
    **for each** attribute $a_j^t$ **in** $D_t$ **do**
      **for each** fact $i$ **in** $\mathcal{C}$ **do**
        **if** (fact $i$ takes $a_j^t$) **then**
          $z_{ij}^t \leftarrow 1$;
          Break for;
        **end if**
      **end for**
    **end for**
    $Z \leftarrow$ **merge**($Z$,$Z_t$);
  **end for**
  $B \leftarrow ZZ'$;
  **for** $(i = 1; i \leq p; i++)$ **do**
    **for** $(j = 1; j \leq p; j++)$ **do**
      **if** $(i \neq j)$ **then**
        $x_{ij} \leftarrow 0$;
        **else** $x_{ij} \leftarrow b_{ij}$;
      **end if**
    **end for**
  **end for**
  $S \leftarrow \frac{1}{d} Z'ZX^{-1}$;
  $S \leftarrow$ **diagonalize**($S$);
  **for** $(\alpha = 1; \alpha \leq p-d; \alpha++)$ **do**
    $\lambda_\alpha \leftarrow s_{\alpha\alpha}$;
  **end for**
**End**

Fig. 3. Algorithm CubeToMCA.

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 [6], [7], [8], [9] to determine the number of axes to keep. In [4], 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.

## IV. Using Test-Values to Characterize Axis

Constructed factorial axes can be characterized by attributes coming from initial OLAP dimensions. In a factorial analysis, relative contributions of variables are usually used to give sense to factorial axes. A relative contribution shows the percent of inertia of a particular axis explained by an attribute. The larger relative contribution of a variable to an axis is, the more it gives sense of this axis. In our approach, we choose to characterize the $k$ selected factorial axes by using *test-values* proposed by Lebart *et al.* in [5].

Let $I(a_j^t)$ be the set of facts having $a_j^t$ as attribute in the dimension $D_t$. $n_j^t$ is the number of elements in $I(a_j^t)$. $n_j^t$ corresponds to the number of facts in $\mathcal{C}$ having $a_j^t$ as attribute (weight of $a_j^t$ in the cube).

$$n_j^t = \text{Card}(I(a_j^t)) = \sum_{i=1}^{n} z_{ij}^t \quad (3)$$

We consider $\psi_{\alpha i}$ the coordinate of fact $i$ according to the axis $F_\alpha$. Therefore, the coordinate of the attribute $a_j^t$ according to $F_\alpha$ is:

$$\varphi_{\alpha j}^t = \frac{1}{n_j^t \sqrt{\lambda_\alpha}} \sum_{i \in I(a_j^t)} \psi_{\alpha i} \quad (4)$$

Under a null hypothesis $H_0$, if the $n_j^t$ facts are selected randomly in the set of $n$ facts, the mean of their coordinates in $F_\alpha$ is represented by a random variable $Y_{\alpha j}^t$:

$$Y_{\alpha j}^t = \frac{1}{n_j^t} \sum_{i \in I(a_j^t)} \psi_{\alpha i} \quad (5)$$

where its mean is $\text{E}(Y_{\alpha j}^t) = 0$, and its variance is:

$$\text{VAR}_{H_0}(Y_{\alpha j}^t) = \frac{n - n_j^t}{n-1} \frac{\lambda_\alpha}{n_j^t} \quad (6)$$

Knowing that $\varphi_{\alpha j}^t = \frac{1}{\sqrt{\lambda_\alpha}} Y_{\alpha j}^t$, the mean of $\varphi_{\alpha j}^t$ is therefore equal to zero ($\text{E}(\varphi_{\alpha j}^t) = 0$) and its variance is:

$$\text{VAR}_{H_0}(\varphi_{\alpha j}^t) = \frac{n - n_j^t}{n-1} \frac{1}{n_j^t} \quad (7)$$

Therefore, the test-value of $a_j^t$ is written as follows:

$$V_{\alpha j}^t = \sqrt{n_j^t \frac{n-1}{n - n_j^t}} \varphi_{\alpha j}^t \quad (8)$$

$V_{\alpha j}^t$ measures the number of standard deviations between the attribute $a_j^t$ (the gravity center of $n_j^t$ facts) and the center of factorial axis $F_\alpha$. The position of an attribute is interesting for a given axis $F_\alpha$ since its facts' cloud is located in a narrow zone in 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 wether an attribute has a *significant* position on a given factorial axis. These elements motivate us to use test-values to characterize factorial axes provided by MCA.

In general, an attribute is considered significant for an axis if the absolute value of its test-value is higher than $\tau = 2$. This corresponds roughly to an error threshold of $5\%$. We note that a low error threshold corresponds to a high value of $\tau$. In our case, for one attribute, the test confidence of hypothesis $H_0$ can be affected by a possible error. This error will increase by performing $p$ tests for all the cube attributes. In order to minimize this accumulation of errors, we propose to fix an error threshold of $1\%$ which correspond to $\tau = 3$. We also emphasize that a given axis can be characterized by too much attributes according to their test-values. Therefore, instead of taking all these attributes, we can consider only a subset of most characteristic ones. We select those having the highest absolute test-values. Finally, in order to detect interesting facts in a data cube, for each dimension $D_t$, we select in the following set the most characteristic attributes.

$$\Phi_t \subseteq \left\{ \begin{array}{l} a_j^t, \text{ where } \forall\ j \in \{1, \ldots, p_t\}, \\ \exists\ \alpha \in \{1, \ldots, k\} \text{ such as } |V_{\alpha j}^t| \geq 3 \end{array} \right\} \quad (9)$$

## V. A Case Study

To test and validate our approach, we apply it on a 5-dimensional cube ($d = 5$) constructed from the *Census-Income Database*[1] of the *UCI Knowledge Discovery in Databases Archive*[2]. Basically, this database contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the *U.S. Census Bureau* about demographic and employment related variables. The constructed cube contains 199 523 facts where each fact represents a particular profile of a sub population measured by *Wage per hour*. Table I details the cube's dimensions.

| Dimension | $p_t$ |
|---|---|
| $D_1$ : Education level | $p_1 = 17$ |
| $D_2$ : Professional category | $p_2 = 22$ |
| $D_3$ : State of residence | $p_3 = 51$ |
| $D_4$ : Household situation | $p_4 = 38$ |
| $D_5$ : Country of birth | $p_5 = 42$ |

TABLE I
DIMENSIONS OF THE *Census-Income*'S CUBE.

According to the binary coding of Equation (1), the *Census-Income* data cube is converted to a complete disjunctive table $Z = [Z_1, Z_2, Z_3, Z_4, Z_5]$. $Z$ contains 199 523 rows and $p = \sum_{t=1}^{5} p_t = 170$ columns. MCA provides $p - d = 165$ factorial axes $F_\alpha$ from $Z$. Figure 4 displays the first factorial plane (first and second axis) provided by MCA. 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 also note that in a case of a uniform distribution of eigenvalues, we normally get 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.

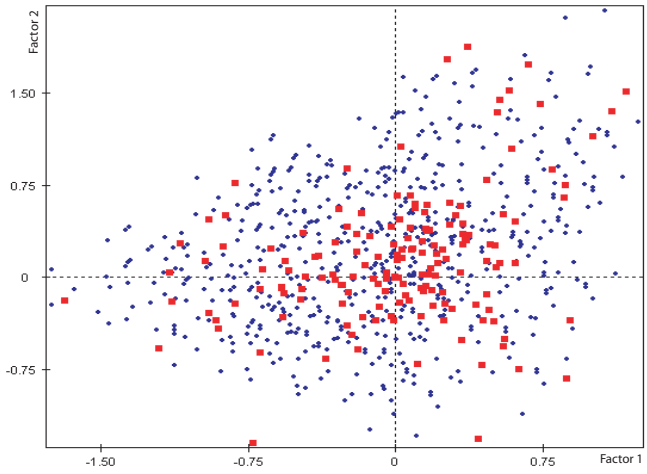


Fig. 4. First factorial plane constructed by MCA.

The organized *Census-Income* data cube is obtained by sorting the attributes of its dimensions. For each dimension $D_t$ its attributes are sorted according to the increasing values of $V_{1j}^t$, then according to $V_{2j}^t$, and then according to $V_{3j}^t$. Table II shows the new attributes' order of the *Professional category* dimension ($D_2$). Note that $j$ is the index of original alphabetic order of attributes. This order has changed according to a sort of test-values. In Figures 5 and 6, 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 ($D_5$). Representation 5 displays the initial view according to the alphabetic order of attributes, whereas representation 6 displays the same view where attributes are rather sorted according to their test-values.

Remember that the objective of our 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 6 is more suitable to interpretation than representation 5. We clearly distinguish in Figure 6 four dense regions of full cells.

According to Equation (9), 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 character-

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

| | | 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 II
ATTRIBUTE'S TEST-VALUES OF *Professional category* DIMENSION.

istic attributes corresponds to those grayed in Table II:

$$\Phi_2 = \left\{ \begin{array}{l} \textit{Hospital services, Other professional services,} \\ \textit{Public administration, Medical except hospital,} \\ \textit{Education, Finance insurance, Social services,} \\ \textit{Forestry and fisheries, Communications,} \\ \textit{Entertainment, Agriculture Construction,} \\ \textit{Manufact. durable goods,} \\ \textit{Manufact. nondurable goods,} \\ \textit{Utilities and sanitary services, Wholesale trade,} \\ \textit{Transportation, Retail trade,} \\ \textit{Business and repair services} \end{array} \right\}$$

In the same way, we apply the test of Equation (9) on other dimensions. In Figure 6, we clearly see that the zones of facts corresponding to characteristic attributes of dimensions $D_2$ and $D_5$ seem to be more interesting and denser than other regions of the data space representation. These zones contain 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. *Asians* people from India, Iran, Japan and China are rather concentrated in commerce and trade activities.

## VI. RELATED WORK

Several works have already treated the issue of enhancing space representations of data cubes. These works were undertaken following different motivations and adopted different ways to address the problem. While some are interested to technical optimization (storage space, queries response time, etc.), others have rather focused on OLAP aspects. Our present work fits into the second category. Recall that, in our case, we focus on assisting OLAP users in order to improve and help analysis processes on large and sparse data cubes. We use a factorial approach to highlight relevant facts and provide data representations interesting for analysis. Nevertheless, we dress an overview of main studies as well in the first as in the second category of works.

In [10], Vitter *et al.* proposed to build compact data cubes by using approximation through wavelets. Another data structure, called `Quasi-Cube` [11], compresses data representation by materializing only sufficient parts of a data cube, the remaining parts are approximated by a linear regression. In [12] approximation is performed by statistical techniques to estimate the density function of data. Method `Dwarf` [13] reduces the storage space of a data cube by identifying and factoring redundant tuples in the fact's table. Wang *et al.* propose to factorize these redundancies by exploiting the notion of `BST` [14] (*Base Single Tuple*). Therefore, a more condensed data cube (`MinCube`) was proposed. In [15], Feng *et al.* introduce `PrefixCube`, a data structure constructed upon only one `BST` by an initial cube dimension. The `Quotient Cube` [16] summarizes semantic contents of a data cube and partitions it into cells with identical values. The best partition corresponds to the minimal lattice structure that allows to browse the reduced cube. In [17], `Quotient Cube` was involved and a newer data structure, `QC-Tree`, was proposed. `QC-Tree` is directly constructed from the base table in order to maintain it under updates. Feng *et al.* [18] propose the `Range CUBE` method to compute and compress a data cube without loss of precision. This method identifies correlation in attributes values and compress the input dataset to reduce the computational cost. Ross and Srivastava [19] addressed the cube representation problem in the case of sparse data. They propose a new algorithm, `Partitioned-Cube`, based on partitioning large relations into *fragments* to fit in memory. Operations over the whole cube are performed on each memory-sized fragment independently. In [20], high dimensional datasets are partitioned into a set of disjoint low dimensional datasets also called *fragments*. For each fragment, a local data cube is computed offline and used to compute queries in an online fashion.

Finally, our approach shares already the same motivation of Choong *et al* [21]. Authors also address the problem of high dimensionality of data cubes. They try to enhance analysis processes by preparing datasets into appropriate representations. Thus, a 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 and considers only integer measures.

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. We do not to reduce it, but we reduce its negative effects on interpretations and OLAP analysis of data. Thus, we use MCA to arrange differently facts and highlight their relevant relationships in a data cube within a visual effect that gathers

| | 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 | | | | | | | | | | | | | | | | | | | | |

Fig. 5. Initial data representation of the *Census-Income*'s data cube.

them as well as possible in the data space representation.

## VII. Conclusion

In this paper we introduced an approach to enhance the space representation of large and sparse data cubes. This approach enables an assistance to OLAP users and helps to explore huge volumes of data. Indeed, our approach identifies and highlights interesting facts and displays them in an appropriate representation by exploiting results of MCA. This representation provides better property for data visualization since it gathers full cells expressing interesting relationships of data. Interesting facts are associated to characteristic attributes selected from the cube's dimensions. These attributes are detected according to their test-values on factorial axes. Furthermore, interesting facts are gathered together in the data representation space. This can solve the problem of high dimensionality, sparsity of data, and allows to concentrate navigation of data on regions containing relevant information.

Some extensions are possible to improve our approach. In our future works, we intend to construct a criterion to measure the quality of a data cube representation. This criterion may enable evaluation of the performance of our approach. We also plan to perform experiments over different configurations of data cubes in order to study the efficiency of our method according to sparsity, number of cells, number of dimensions, and number of facts.

Currently, we are also studying some possible extensions for this work. We consider the problem of optimizing complexity

| | 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 |

Fig. 6. Organized data representation of the *Census-Income*'s data cube.

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. We choose the Web technology to emphasize on the on line and interactive aspect of the approach.

## REFERENCES


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

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

[3] S. Chaudhuri and U. Dayal, "An Overview of Data Warehousing and OLAP Technology," *SIGMOD Record*, vol. 26, no. 1, pp. 65–74, 1997.

[4] J.P. Benzecri, *Correspondence Analysis Handbook*, hardcover ed., Hardcover, Ed. Marcel Dekker, January 1992.

[5] L. Lebart, A. Morineau, and M. Piron, *Statistique exploratoire multidimensionnelle*, 3rd ed. Paris: Dunold, 2000.

[6] R. Cattell, "The Scree Test for the Number of Fators," *Multivariate Behavioral Research*, vol. 1, pp. 245–276, 1966.

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

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

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

[10] 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)*. Philadelphia, Pennsylvania, U.S.A.: ACM Press, June 1999, pp. 193–204. [Online]. Available: citeseer.ist.psu.edu/vitter99approximate.html

[11] D. Barbará and M. Sullivan, "Quasi-Cubes: Exploiting Approximations in Multidimensional Databases," *SIGMOD Record*, vol. 26, no. 3, pp. 12–17, 1997. [Online]. Available: http://www.acm.org/sigmod/record/issues/9709/dbarbara.ps

[12] 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)*, San Diego, California, U.S.A., August 1999, pp. 223–232. [Online]. Available: ftp://ftp.research.microsoft.com/pub/tr/tr-99-13.pdf

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

[14] 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)*, San Jose, California, U.S.A., February–March 2002, pp. 155–165. [Online]. Available: citeseer.ist.psu.edu/wang02condensed.html

[15] 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)*, Washington D.C., U.S.A., November 2004, pp. 38–47.

[16] 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.

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

[18] 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)*, Boston, Massachusetts, U.S.A., March–April 2004, pp. 658–670. [Online]. Available: citeseer.ist.psu.edu/682647.html

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

[20] 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)*, Toronto, Canada, August 2004, pp. 528–539. [Online]. Available: citeseer.ist.psu.edu/642738.html

[21] Y. W. Choong, D. Laurent, and P. Marcel, "Computing Appropriate Representations for Multidimensional Data," *Data & knowledge Engineering Journal*, vol. 45, no. 2, pp. 181–203, 2003. [Online]. Available: http://www.cs.brown.edu/courses/cs227/Papers/Visualization/Choong.pdf