

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/295399010>

The perception of corruption in health: AutoCM methods for an international comparison

Article in *Quality and Quantity* · February 2016

DOI: 10.1007/s11135-016-0315-4

CITATIONS

0

READS

22

8 authors, including:



[Massimo Buscema](#)

Semeion Centro Ricerche

212 PUBLICATIONS 1,550 CITATIONS

[SEE PROFILE](#)



[Lara Gitto](#)

University of Rome Tor Vergata

47 PUBLICATIONS 94 CITATIONS

[SEE PROFILE](#)



[Andrea Marcellusi](#)

University of Rome Tor Vergata

71 PUBLICATIONS 220 CITATIONS

[SEE PROFILE](#)



[Francesco Saverio Mennini](#)

University of Rome Tor Vergata

95 PUBLICATIONS 296 CITATIONS

[SEE PROFILE](#)

*The perception of corruption in health:
AutoCM methods for an international
comparison*

**Paolo Massimo Buscema, Lara Gitto,
Simone Russo, Andrea Marcellusi,
Federico Fiori, Guido Maurelli, Giulia
Massini & Francesco Saverio Mercuri**

Quality & Quantity
International Journal of Methodology

ISSN 0033-5177

Qual Quant
DOI 10.1007/s11135-016-0315-4



Your article is protected by copyright and all rights are held exclusively by Springer Science +Business Media Dordrecht. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".

The perception of corruption in health: AutoCM methods for an international comparison

Paolo Massimo Buscema^{1,2} · Lara Gitto³ · Simone Russo^{3,4} ·
Andrea Marcellusi³ · Federico Fiori³ · Guido Maurelli¹ ·
Giulia Massini¹ · Francesco Saverio Mennini^{3,5}

© Springer Science+Business Media Dordrecht 2016

Abstract According to an OCSE study, corruption in healthcare costs about €56 billion yearly, corresponding to €160 million daily (European Commission, Accompanying document on the draft Commission Decision on Establishing an EU Anti-corruption Reporting Mechanism (“EU Anti-Corruption Report”), 2011). In Italy the situation is even more alarming. A recent study (WHO, The World Health Report: Health Systems Financing: The Path to Universal Coverage, 2010) has estimated that waste, inefficiency and corruption cost our National Health Service about €20 billion a year, corresponding to

✉ Paolo Massimo Buscema
m.buscema@semeion.it

Lara Gitto
gitto@ceis.uniroma2.it

Simone Russo
simone.russo82@hotmail.it

Andrea Marcellusi
andrea.marcellusi@uniroma2.it

Federico Fiori
federico.fiori14@gmail.com

Guido Maurelli
g.maurelli@semeion.it

Giulia Massini
g.massini@semeion.it

Francesco Saverio Mennini
f.mennini@uniroma2.it

¹ SEMEION Research Centre of Sciences of Communication, Rome, Italy

² Department of Mathematical and Statistical Sciences, University of Colorado, Denver, CO, USA

³ CEIS-Economic Evaluation and HTA (EEHTA), University of Rome Tor Vergata, Rome, Italy

⁴ Department of Statistics, University of Rome La Sapienza, Rome, Italy

⁵ Department of Accounting and Finance, Kingston University, London, UK

about 20 % of the total health expense. By nature, corruption is a hidden, complex and difficult to quantify reality. The measurement of this phenomenon may only be approximate and risks not to include important elements such as social costs of corruption that cannot be quantified. For this reason, there is no mathematical formula able to supply a country's level of corruption. However, in the course of time, some research demonstrated that people's perception of corruption gives a rather reliable estimate of the nature and extent of some corruption phenomena in a given context. In this perspective, the objective of this study is to understand how the phenomenon of corruption is perceived across several world countries; in other words, is there a common view concerning corruption? Secondly, we aim at identifying any management "models" of this phenomenon, that may be applied to different social and economic contexts. These objectives are considered through the application and the comparison of results of two different methods of data mining, such as the principal component analysis and the auto-contractive map, based on the analysis of a dataset.

Keywords Healthcare corruption · Data Mining · Artificial Neural Network · Auto-Contractive Map

1 Introduction

The term "corruption" is generally referred to two kinds of actions: corruption, strictly speaking meant as an "abuse of power for private profit" (Transparency International 2009), and the violation of ethical principles that, even if tolerated, allowed or not prosecuted by law, in any case damage the effectiveness and quality of services. Therefore, according to this definition, corruption is also a form of moral aberration causing, in addition to economic damage, shortcomings, inefficiency, waste, thus limiting growth and investment in social services (ISPE Sanità 2014).

Specifically, there are different definitions of the term corruption. One of these is based on the quantity of money involved, and the sector in which corruption takes place. This definition includes "Big corruption", "Small corruption" and "Political corruption". The first one refers to actions affecting the functioning of administrations at central level, at the expense of the common good; the second kind of corruption corresponds to daily abuse of power, taking place in the relationships between public officials and common citizens who often try to access basic goods and services; the third one consists in the manipulation and abuses of decision-makers on institutions and procedures to increase their power, richness and status.

Healthcare is one of the most corrupted sectors (after construction industry, energy production and mine industry) affecting citizens' health and quality of life.

According to the World Health Organization (WHO 2010), "many—if not all-countries do not succeed in fully exploiting the available resources, due to the bad administration of contracts/purchases, the irrational use of drugs, the bad management and allocation of technical and human resources or a discontinuous financial and administrative management".

From this point of view, the European situation is rather worrying.

According to an OCSE study, corruption in healthcare costs about €56 billion yearly, corresponding to €160 million daily (European Commission 2011). In Italy the situation is even more alarming. A recent study (WHO 2010) has estimated that waste, inefficiency and corruption cost our National Health Service about €20 billion a year, corresponding to about 20 % of the total health expense.

By nature, corruption is a hidden, complex and difficult to quantify reality. The measurement of this phenomenon may only be approximate and risks not to include important elements such as social costs of corruption that cannot be quantified. For this reason, there is no mathematical formula able to supply a country's level of corruption.

However, in the course of time, some research demonstrated that people's perception of corruption gives a rather reliable estimate of the nature and extent of some corruption phenomena in a given context (Autorità Nazionale AntiCorruzione 2014; Transparency International 2013a; Transparency International 2013b; Governo Italiano, Presidenza del Consiglio dei Ministri 2012).

In this perspective, the objective of this study is to understand how the phenomenon of corruption is perceived across several world countries; in other words, is there a common view concerning corruption?

Secondly, we aim at identifying any management "models" of this phenomenon, that may be applied to different social and economic contexts.

These objectives are considered through the application and the comparison of results of two different methods of data mining, such as the principal component analysis (PCA) and the auto-contractive map, based on the analysis of a dataset including data collected by *Transparency International* on citizens' perception of corruption.

2 Data

The data employed in the present analysis have been collected for the Global Corruption Barometer (GCB) of Transparency International (Transparency International 2013b). The edition of 2013 supplied the answers of 114,270 people from 107 countries.

GCB inquires on different aspects of corruption: through the consideration of citizens' opinions it detects the presence of corruption in a given context, evaluates population's involvement with corruption in the year prior to the interview, and tries to understand on which institutions, including healthcare) the problem has a wider impact. Furthermore, it detects the opinion of the people interviewed on the actions taken by administrations in the fight against corruption. A short list of questions with related answers, used to build the dataset, is indicated here below:

Q1. Over the past 2 years, how has the level of corruption in this country changed?

% Increased and Increased a lot.

Q2. To what extent do you think that corruption is a problem in the public sector in this country?

Scale 1–5, aggregate score.

Q3. In your dealings with the public sector, how important are personal contacts to get things done?

% Important or very important, aggregate.

Q4. To what extent is this country's government run by few big entities acting in their own best interests?

% Large extent or entirely, aggregate.

Q5. How effective do you think your government's actions are in the fight against corruption?

% Ineffective.

Q6. Perception of corruption, by institution

% of people that think corruption exists in medical and health institutions.

Q7. Have you paid a bribe to at least one of the 8 listed services in the past 12 months?
% people that paid a bribe for medical and health service.

Q8. What was the most common reason for paying the bribe/bribes?
% It was the only way to obtain a service.

Q9. Can ordinary people make a difference in the fight against corruption?
% Disagree and Strongly disagree.

Can ordinary people make a difference in the fight against corruption

Q10. There are different things people could do to fight corruption. Would you be willing to do any of the following (answer 'yes' or 'no'): Sign a petition asking the government to do more to fight corruption; Take part in a peaceful protest or demonstration against corruption; Join an organisation that works to reduce corruption as an active member; Pay more to buy goods from a company that is clean/corruption-free; Spread the word about the problem of corruption through social media; Report an incident of corruption.

Q11. Would you report an incident of corruption?
% Yes.

Q12. Have you ever been asked to pay a bribe?
% Yes.

In addition to the indicators detected by the answers to the questionnaire, other three variables have been taken into account in the dataset (Transparency International 2012) to define a ranking of the countries according to the degree of corruption perception among public and political officials. First, the corruption perception index (CPI) has been considered: it is a composed index whose value ranges from zero (totally corrupt country) to ten (total absence of corruption). CPI refers to the same countries taken into account by GCB and is always based on perception only. It is made up of 14 different surveys and studies carried out in 12 independent institutions.

The other two variables considered represent the percentage of total health expenditure (both public and private) on the gross domestic product (The World Bank 2014a) and the percentage of public health expense on total health expense (The World Bank 2014b).

The resulting dataset is made up of 107 observations, corresponding to the examined countries, and 14 variables including the most important indicators.

3 Methods

Two different techniques have been employed in the analysis: the classical Principal Components Analysis (PCA) and the auto contractive map (Auto-CM) method.

PCA is used in multivariate statistical analysis to simplify the source data. It explains the structure of the variance and covariance of a dataset through a few linear combinations of the original variables.

In fact, the variability of a complex set of p indicators can be largely explained by a small number, $k < p$ of variables, defined as “principal components” or “factors”.

The k principal components contain all the information available in the p initial variables and may, therefore, allow to replace the original dataset; the latter consists of n measurements of p variables, with a dataset alternative, consisting of n measurements of $k < p$ principal components.

In other words, PCA allows to decrease the complexity of some phenomena, by considering that just two or three latent dimensions, by themselves, are able to explain the whole variability of the phenomenon under consideration.

Instead, the Auto-CM method was invented, tested and implemented in C language by Paolo Massimo Buscema in 1998 at Semeion Research Centre (Buscema 1998).

Auto-CM is characterized by a three-layer architecture: an Input layer, where the signal is captured from the environment, a Hidden layer, where the signal is modulated inside the CM, and an Output layer, by means of which the CM influences the environment on the basis of previously received stimuli. Each layer contains an equal number of N units. Therefore, the whole CM is made up of $3 N$ units. The connections between the Input and the Hidden layers are mono-dedicated, whereas those between the Hidden and the Output layers are at maximum gradient (Buscema et al. 1998).

With respect to the number of units, the corresponding number of connections N_c , is given by: $N_c = N(N + 1)$, see Fig. 1.

All CM connections may be initialized either by assigning the same value to each of them, or by assigning values at random. The best practice is to initialize all the connections with the same, positive value, close to zero.

The learning algorithm of CM may be summarized in a sequence of four orderly steps:

1. Signal Transfer from the Input into the Hidden layer
2. Adaptation of the connection values between the Input and the Hidden layers*
3. Signal Transfer from the Hidden into the Output layer*
4. Adaptation of connection value between the Hidden and the Output layers.

(*) steps 2 and 3 may take place in parallel.

We define $m[s]$ as the units of the Input layer (sensors), scaled between 0 and 1; $m[h]$ as the units of the Hidden layer; and $m[t]$ as the units of the Output layer (system target). We define v , the vector of mono-dedicated connections; w , the matrix of the connections between the Hidden and the Output layers; and n , the discrete time of the weights evolution.

There are four signal forward transfer and learning equations:

- signal transfer from the Input to the Hidden layer:

$$m_{i(n)}^{[h]} = m_{i(n)}^{[s]} \left(1 - \frac{V_{i(n)}}{C} \right) \tag{1}$$

where C Contractive Factor, is a positive real number.

- adaptation of the connections $V_{i(n)}$ through $\Delta V_{i(n)}$ trapping the energy difference generated by the Eq. (1):

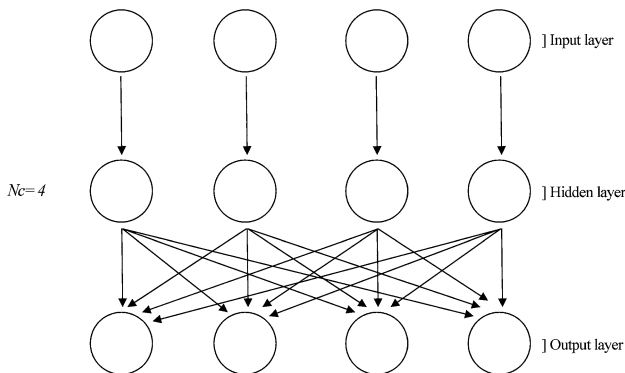


Fig. 1 Example of Auto-CM with $N = 4$

$$\Delta v_{i(n)} = \left(m_{i(n)}^{[s]} - m_{i(n)}^{[h]} \right) \left(1 - \frac{v_{i(n)}}{C} \right) \tag{2}$$

$$v_{i(n+1)} = v_{i(n)} + \Delta v_{i(n)} \tag{3}$$

– signal transfer from the Hidden to the Output layer:

$$Net_{i(n)} = \sum_j^N m_{j(n)}^{[h]} \left(1 - \frac{w_{ij(n)}}{C} \right) \tag{4}$$

$$m_{i(n)}^{[r]} = m_{i(n)}^{[h]} \left(1 - \frac{Net_{i(n)}}{C} \right) \tag{5}$$

– adaptation of the connections $w_{i,j(n)}$ through $\Delta w_{i,j(n)}$ trapping the energy difference generated by the Eq. (5):

$$\Delta w_{i,j(n)} = \left(m_{i(n)}^{[h]} - m_{i(n)}^{[r]} \right) \left(1 - \frac{w_{ij(n)}}{C} \right) m_{j(n)}^{[h]} \tag{6}$$

$$w_{i,j(n+1)} = w_{i,j(n)} + \Delta w_{i,j(n)} \tag{7}$$

The value $m_{j(n)}^{[h]}$ of the Eq. (6) is used to make the change of the connection $W_{i,j(n)}$ proportional to the quantity of energy liberated by node $m_{j(n)}^{[h]}$ in favour of node $m_{i(n)}^{[r]}$ (refer to Buscema and Sacco 2010; Buscema et al. 2014 for related mathematical steps).

The Auto Contractive Maps do not behave as other regular artificial neural networks. In fact, they are able to start from all connections created with the same values. In this way they do not suffer the problem of symmetric connections.

During the training phase, they develop only positive values for each connection. Therefore, they do not present inhibitory relations among the nodes, but only different strengths of excitatory connections.

They can also work in hard conditions, that is, when the connections of the main diagonal of the second connections matrix are removed. When the learning process is organized in this way, the Auto-CM seems to find a specific relationship between each variable and the others. Consequently, from an experimental point of view, it seems that the ranking of its connections matrix is equal to the ranking of the joint probability between each variable and the others.

After the learning process, each Input vector belonging to the learning set will generate a null Output vector. Therefore, the energy minimization of the training vectors is represented by a function through which the trained connections completely absorb the Input formation vectors. At the end of the training phase ($\Delta W_{ij} = 0$), all the components of the weights vectors V reach the same value:

$$\lim_{n \rightarrow \infty} v_{i(n)} = C \tag{8}$$

The matrix w represents the Auto-CM knowledge about all the dataset. It is possible to convert the matrix w in the probabilistic joint association among the variables m :

$$P_{ij} = \frac{w_{i,j}}{\sum_j^N w_{i,j}} \tag{9}$$

$$P\left(m_j^{[s]}\right) = \sum_j^N p_{ij} = 1 \tag{10}$$

The new matrix p can be read as the probability of transition from any state-variable to any other else:

$$P\left(m_i^{[t]}|m_j^{[s]}\right) = p_{ij} \tag{11}$$

At the same time, the matrix w can be converted into a non Euclidean distance (U-metric) when we train the CM with the main diagonal of matrix w fixed at value N .

Now, if we consider N as a limit value for all the weights of matrix w , we can write:

$$d_{i,j} = N - w_{ij} \tag{12}$$

The new matrix d is also a squared symmetric matrix where the main diagonal represents the zero distance between each variable from itself.

3.1 The contractive factor

We now discuss in detail the interpretation of the squared weights matrix of the Auto-CM system (Buscema and Grossi 2008; Buscema et al. 2008).

We should assume that each variable of the dataset is a vector made up of all its values. In this perspective, the dynamic value of each connection between two variables represents the local velocity of their mutual attraction caused by the degree of similarity of their respective vectors: the more the vectors' similarity, the higher their attraction velocity. When two variables are attracted to each other, they proportionally “contract” their original Euclidean space. The limit case takes place when the two variables are identical: the space contraction should be infinite and the two variables should collapse at the same point. We can extract from each weight of a trained Auto-CM this specific contractive factor:

$$F_{ij} = \left(1 - \frac{W_{ij}}{C}\right)^{-1} ; \quad 1 \leq F_{ij} \leq \infty \tag{13}$$

This equation is interesting for the following three reasons:

- It is the inverse of the equation used as contractive factor that rules the Auto-CM training.
- Considering the Eq. (3), each mono-connection v_i will reach value C at the end of the training. In this case, the contractive factor will be infinite because the two variables connected by the weight are indeed the same variable.
- Instead, considering the Eq. (7), each weight w_{ij} will always be smaller than C at the end of the training. This means that the contractive factor for each weight of the matrix we are considering will always be bounded (for mathematical see Buscema et al. 2014). In fact, in the case of weight $w_{i,i}$, the variable is connected to itself, but the same variable has also received the influences of other variables (it should be reminded that matrix w is a squared matrix where each variable is linked to the other). Consequently, this variable will not have to be exactly the same.

At this point, we are able to calculate the contractive distance among each variable and the others, by adjusting the original Euclidean distance by a specific contractive factor. The Euclidean distance among the variables in the data set is given by the following equation:

$$d_{i,j}^{[Euclidean]} = \sqrt{\sum_k^R (x_{i,k} - x_{j,k})^2} \tag{10a}$$

Consequently, the Auto-CM matrix of the distances among the same variables is the following:

$$d_{i,j}^{[AutoCM]} = \frac{d_{i,j}^{[Euclidean]}}{F_{i,j}} \tag{11a}$$

3.2 Auto-CM and minimum spanning tree

Equation (12) transforms the squared weights matrix into a squared matrix of distances among nodes. Consequently, each distance between a pair of nodes becomes the weighted edge between this pair of nodes. At this point the matrix d may be analysed through the graph theory.

A graph is a mathematical abstraction very useful for solving different kinds of problems. It basically consists of a set of vertices and a set of edges, where an edge represents the connection between two vertices in a graph. More precisely, a graph is a pair (V, E) where V is a finite set and E is a binary relation on V , to which a scalar value can be associated (in this case the weights are the distances $d_{i,j}$). V is a set of vertices and E is a set of edges, where an edge is a pair (u, v) with u, v in V . In a directed graph, the edges are ordered pairs, obtained connecting a source vertex to a target vertex. In an undirected graph the edges are not ordered in pairs, and the two vertices can be connected in both directions. Consequently, in this situation (u, v) and (v, u) are two different ways of writing the same edge.

The graph does not supply information on what an edge or a vertex actually represent. They could be cities with road connections, Web pages with hyperlinks, etc. These details are not included in the graph for an important reason, because they are not required in the graphic abstraction.

The representation of an adjacency matrix of a graph is a two-dimensional matrix $V \times V$, where the rows represent the list of vertices and the columns the edges among vertices. To each element of the matrix a Boolean value is assigned, indicating if the edge (u, v) is in the graph.

Table 1 Adjacency matrix of a distance matrix

	A	B	C	D	...	E
A	0	1	1	1	1	1
B	1	0	1	1	1	1
C	1	1	0	1	1	1
D	1	1	1	0	1	1
...	1	1	1	1	0	1
E	1	1	1	1	1	0

A distance matrix among vertices V represents an undirected graph, where each vertex is connected with all the others except for itself (see Table 1).

It is now interesting to introduce the concept of Minimum Spanning Tree (Buscema and Sacco 2010).

The Minimum Spanning Tree is defined as follows: it consists in finding an acyclic subset T of E connecting all the vertices together in the graph, whose total weight is reduced to a minimum, where the total weight is given by:

$$d(T) = \sum_{i=0}^{N-1} \sum_{j=i+1}^N d_{ij}, \quad \forall d_{-}(i,j) \quad (14)$$

T is the spanning tree and MST is T with the minimum weighted sum of its edges.

$$MST = \text{Min}\{d(T_k)\} \quad (15)$$

Given an undirected graph G , representing a matrix d of distances, with V vertices completely connected between them, the total number of their edges (E) will be the following:

$$E = \frac{V(V-1)}{2} \quad (16)$$

And the number of its possible tree will be the following:

$$T = V^{V-2} \quad (17)$$

In 1956 Kruskal found an algorithm able to determine the MST of any undirected graph (Kruskal 1956). Obviously, it generates a possible MST. In fact, in a weighted graph more MSTs are possible.

From a conceptual point of view, MST represents the state in which the energy of a structure is minimized. For example, if we consider the atomic elements of a structure as vertices of a graph, and the strength among them as the weight of each edge, connecting a pair of vertices, MST represents the minimum energy required to allow all the elements of the structure to stay together.

In a closed system, all the components tend to minimize the total energy. In this way, in specific cases, MST represents the most probable state to which a system tends.

In order to define MST of an undirected graph, each edge of the graph shall be weighted. Equation (12) shows a way to weigh each edge whose nodes represent the variables of a dataset and whose weights provide the parameters.

Obviously, it is possible to use any kind of Auto-Associative ANN or any kind of Linear Auto- Associator to generate a weighted matrix of variables in an assigned dataset (Buscema et al. 2014).

In most cases, the mean squared error stops and starts decreasing after few cycles, especially when the orthogonality of the records increases. This usually takes place when it is necessary to weigh the distance among the records of the assigned dataset. In fact, in this case, it is necessary to create the transposed matrix of the assigned dataset.

If a linear Auto-associator is used for the purpose, all the non linear association among the variables will be lost. Therefore, Auto-CM seems to be the best choice to calculate a complete and non-linear matrix of weights among the variables or the records of any assigned dataset.

3.3 A few qualitative characteristics of MST

When we have a distance matrix among the nodes $d_{i,j}$ with $i,j \in [1, 2, \dots, N]$, the MST of the implicit graph is easy to define, using the Kruskal algorithm. The adjacency matrix should then be analysed. The easiest criterion to study the adjacency matrix is to sort the number of connections of each node. This algorithm defines the connectivity of each node:

$$C_i = \sum_j^N l_{i,j}$$

where: if $l_{i,j} \in \text{MST}$ then $l_{i,j} = 1$, if $l_{i,j} \notin \text{MST}$ then $l_{i,j} = 0$, $l_{i,j}$ = possible direct connection between Node_{*i*} and Node_{*j*}

- the nodes with just one connection are known as “leaves”. The leaves define the boundaries of the MST graph.
- the nodes with two connections are known as “connectors”.
- the nodes with more than two connections are known as “hubs”. Each hub has a specific degree of connectivity.

$$\text{HubDegree}_i = C - 2 \tag{18}$$

A second indicator quantifying a MST graph is the *cluster* of the strength of each of its nodes. The cluster of the node strength is proportional to the number of its links and the number of the directly connected node links:

$$S_i = \frac{C_i^2}{\sum_{j=1}^C C_j} \tag{19}$$

3.4 Comparing Auto-CM and PCA

Auto-CM is a very good alternative to the PCA for different reasons, especially for its limitations that are not present in the Auto-CM.

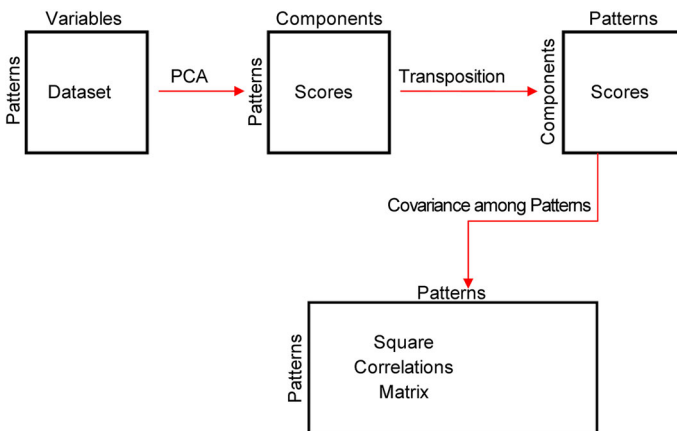


Fig. 2 Transformation process of the initial dataset in the square matrix of correlations through the PCA

For example, PCA:

- is characterized by a high *sensitivity* in presence of anomalous data. In fact, just one case explained with different data from those contained in the rest of the data set, may substantially change the direction of the principal components, or the definition itself of the latent variables;
- has *difficulty in preserving the geometrical structure* of the original space in case there are relations among non linear variables. This does not take place in the Auto-CM model;
- has *difficulty in establishing precise associations* among variables whose only knowledge is adjacency; is characterized by a *mapping generally based on a specific kind of “distance”* among the variables (for example, Euclidean, correlation, etc.). This generates a static projection of possible associations, causing the loss of intrinsic dynamics due to active interactions of the variables in the life systems, characterizing the real world. On the contrary, the Auto-CM succeeds in trapping these dynamics, as it is not based on a distance, but on the weight of the variables.

In Fig. 2, we describe the process used to transform the initial dataset in the final square correlations matrix used to find the MST through the PCA. From the initial matrix, the PCA estimates the scores of the components identified. This second matrix has to be transposed in order to calculate the correlations between the observations and construct the weights matrix used to find the MST.

3.5 MST similarity index

The MST Similarity Index is a new method to measure the capability of any unsupervised algorithm to explain statistically any dataset. The prerequisite to apply this technique is that the algorithm describes the assigned dataset through a square and symmetric matrix of parameters, where each parameter defines the association of each couple of variables of the dataset. In fact, in a second step, a Minimum Spanning Tree (MST) will be calculated from the symmetric and square matrix of the parameters of the analyzed algorithm, in order to measure the amount of similarities that the analyzed algorithm has detected from the assigned data set. We have named this evaluation method MST Similarity Index, and the author is Paolo Massimo Buscema.

This new method generates two indices measuring different properties of the algorithm used for the analysis of the dataset:

- a. The main weighted similarity (named “Main Fitness”): it is an index measuring the amount of weighted similarity that the algorithm was able to code in its MST;
- b. The global weighted similarity (named “Global Similarity”): it is an index measuring the similarity detected by the algorithm in all the derived MSTs that is possible to generate from the matrix of parameters of the algorithm.

First of all, we have to define the concept of “intrinsic dataset similarity”.

Let us image a dataset Ds of N variables and M patterns:

$$Ds = \{x_i\}_{i=1}^M; \tag{20}$$

where $\|x\| = N$.

We need to scale linearly each variable of the dataset into the unit interval $[0,1]$:

$$Ds = \{x_i\}_{i=1}^M \rightarrow D = \{v_i\}_{i=1}^M; \tag{21}$$

where $v \in [0, 1]$.

At this point, using the intersection operator, we can calculate the “intrinsic similarity” of the assigned dataset:

$$Sim(D) = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \sum_k^M v_{i,k} \cdot v_{j,k}; \tag{22}$$

where $\forall v \in [0, 1]$;

The second step consists to scale linearly the square and symmetric matrix of parameters, w , found by the analyzed algorithm into the unit interval $[0,1]$:

$$w_{i,j} \rightarrow m_{i,j}; \tag{23}$$

where $m \in [0, 1]$; $i, j \in N$.

The third step consists to calculate the MST among the variables of the assigned dataset, using the symmetric and square matrix of parameters, m , generated by the analyzed algorithm:

$$Mst(V, E) = f(A, D, m_{i,j}); \tag{24}$$

where D the assigned dataset scaled in the unit interval, A the analyzed algorithm, $m_{i,j}$ the parameters found by the algorithm, i, j the indexes of the parameters $i, j \in N$, V the notes (variables) of the tree, E The undirected links of the tree.

At this point we can calculate the “weighted similarity” that the algorithm was able to code into the Mst:

$$Sim(A) = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \sum_k^M m_{i,j} \cdot (v_{i,k} \cdot v_{j,k}); \quad \text{if } (v_{i,k}, v_{j,k}) \in E \tag{25}$$

Finally, we can define the “Main Fitness” of the analyzed algorithm:

$$Main_Fitness(A) = \frac{Sim(A)}{Sim(D)}. \tag{26}$$

On this basis, we can define also another index able to consider all the MSTs that it is possible to generate from the weights matrix of each algorithm. The procedure is the following:

- a. Extract from the square and symmetric matrix of the analyzed algorithm parameters, m , the first MST;
- b. Remove the matrix parameters, m , all the connections used for the first MST;
- c. Try to build a new MST using only the not removed parameters;
- d. If it is impossible to build a new MST stop the process, else extract and remove the new parameters from the matrix, m , and go back to the step c;

This procedure defines for any analyzed algorithm the maximum number of MSTs allowed from its matrix of parameter, m : more is the number of MSTs generated, the better is the distribution of the parameters m in the matrix.

We have named this index “Global Weighted Similarity”.

The equations to calculate the Global Weighted Similarity for each algorithm is the following:

Global Similarity found by the algorithm (A), in all MSTs:

$$Global_Sim(A, MSTs) = \frac{1}{Sim(D)} \sum_{n=1}^{NumMst} \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \sum_{k=1}^M m_{n,i,j}(v_{i,k} \cdot v_{j,k}); \text{ if } (v_{i,k}, v_{j,k}) \in E^{Mst_n}; \tag{27}$$

where $Sim D$ see Eq. 22, $NumMst$ total number of MSTs found by the algorithm in the m matrix, Mst_n the n th MST found by the algorithm in the m matrix.

In few words: Eqs. (26) and (27) measure two different properties of the algorithms through which we analyze any dataset:

- a. The amount of main weighted similarity detected into the assigned dataset by the main MST found by the algorithm used for the analysis (Eq. 26);
- b. The amount of global weighted similarity found by the algorithm used because of the rate of uncorrelated distribution of its parameters into the matrix m (Eq. 27).

The Similarity Sum is basically a good measure for 3 main reasons. First of all, the similarity indicates how much similar information is compared. Therefore, the more similar information is compared, the better the analysis. Furthermore, as the sum is iterated for all MSTs, the more MSTs can be derived, the better the weights are distributed. Finally, the more similarities can be counted at different levels, the more correctly the system distributed the weights and, therefore, carried out a good job. For these reasons, the results obtained by the comparisons made in our analysis can be deemed acceptable to explain the differences between the two methods.

In order to validate the method used to make the comparison, some easy to understand “toy” datasets have been used. Their distributions are already known. Once the validity of the method has been demonstrated, it has been possible to make an analysis on the dataset concerning corruption.

4 Results

4.1 The similarity index

The values of the two indices of similarity clearly show the greater capability of Auto-CM to explain statistically the dataset.

In particular, the comparison between the two values of the Main Similarity reported in Table 2, shows that the Auto-CM find a main MST able to detect a larger amount of main weighted similarity.

The lower number of total MSTs found by Auto-CM compared to the number found by PCA (Tables 2 and 3), shows the increased efficiency and capability to explain the dataset of this method because with a smaller number of steps still detects a greater amount of global weighted similarity.

Table 2 Main similarity, global similarity and number of MSTs for Auto-CM and PCA

MST index	Main similarity	Global similarity	Number of MSTs
Auto-CM	0.021768	0.160994	32
PCA	0.019052	0.141284	38

Table 3 Global similarity and number of MSTs for Auto-CM and PCA

Name	Criterion	Num MSTs	Global similarity	1	2	3	4	5	6	7	8	9	10	11
Auto- CM	Fitness	32	0.160994	0.021768	0.021436	0.021294	0.021126	0.021292	0.02089	0.020655	0.020363	0.020447	0.020557	0.020186
PCA	Fitness	38	0.141284	0.019052	0.018588	0.018478	0.018025	0.017486	0.017439	0.017157	0.017318	0.017014	0.016709	0.016704
Name	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Auto- CM	0.020108	0.019852	0.019686	0.019651	0.01952	0.019708	0.019401	0.019331	0.019243	0.018971	0.018773	0.018782	0.018539	0.018391
PCA	0.016606	0.016196	0.015934	0.016031	0.015922	0.015787	0.015793	0.015551	0.015351	0.015102	0.015083	0.01507	0.015058	0.014736
Name	26	27	28	29	30	31	32	33	34	35	36	37	38	
Auto-CM	0.0184	0.018011	0.018147	0.017915	0.017512	0.017439	0.01725							
PCA	0.014624	0.014712	0.014573	0.014286	0.014044	0.013855	0.013844	0.013527	0.013735	0.013409	0.013084	0.012959	0.012947	

Table 4 Correlation between Auto-CM and PCA matrices

	R^2
Auto-CM-PCA matrices	0.350352

Finally, the low correlation between the matrices of the two different algorithms (Table 4) is a further demonstration of the greater capability of Auto-CM to detect elements and similarities that the PCA does not capture.

The graphical representation shows with clarity the pattern that we expect to obtain.

As we have already said explaining how the dataset has been built, data are related to countries located in different continents: in some countries, the sense of public responsibility is higher and, consequently, the level of corruption in basic services as health is lower; in other countries, characterized by political instability (as, for example, developing countries), public institutions are weaker and cannot develop measures likely to contrast efficiently the corruption in sectors as health and other public services.

In constructing the MST graph, the software observes the numerical values of CPI only, together with other variables defined earlier that are employed to describe how people living in countries considered in the analysis “feel” the level of corruption.

Nonetheless, it succeeds in identifying common patterns in the perception of corruption employing all information, even the “hidden information” in the dataset.

4.2 Auto-CM

By analysing the MST graph, processed through the Auto-CM network in Fig. 3, at least 6 main hubs can be identified.

As explained in the methodology, the Auto-CM creates connections according to a similarity criterion. First of all, by observing the graph in detail, we note that the network connects geographically close countries where similar answers have been supplied.

There are connections among almost all African countries such as Malawi, Kenya, Zimbabwe, Tanzania, Cameroon, Sudan, Libya, Ethiopia, South Africa and others.

In the same way, there are connections among Slovakia, Croatia, Czech Republic, Hungary, Romania, Bosnia, Slovenia, etc. The Northern European social democracies (Finland, Norway and Denmark) are rather connected and similar to other countries like Canada, the United Kingdom and Australia, in addition to France, Germany and Switzerland.

Italy belongs to the same group as Spain, Portugal and Greece, the most crisis-stricken Eurozone countries, along with Kosovo and some Eastern European countries.

The answers to the questions considered for the dataset are almost exclusively based on perceptions. Therefore, it is not possible to refer to “corruption models”, but, rather, to how corruption, at a general level, is perceived.

Based on this graph, it is possible to ascertain how people from geographically close countries or with similar socio-economic conditions, perceive corruption in a similar way.

A different picture can be observed when representing the graph obtained through PCA method.

In Fig. 4, after extracting two factors, the matrix of their Euclidean distances has been calculated with the factors’ coordinates, and a corresponding Minimum Spanning Tree (MST) has been processed.

It is not possible to identify any defined group in the graph, but, rather, a long “chain” with little nodes.

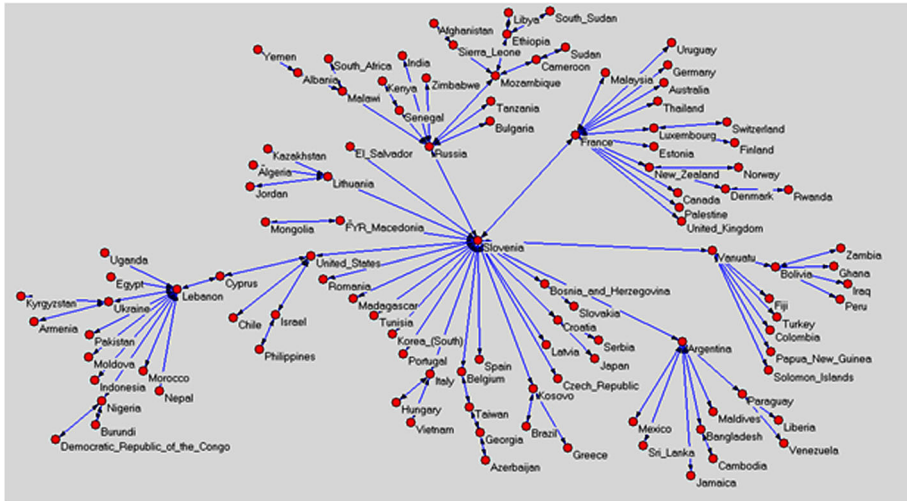


Fig. 3 MST resulting from the Auto-CM on countries (dataset 14×107)

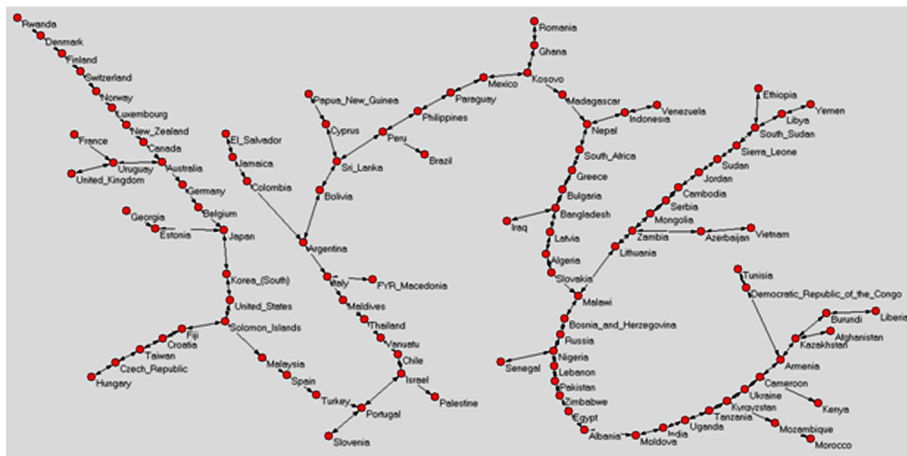


Fig. 4 MST resulting from PCA on Countries (dataset 14×107)

In some way, the collinearity corresponds to what emerged in the previous analysis. Northern European countries as Denmark, Finland, Switzerland and Norway are connected each other. Through Australia, that creates a little connection node, they are linked to other European countries as Germany and Belgium, followed by Asian countries as Japan and Korea.

In other groups, we have a different situation. In fact, Eastern European countries, that in the Auto-CM graph create a rather compact group, here are scattered along the chain.

We can note the same situation for African countries that, in addition to a small group including Lybia, Sudan, Ethiopia, South Sudan and Sierra Leone, are separated in different points.

The graph shows that PCA supplies a MST characterized by a sequence of countries, with very few nodes. This makes its interpretation difficult, although their adjacency is rather logical, while the MST resulting from the Auto-CM method shows different nodes to which countries with similar characteristics are connected.

Therefore, the second graph clearly shows a greater strength and descriptive clarity.

Even visually, the Auto-CM produces clearer results and easier to understand information than those produced by PCA.

5 Conclusion

The objective of the study was to analyse the phenomenon of corruption in the health sector and to understand how it is perceived across several world countries; in fact, in case a common view concerning corruption can be identified, it could be possible to develop policy models to contrast it, favouring the cooperation across countries which present similarities.

The Auto-CM method, widely described in the paper and its further specification based on the Similarity Index, proved effective in outlining such similarities.

Auto-CM has shown remarkable differences comparing to PCA method, when analyzing, as in this case study, data on perceived corruption in public services. Considering the outcome of the dataset analysis, the Auto-CM method succeeds in supplying significant results.

This is also confirmed at the analytical level—given the particularly high Similarity Sum of recurrent MST—and, above all, from a graphic and visual point of view. The building of trees is characterized by nodes and links that allow an easy reading of the results. This is partly due to a low sensitivity to abnormal data, to a higher capacity of preserving the geometrical structure of the original space, when the relations between non linear variables are present and to the fact that the mapping is not based on a specific kind of “distance” among variables (for example Euclidean, correlation, etc.), but, rather, on the weight of variables. In this way, the intrinsic dynamics are not lost due to active interactions of the variables in the life systems characterizing the real world.

Auto-CM correctly identifies the similarities between geographically close contexts and realities, where similar answers concerning corruption in the health sector have, evidently, been supplied by the people interviewed.

The analysis produces a graph where different sets can be identified: a group of countries containing almost all African countries; a group of Eastern European countries; a group of northern European social democracies, together with other countries like Canada, the United Kingdom and Australia and, close to them, France, Germany and Switzerland, along with Italy, belonging to the same group as Spain, Portugal and Greece—the most crisis-stricken Eurozone nations.

The groups of countries are ordered according to geographical proximity or socio-economic variables, as the average income or the level of development, measured through the Human Development Index (HDI).

The European countries, characterized by a high or very high HDI are close each other, while those countries characterized by a average or low HDI are spread over the graph.

Considering the purpose of the analysis (the identification of possible patterns of corruption), based on the results obtained it could be stated that citizens of countries with a similar HDI, perceive corruption in health in the same way.

However, this is only one of the possible interpretations of the chart obtained, and certainly not the only explanation to be taken into account.

Moreover, there is another relevant limitation to consider: since the dataset employed is almost exclusively based on data from an opinion survey, is not possible to clearly define a “corruption model”, aimed at explaining the causes of corruption but, as it has been stressed, only seeing how the phenomenon is perceived across different countries. In countries culturally characterized by high levels of corruption, corruption may not be even properly recognized. In the same way, in countries with a very low level of corruption, a single event may highly influence public opinion.

By identifying the similarities between countries it might be possible to plan intervention strategies common to more countries, particularly to those ones that show a similar attitude in perceiving corruption.

This is the novelty and the main contribution offered by this paper: together with the description of the phenomenon, there is the application of an innovative method of analysis, likely to be successfully replicated in more detailed studies.

References

- Autorità Nazionale AntiCorruzione (ANAC): Corruzione sommersa e corruzione emersa in Italia. Modalità di misurazione e prime evidenze empiriche. <http://www.anticorruzione.it/portal/rest/jcr/repository/collaboration/Digital%20Assets/anacdocs/Attivita/Pubblicazioni/RapportiStudi/Metodologie-dimisurazione.pdf> (2014). Accessed 30 May 2015
- Buscema, M.: Constraint satisfaction neural networks. In: Buscema, M. (ed.) *Special Issue on Artificial Neural Networks and Complex Social Systems*, pp. 389–408. *Substance Use & Misuse*, New York (1998)
- Buscema, M., Grossi, E.: The semantic connectivity map: an adapting self-organizing knowledge discovery method in data bases. Experience in Gastro-oesophageal reflux disease. *Int. J. Data Min Bioinform* **2**, 362–404 (2008)
- Buscema, P.M., Sacco, P.L.: Auto-contractive maps, the h function, and the maximally regular graph (MRG): a new methodology for data mining. In: Capecchi, V., Buscema, P.M., Contucci, P., D'Amore, B. (eds.) *Applications of Mathematics in Models. Artificial Neural Networks and Arts*. Springer, New York (2010)
- Buscema, P.M., Consonni, V., Ballabio, D., Mauri, A., Massini, G., Breda, M., Todeschini, R.: K-CM: A new artificial neural network. Application to supervised pattern recognition. *Chemom Intell Lab Syst* **138**, 110–119 (2014)
- Buscema, M., Grossi, E., Snowdon, D., Antuono, P.: Auto-contractive maps: an artificial adaptive system for data mining. An application to alzheimer disease. *Curr. Alzheimer Res.* **5**, 481–498 (2008)
- European Commission: Commission Staff Working Paper. Accompanying document on the draft Commission Decision on Establishing an EU Anti-corruption Reporting Mechanism (“EU Anti-Corruption Report”). In: European Commission, Brussels, pp. 1–69 (2011)
- Governo Italiano, Presidenza del Consiglio dei Ministri: La corruzione in Italia. Per una politica di prevenzione. Analisi del fenomeno, profili internazionali e proposte di riforma. Rapporto della Commissione per lo studio e l'elaborazione di proposte in tema di trasparenza e prevenzione della corruzione nella pubblica amministrazione. http://trasparenza.formez.it/sites/all/files/Rapporto_corruzioneDEF_ottobre%202012.pdf (2012). Accessed 30 May 2015
- Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.* **7**, 48–50 (1956). doi:10.1090/S0002-9939-1956-0078686-7
- Mennini, F.S., Maurelli, G., Gitto, L., Ruggeri, M., Russo, S., Cicchetti, A., Buscema, P.M.: Modello per l'individuazione di inefficienze e corruzione in sanità. In: ISPE Sanità, Libro bianco sulla corruption in sanità, pp. 120–145. ISPE (2014). http://www.ispe-sanita.it/1/libro_bianco_3743257.html
- The World Bank: Health expenditure, total (% of GDP). <http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS> (2014a). Accessed 30 May 2015
- The World Bank: Health expenditure, public (% of total health expenditure). <http://data.worldbank.org/indicator/SH.XPD.PUBL/countries> (2014b). Accessed 30 May 2015

- Transparency International: The Anti-Corruption Plain Language Guide, July 2009. https://www.transparency.de/fileadmin/pdfs/Themen/Wirtschaft/TI_Plain_Language_Guide_280709.pdf (2009). Accessed 30 May 2015
- Transparency International: Corruption Perceptions Index 2012. Transparency International. <https://www.transparency.org/cpi2012/results> (2012). Accessed 30 May 2015
- Transparency International: Global Corruption Barometer 2013. Transparency International (2013). ISBN: 978-3-943497-36-6 2013. http://www.transparency.org/gcb2013/in_detail (2013). Accessed 30 May 2015
- Transparency International: Corruption Perception Index 2013. Transparency International (2013a). ISBN: 978-3-943497-49-6. http://issuu.com/transparencyinternational/docs/cpi2013_brochure_single_pages?e=2496456/5813913 (2013). Accessed 30 May 2015
- Transparency International: Global Corruption Barometer 2013. Transparency International (2013b). ISBN: 978-3-943497-36-6 2013. http://www.transparency.org/gcb2013/in_detail (2013). Accessed 30 May 2015
- World Health Organization (WHO): The World Health Report: Health Systems Financing: The Path to Universal Coverage. WHO (2010). ISBN:978-92-4-156402-1