Far East Journal of Applied Mathematics
Volume 47, Number 1, 2010, Pages 35-49
Published Online: November 18, 2010
This paper is available online at http://pphmj.com/journals/fjam.htm

MULTI-DIMENSIONAL ANALYSIS ON ADVERTISING MARKET

MARCO FRANCIOSI and GIULIA MENCONI*

Dipartimento di Matematica Applicata "U. Dini"
Università di Pisa
Italy

Istituto Nazionale di Alta Matematica
Roma, Italy
e-mail: menconi@mail.dm.unipi.it

Abstract

Multimedia sequential data represent the behavior of multiple measurements on some process and may be analyzed as multi-dimensional time series via entropy and statistical linguistic techniques. We introduce three markers: influence area, consistency and diversification. The former two refer to the quality of the dynamic change of the data with time; the last one measures the variability of recurrent patterns. These markers are useful in classification or clustering of large databases, prediction of future behavior and attribution of new data. We show an application concerning different investment strategies in purchasing commercials in advertising market.

2010 Mathematics Subject Classification: 94A17, 62H30.

Keywords and phrases: time series analysis, entropy and other measures of information, econophysics, computational methods in statistical physics, nonlinear dynamics.

The work of G.M. is supported by a post-doc research scholarship "Compagnia di San Paolo" awarded by the Istituto Nazionale di Alta Matematica "F. Severi".

*Corresponding author

Received July 19, 2010

1. Introduction

This paper aims at introducing measures of complexity of multimedia sequential data, based on entropy analysis and statistical linguistic techniques. We also show an application in order to figure out leading features driving the dynamical change in such data.

By multimedia data (multi-dimensional time series, in the following), we mean a finite set

$$\mathbb{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}$$

of N data, where each data X_j is a finite array of real numbers coming from subsequent discrete measures of some empirical phenomena (e.g., weekly data): $X_j = (x_{j,1} \cdots x_{j,t})$.

Multi-dimensional time series appear when dealing with multiple measurements on some objects/phenomena, each one focused on some *a priori* structure of the process under study. Examples of such multimedia mining are given when considering different measurements (such as temperature, pressure, velocity) of the same physical phenomenon or taking different clinical data (such as pulse-rate, blood pressure, oxygen saturation, etc.) of one single patient (see for instance [1]). Other examples refer to the analysis of financial markets, where it is worth to look at different behaviors of a company.

A standard approach in studying time series consists in using a symbolic translation of the experimental outcomes (e.g., following [5]) and then applying any symbolic data analysis. In the framework of symbolic sequences, two branches of nonlinear dynamics may be of help in facing such multimedia sequences.

The first tool we use is based on an asymptotic measure of the density of the information content. In an experimental setting, information content may be approximated by means of compression algorithms (see for instance [2]). The notion of information content of a finite string can be used also to face the problem of giving a notion of randomness. Namely, this leads to the notion of *entropy* $h(\sigma)$ of a finite string σ, which is a number that yields a measurement of the complexity of σ

(see Subsection 2.1 for details). Intuitively, the greater the entropy of a string, the higher its randomness is, in the sense that it is poorly compressible.

The second tool is given by statistical linguistic techniques such as the Zipf scaling law, which offers a nice methodology that can be applied in order to characterize specific aggregation patterns or to identify different “languages”. The so-called Zipf analysis [8] is useful to understand how variable the distribution of patterns is within a symbol sequence. The basic idea is that the more variable the observed sequences are, the more variable the measurements and the more complex the obtained language are.

Entropy and Zipf analysis may be performed towards each array of a given multi-dimensional time series $\mathbb{X}$, allowing a global perspective of $\mathbb{X}$ to be achieved. This analysis is particularly useful when arrays X_j are pairwise incomparable (e.g., they represent different physical measures of some phenomenon) or if the values acquired in different series are different in magnitude (i.e., given X_j and X_h, it is $x_{j,i} \gg x_{h,i}$ for every *i*).

As a preliminary application of the above methodology, we study here some economic data coming from advertising market. In our example, we shall deal with the measurements of money investments of some business companies when purchasing commercials on different media sources (radio, newspapers, magazines, TV, free press, cinema, etc.). Our analysis results in a classification of companies as more/less dynamic in investment strategies, which is different from a mere classification based on the cumulative invested money. This may also suggest new marketing lines, focused on a customized strategy of investment offers from the viewpoint of commercials’ sellers.

1.1. Notations

Throughout this paper, we shall use the following notations for time series:

- X : one-dimensional time series
- $\mathbb{X}$: multi-dimensional time series
- σ, S: finite symbolic sequence
- $\mathbb{S}$: multi-dimensional symbolic sequence

2. One-dimensional Symbolic Data

Consider a one-dimensional t-step long time series $X = (x_1 x_2 \cdots x_t)$. We first translate X into a finite symbol sequence S. This, for instance, may be done either by means of some method of symbolic representation (such as SAX [5]) or by selecting as 'event 1' only the values exceeding some threshold and 'event 0', otherwise. Anyway, we obtain a sequence S which is the symbolic translation of series X.

The symbolic sequence S may be considered as a phrase written in some language. The more variable the observed data, the more complex the obtained language is. Entropy is a way to characterize the way the phrase S is built, while Zipf analysis refers to the typical recurrent words.

2.1. Entropy

One of the most significant tools from the modern theory of nonlinear dynamics used to analyze symbolic time series is related to the notion of *information content* of a finite sequence as introduced by Shannon in [7]. The intuitive notion of information content of a finite word can be stated as "the length of the shortest message from which it is possible to reconstruct the original word". The method we use to study the information content of a finite sequence is related to *compression algorithms*. The compression of a finite sequence reflects the intuitive meaning of its information content.

Let $\sigma = (s_1 s_2 \cdots s_t)$ be a t-long sequence written in the finite alphabet $\mathcal{A}$. Let $\mathcal{A}^t$ be the set of t-long sequences written using $\mathcal{A}$ and let the space of finite sequences be denoted by $\mathcal{A}^* := \bigcup_t \mathcal{A}^t$. A *compression algorithm* on a sequence space is any injective function $Z : \mathcal{A}^* \to \{0, 1\}^*$, that is, a binary coding of the finite sequences written on $\mathcal{A}$. The information content of a word σ w.r.t. Z is the binary length of $Z(\sigma)$, the compressed version of σ. Hence we define $I(\sigma) \doteq |Z(\sigma)|$, the information content of σ.

The notion of information content of a finite string may be used also to face the problem of giving a notion of randomness. Namely, we can think a string to be more random as less efficient is the compression achieved by a compression algorithm. This leads to the notion of *entropy* $h(\sigma)$ *of a finite string*, defined as the compression ratio (i.e., the information content per unit length):

$$h(\sigma) \doteq \text{Entropy of } \sigma = \frac{I(\sigma)}{|\sigma|} = \frac{|Z(\sigma)|}{t}.$$

It holds that $0 < h(\sigma) \leq 1$ and moreover, the greater the entropy of a string, the higher its randomness is, in the sense that it is poorly compressible.

Finally, a crucial remark on data pre-processing. Frequently, an experimental (one-dimensional) time series $Y = (y_1, ..., y_{t+1})$ is short and no longer prolongable. Moreover, (even if Y is sufficiently long), it is usually far from being stationary. Even in the case of finite sequence σ, the property of being stationary allows a proper connection of the entropy of σ to the theoretical results on nonlinear dynamical systems. A standard way to make it close to be stationary is to consider the difference series $Dif(Y) = (d_1, ..., d_t)$, where $d_j \doteq y_{j+1} - y_j$ and to apply the symbolic analysis to that $Dif(Y)$. In the infinite-length case, the entropy of Y and $Dif(Y)$ coincide and this motivates the use of $Dif(Y)$ also in the finite case.

2.2. Linguistic analysis

A time series X may be read as a sequence of measurements governed by some dynamic rules driving the time change in the measured values. Notwithstanding the entropy measures the rate of variability in the series, other crucial hints about the series may come from statistical analysis of the patterns described by the series, as words in a language generating the symbolic string associated to the series. The so-called Zipf analysis [8] is useful to understand how variable the distribution of patterns within a symbol sequence is.

Given a finite symbol sequence σ of length t, let us fix a word size $p < t$ and let us consider the frequency of all words of length p within σ. Let us order such words according to their decreasing frequencies. This way, each word has a rank $r \geq 1$. The Zipf scaling principle asserts that in natural languages, the frequency of occurrence $f(r)$ of word of rank r is s.t. $f(r) \sim r^{\rho}$, where $\rho \leq 0$. In an experimental setting, the value of Zipf coefficient ρ may be calculated via linear regression on the frequency/rank values in bilogarithmic scale. A low scaling coefficient is connected to high variability of words: were the words uniformly distributed, the scaling coefficient would be zero. Thus, the more variable the observed sequences are, the more complex the obtained language is, and the more variable the measurements are. This rank-ordering statistics of extreme events, originally created to study natural and artificial languages, had interesting applications in a great variety of domains, from biology [6], to computer science [3], to signal processing [4].

3. Multi-dimensional Data

In this section, we show how to extend the above mentioned tools to multi-dimensional time series. We may assume that the one-dimensional series have comparable length. We do not require them to have the same length t, but we require that each length of $t_1, \ldots, t_N$ is of the same order t and all the measurements refer to the same time lag of observation of the phenomenon. This discrepancy may be overcome by adding null values when there is lack of measurements, when this does not affect the sense of the analysis.

Given an alphabet size L, we associate to each multi-dimensional time series $\mathbb{X} = (X_1, \ldots, X_N)^T$ a multi-dimensional symbolic sequence $\mathbb{S} \doteq (S_1, \ldots, S_N)^T$, where S_j is the symbolic sequence associated to the one-dimensional sequence X_j.

3.1. Influence area

Given a multi-dimensional time series $\mathbb{X}$ and its symbolic translation $\mathbb{S} = (S_1, \ldots, S_N)^T$, we can compute the entropy of each component and obtain the entropy vector:

$$H(\mathbb{X}) = \begin{pmatrix} h(S_1) \\ \vdots \\ h(S_N) \end{pmatrix}. \tag{1}$$

Notice that the vector $H(\mathbb{X})$ yields a simple way to characterize the behavior of the series $\mathbb{X}$ and it is not uncommon to see that symbolic series associated to experimental measurements with different magnitudes may have almost the same entropy (for instance, take a series and create a new one just doubling the values of the first one, then in the symbolic model they are the same sequences).

Assume that the entropy vector is not null. We may investigate what the relative influence of each component is, by means of the following symplex analysis.

Choose some component, say the Nth.

We consider the $(N-1)$-dimensional symplex

$$\Delta^N = \left\{ \begin{pmatrix} y_1 \\ \vdots \\ y_{N-1} \end{pmatrix} : \sum_{i=1}^{N-1} y_i \leq 1 \text{ and } y_i \geq 0, \ \forall i \right\}$$

and the natural projection P of the vector $H(\mathbb{X})$ onto Δ^N, i.e.,

$$P = P(\mathbb{X}) = \begin{pmatrix} \frac{h(S_1)}{\| H(\mathbb{X}) \|_1} \\ \vdots \\ \frac{h(S_{N-1})}{\| H(\mathbb{X}) \|_1} \end{pmatrix}, \tag{2}$$

where $\| H(\mathbb{X}) \|_1 = \sum_{i=1}^{N} h(S_i)$ is the ℓ_1 norm of $H(\mathbb{X})$.

The position of the point P w.r.t. the vertices and the centroid G of Δ^N is a static feature of the process represented by $\mathbb{X}$, showing which one of the N components is leading the dynamics. Indeed, the vertex $V_N = (0, ..., 0)^T$ is associated to the Nth component whereas the vertices $V_1 = (1, 0, ..., 0)^T$, $V_2 = (0, 1, 0, ..., 0)^T$, ... and $V_{N-1} = (0, 0, ..., 1)^T$ correspond to the components labeled by $1, 2, ..., N-1$.

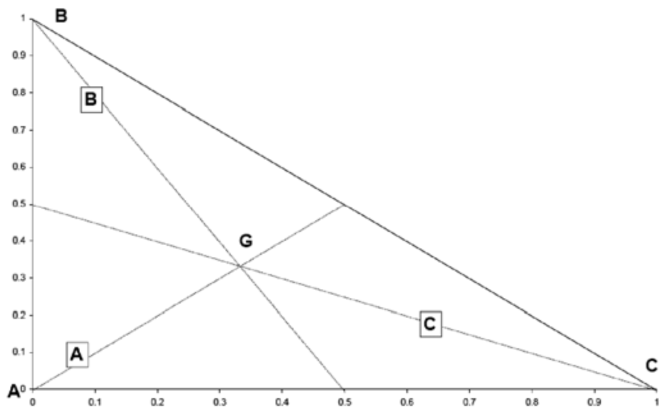

Figure 1. Symplex Δ^3 in $\mathbb{R}^2$ and influence areas of vertices A, B, C w.r.t. centroid G.

For each vertex V_j, consider the hyperplanes connecting $N-2$ other vertices to the centroid and not containing V_j. They partition the symplex Δ^N into N

regions, representing the *influence areas* of each vertex (see an example for Δ^3 in Figure 1). Therefore, if the influence area relative to point P is that of vertex V_d, then the dynamics of $\mathbb{X}$ is driven by the dth component (called *leading component*), in the sense that the dth entropy coefficient is prevailing on the others and the dynamic of that component is to be taken under observation more than the others'. We denote by $\mathcal{I}(\mathbb{X})$ both the leading component and the influence area of multi-dimensional time series $\mathbb{X}$.

3.2. Consistency

A mobile window entropy analysis allows extracting a dynamic feature showing the trend of entropy production of series $\mathbb{X}$.

Again, consider some multi-dimensional series $\mathbb{X} = (X_1, \ldots, X_N)^T$. Let t be the length of each component time series. Fix k be some positive integer. From each series S_j $(j = 1, \ldots, N)$ within $\mathbb{S}$, the symbolic model of $\mathbb{X}$, we may extract k subseries $W_1, \ldots, W_k$ in many ways: for instance, overlapping windows, non-overlapping windows, random starting points (fixed once for each collection $\mathcal{B}$ of multi-dimensional time series), etc. We only require all the k subseries having the same length; this implies that the choice of k should keep the length of the subseries sufficiently long for the entropy analysis to be meaningful. For each window, we calculate the entropy. We repeat the same for every series in $\mathbb{X}$. We obtain a matrix of entropy vectors in $[0, 1]^{N \times k}$, the moving vector of $\mathbb{X}$ denoted by $\mathcal{M}(\mathbb{X})$ whose rth row is $\mathcal{M}_r \doteq (h(W_{r,1}), \ldots, h(W_{r,k}))$.

As shown above, we may associate to series $\mathbb{X}$ a *sequence* of k points in the symplex Δ^N:

$$\mathcal{W} = (P_1, \ldots, P_{k-1}), \tag{3}$$

where P_1 is the point in the symplex corresponding to the entropy vector $(h(W_{1,1}), \ldots, h(W_{1,N}))^T$, the one relative to the first window, etc. We say that sequence $\mathcal{W}$ is the *entropy walk* relative to each multi-dimensional time series.

The entropy value is a marker of the dynamic change in the time series. The higher the entropy is, the higher is the variability of the series, therefore the more "impredictable" is the future of the series. The entropy walk is a way to look how the entropy changes with time within the sequence. Were the points almost colinear, the entropy change is balanced and the dynamic change is homogeneous; were the points more scattered, the dynamic rules changed and the process may need a finer observation.

We shall distinguish among series where recent measurements are consistent with past changes from series showing an abrupt change in dynamics in the last time steps.

We calculate $\mathcal{R}$, the linear regression of the first $k-1$ points defining the entropy walk $\mathcal{W}$ in Δ^N (all but the last one). Let $P_k \in \Delta^N$ be the point associated to the *k*th (last) window. We aim at understanding whether it comes from a dynamics in common with the one driving the past walk, that is, we aim at verifying how much dynamics the last window shares with previous $k-1$ windows.

There are many different ways to do it; we decided to apply the following criterium:

If the distance of P_k from the linear regression is not greater than the mean distance of points within the entropy walk, then we say that the point P_k is *consistent* with the walk $(\mathcal{C} = 1)$. Otherwise, it is not consistent $(\mathcal{C} = 0)$.

Each series is then characterized by the following two features:

- $\mathcal{I}$ is the influence area of the $(k-1)$th window
- $\mathcal{C}$ is the consistency binary value of the series

As a final remark, in case the consistency is null, we may apply a second order analysis and use the influence area of the last window as lighter marker of dynamic change: were it different from $\mathcal{I}$, then the process under examination is undergoing an abrupt change. In the case if the influence area of P_k coincides with the past one, then we may say that the change is still slightly acceptable.

3.3. Global linguistic analysis

For what concerns multi-dimensional time series, we recall that they are assumed to be short, therefore the statistics is quite poor. Nevertheless, what may be distinctive is the use they do of the distinct words. Moreover, we define a marker of pattern differentiation as follows. Fix once and for all a pattern size *p* which is sufficiently long w.r.t. the order of the series length *t*. Given a multi-dimensional series $\mathbb{X} = (X_1, \ldots, X_N)$, we calculate the Zipf coefficient for each component $(\rho_1, \ldots, \rho_N)$ and denote by $\mathcal{D}$ the *diversification*:

$$\mathcal{D}(\mathbb{X}) = 1 + \frac{1}{N}\sum_{j=1}^{N}\rho_j. \tag{4}$$

This way, the mean Zipf coefficient gives an estimate of the degree of differentiation in the use of most frequent patterns of length *p* within series in $\mathbb{X}$. For values of $\mathcal{D}$ close to 1, there is a high diversification of patterns that tend to be used indifferently, since their distribution is almost uniform. For values of $\mathcal{D}$ close to 0, the language of the *p*-patterns is rich and there exist some rules giving more importance to some patterns despite others, therefore the distribution of words is no longer balanced. If $\mathcal{D} < 0$, then the words are extremely unbalanced and typically there are a few words used recurring very frequently while most of the words are rarely used.

4. Application to Commercials' Market

We applied our method to three-dimensional time series related to 40 companies. The data come from Nielsen Media Research data base of weekly investments in purchasing commercials on three Italian media from 1996 to 2006. Each company is an automotive brand and each one-dimensional series has 585 non-negative data. The components are the money spent by each brand for purchasing commercials on radio, on magazines and on newspapers, respectively. The data were collected in the framework of a collaboration of Department of Applied Mathematics in Pisa with A. Manzoni and C.S.p.A. in Milan. The experimental application we are showing here is part of a joint work with Massimo Colombo, Guido Repaci and Giovanni Sanfilippo.

The original series were pre-processed in order to make them more stationary; consequently, we worked on the difference series, as explained in Subsection 2.1. We applied a symbolic filter using SAX procedure [5] with alphabet size $L = 4$. We used symbol '*a*' in case of large decrement; '*b*' in slight decrement; '*c*' in slight increment; '*d*' in huge increment.

The entropy was calculated using a Lempel-Ziv based algorithm introduced by Benci et al. in [2]. We recall that any optimal compression algorithm (i.e., on almost every infinite sequences, the entropy of the source is reached) may be used.

The series are three-dimensional, therefore the symplex we use is $\Delta^3 \subset \mathbb{R}^2$, where the vertices are *A* (relative to radio component), *B* (relative to magazine component) and *C* (relative to newspaper component).

The window analysis was exploited over the period 1996-2006 using 5 windows approximately 7-years long (350 measurements) and overlapping for 6 years (first year out, new year in). The measurements concerning year 2006 were used to build final window.

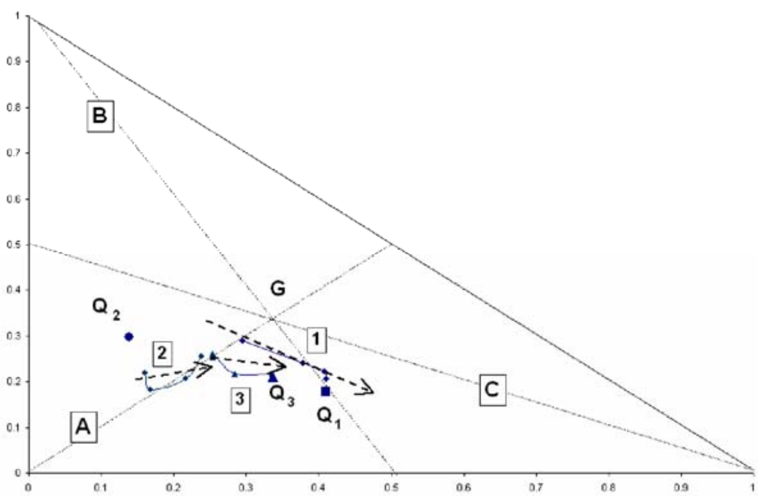

Figure 2. Symplex Δ^3 in $\mathbb{R}^2$ for the multi-dimensional series of 42 brands: entropy walk and trend. Example for three brands: b_1 (plotted with □), b_2 (plotted with ○) and b_3 (plotted with Δ) (see text). Vertex *A* is relative to radio component, *B* is relative to magazine component and *C* is relative to newspaper component.

In Figure 2, only three brands b_1, b_2 and b_3 have been considered to exemplify the entropy analysis. The entropy walks are plotted by solid lines and the tendency of each brand are arrows. Brand b_1 is plotted by symbol □, b_2 by symbol ○ and b_3 by symbol Δ. Three new points Q_1, Q_2 and Q_3 represent the position in Δ^3 of the final subseries used to calculate the consistency of the brand. We deduce that Q_1 is not consistent and its influence area also changed, while Q_2 is still not consistent, but within the same influence area. Finally, Q_3 is consistent.

Now, let us expose the results of the linguistic analysis on such a collection of sequential data. The diversification shall be paired to the above dynamical features to draw a finer picture of the data.

We exploited Zipf analysis on the symbolic series built on the alphabet $\{a, b, c, d\}$ as explained at the beginning of this section. We analyzed the frequency of words of length $p = 12$ (corresponding to 3 subsequent monthly investments) modulo permutations of the four symbols. That is, any two words of length $p = 12$ are equivalent if they have the same content in symbols *a*, *b*, *c* and *d*. They were identified by the 4-tuple (n_a, n_b, n_c, n_d). This choice of length 12 is motivated by the specific context where the multi-dimensional series come from: such words of length 12 represent what type of investments occurred over three months, without paying attention to their exact chronological order. Notice that, due to the short length of the series, the number of different 12-words used by the 42 brands range from 2 to around 60 (over 2148 possible). We found that many words were *rare*, that is, they occurred with frequency lower than 1%; therefore, we decided to calculate Zipf coefficient only for non-rare words. Of course, a finer analysis should also include specific words which have been used more frequently, but this is not what this example is devoted to.

To sum up, the influence areas are either radio (36 brands) or newspapers (4 brands). No brands show magazines as influence area. There are 65% consistent brands (26 over 40), while of the remaining 14 not consistent brands, only 2 changed influence area. Finally, concerning the three categories of diversification (as in Subsection 3.3), the highly diversified brands $(10^{-1} < \mathcal{D} \leq 1)$ are 82.5% of the total, the rich brands (if $|\mathcal{D}| \leq 10^{-1}$) are 10% and totally unbalanced brands (when $\mathcal{D} < -10^{-1}$) are 7.5%.

We built a dendrogram resulting from the three indices we obtained by our analysis. We used hierarchical clustering on Euclidean distance. Each brand was identified by its influence area (discrete values, for the 3 media), consistency (binary value) and diversification (real-valued). The brands are classified into eight groups of investment strategies: from I to III, there are brands with the most vivacious strategies, while from VI to VIII, there are the most static and repeated strategies. Exceptions are singletons IV (brand $b = 31$) and V (brand $b = 37$), representing two luxury brands.

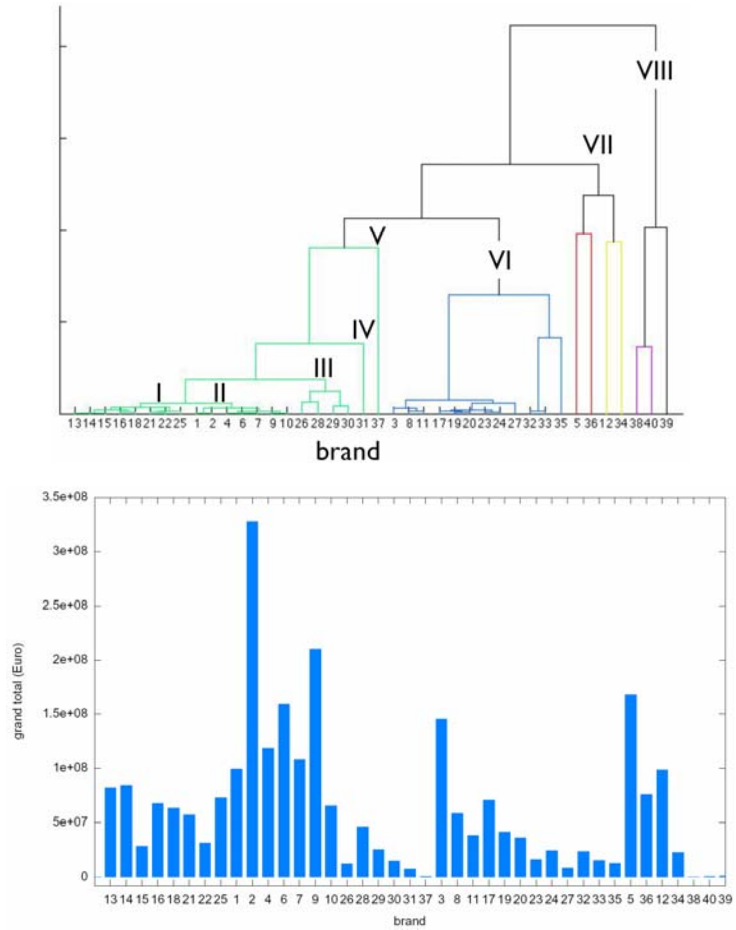

Figure 3. Comparison between the dendrogram for influence area, consistency and diversification (top) and the grand total of money that have been invested on the three media (bottom) for the 40 brands (see text).

Final Figure 3 shows the comparison between such dendrogram (top) and the grand total of money that have been invested on the three media (bottom). It is straightforward that, the more money is available for buying commercials, the more diversified and vivacious the strategy may be, but notice that class I is mainly constituted by the majority of medium-valued investments. The top investor (brand $b = 2$ in outmost class II) is an Italian brand which underwent a heavy renewal in marketing strategies in the last decade. An analogous situation holds for brands investing in a more self-confirming way: classes from VI to VIII contain brands with lower investments, with the two exceptions (brand $b = 3$ in outmost class VI and brand $b = 5$ in outmost class VI) of two extra-European brands which are worldwide leaders for city cars.

5. Conclusion

In this paper, we define a technique of feature extraction for multi-dimensional data via entropy and statistical linguistic techniques.

Given some process on which *N* different measures have been exploited over some time lag, we obtain an *N*-dimensional time series $\mathbb{X} = (X_1, ..., X_N)^T$. After translating the experimental data into symbolic multi-dimensional data, the analysis allows the following features to be associated to $\mathbb{X}$:

Influence area of the data $\mathcal{I}(\mathbb{X})$

It refers to the one-dimensional component $X_{\mathcal{I}}$ in $\mathbb{X}$ whose dynamic is driving the evolution of the overall *N*-dimensional phenomenon.

- Consistency $\mathcal{C}(\mathbb{X})$ w.r.t. *k*-window analysis

It is a binary attribute on whether the dynamics has changed in recent time.

- Diversification $\mathcal{D}(\mathbb{X})$

It formalizes the differentiation in the use of recurrent patterns.

The paper shows an application to economic data. We deal with measurements of money investments of some business companies in advertising market for different media sources. Such example shows that this methodology may offer new perspectives of interpreting the way brands invest their money: it is not completely true that “money leads strategy”, at least in this case study.

References

[1] F. Altiparmak, H. Ferhatosmanoglu, S. Erdal and D. C. Trost, Information mining over heterogeneous and high-dimensional time-series data in clinical trials databases, IEEE Trans. Knowledge Data Engng. 10 (2006), 254-263.

[2] V. Benci, C. Bonanno, S. Galatolo, G. Menconi and M. Virgilio, Dynamical systems and computable information, Discr. Contin. Dynam. Syst. - B 4(4) (2004), 935-960.

[3] M. E. Crovella and A. Bestavros, Self similarity in world wide web traffic: evidence and possible causes, IEEE/ACM Trans. Networking 5(6) (1997), 835-846.

[4] E. Dellandrea, P. Makris and N. Vincent, Zipf analysis of audio signals, Fractals 12(1) (2004), 73-85.

[5] J. Lin, E. Keogh, S. Lonardi and B. Chiu, A symbolic representation of time series, with implications for streaming algorithms, Proceedings ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2003.

[6] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons and H. E. Stanley, Linguistic features of noncoding DNA, Phys. Rev. Lett. 73(23) (1994), 3169-3172.

[7] C. E. Shannon, The mathematical theory of communication, Bell System Tech. J. 27 (1948), 379-423; 623-656.

[8] G. K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, 1949.