Analysis of the emission of American Depositary Receipts of Brazilian companies through the extraction of linguistic summaries

Amal Oudni, Marie-Jeanne Lesot, Maria Rifqi, Rosangela Ballini

Analysis of the emission of American Depositary Receipts of Brazilian companies through the extraction of linguistic summaries

Amal Oudni[1,2] Marie-Jeanne Lesot[1,2]
[1]Sorbonne Universités, UPMC Univ Paris 06
UMR 7606 - LIP6, F-75005, Paris, France.
[2]CNRS, UMR 7606, LIP6, F-75005, Paris, France.
Email: {amal.oudni, marie-jeanne.lesot}@lip6.fr

Maria Rifqi
Université Panthéon-Assas - Paris 2
LEMMA, F-75005, Paris, France
Email: maria.rifqi@u-paris2.fr

Rosangela Ballini
Institute of Economics
University of Campinas
13083-857, Campinas, Brazil
Email: ballini@eco.unicamp.br

***Abstract*—The cross-listing mechanism enables that companies collect funds and investors invest in capital markets of foreign countries. Among the objectives are the increase in the liquidity, the reduction of the risk and of the capital cost. In this context, this paper analyses the relationship of dually listed stocks of Brazilian companies, simultaneously traded on the São Paulo stock Exchange and New York Exchange, through American Depositary Receipts (ADR). In this sense, we evaluate which of the markets has the greatest influence on the pricing of those assets. For this purpose, we extract knowledge in the form of linguistic summaries representing attribute co-variations, enriched by different types of additional information which characterize the context and describe the co-variation type.**

I. Introduction

A financial asset such as stock is often traded in multiple markets. Its price in any given market is discovered by information being gathered and interpreted in one or more of these markets. A well known and widely traded type of stocks in foreign financial markets is the American Depositary Receipt (ADR), which is traded in the U.S. financial market. In this paper, we propose to investigate the price discovery process in the foreign exchange market through the use of data mining tools with high interpretability. We more precisely consider the relationship of dually listed stocks of Brazilian companies traded on Brazilian Securities, Commodities and Futures Exchange (BM&FBOVESPA[1]) and on New York Stock Exchange (NYSE).

A large variety of data mining tools can be considered for this goal. In this paper, we propose to exploit linguistic summaries based on gradual itemsets: first, they are appropriate to manage numerical data, second they provide a simple and understandable representation of information that summarizes and globally characterizes data.

Initially, the linguistic summaries have been introduced in a fuzzy variant by [29] and then further developed and presented in a computational form by [16], [15], using Zadeh's calculus of linguistically quantified propositions [30]. They are generally defined as texts made up of a few sentences in natural language [29], [16], [15]. They can extract different types of information, as for instance co-occurrences in the case of association rules [2], sequential relations in the case of sequential patterns [3] or gradual tendencies between the attributes of the data in the case of gradual itemsets [12], [7]. In this paper, we focus on the latter and their enrichments: we consider linguistic summaries based on gradual itemsets, i.e. of the form "*the more/less A, the more/less B*" where A and B are attributes. They summarize data through the description of their internal tendencies, identified as correlation between attribute values.

Two types of enriched gradual itemsets are considered in the paper. The first type, called characterized gradual itemset, linguistically introduced by the expression "*especially if*" [21]. It can be exemplified with "the more the US closing price increases, the more the Brazilian closing price increases, *especially if* it belongs to the interval $[1, 10]$". The second one is defined as accelerated gradual itemsets [22]. It can be illustrated by an example such as "the more the US closing price increases, the more *quickly* the Brazilian closing price increases", where "*quickly*" provides additional knowledge about the relation between considered attributes.

In this paper, we propose to extend the methods for gradual itemset extraction to the processing of time series, discussing the specific role of the temporal attributes.

The paper is organized as follows. Section II describes the application domain in more details. Section III formally defines gradual itemsets and the two types of considered enriched variants, by describing their principle, their interpretation and the criteria proposed for their evaluation. Section IV presents the real data to which the proposed approach is applied. Section V details and discusses the obtained empirical results describing U.S. and Brazilian financial markets. Finally, Section VI concludes the paper and proposes some future studies.

II. Analysis of the Brazilian Financial Market and Study of its Relationship with the U.S. Market

Many companies list their shares not only on their domestic exchange but also on foreign exchanges. The first sorting is typified by the reinvestment of profits in operational activity

[1]The BM&FBOVESPA (in Portuguese, Bolsa de Valores, Mercadorias & Futuros de São Paulo) is a stock exchange located at São Paulo, Brazil, created by the merger of the São Paulo Stock Exchange (Bovespa) and the Brazilian Mercantile and Futures Exchange (BM&F).

and turnover credit. The second category involves the raising of funds from outside the corporation through loans from financial institutions, the issuance of securities (bonds and commercial papers), and initial public offerings on stock exchanges. This last option includes the possibility for domestic companies to represent their stocks in foreign financial markets, which are called Depositary Receipts (DR) [25].

The pioneering study in this area was realized by [9], who evaluated the integration among stocks traded on different U.S. exchanges. Recent studies have been conducted in light of the increased globalization and integration of markets, as well as the predominance of electronic bids on the stock exchanges and the increasing use of robot traders, like the studies conducted between the ADRs traded on the NYSE and the stocks traded on the stock exchanges of Canada [10], [18], [17], Israel [24], Argentina, Egypt and Spain [23], [1] and China [5]. It is also worth noting works that addressed other dominant satellite markets, such as, with evaluation of the stocks listed on exchanges in Hong Kong and London [28], the stock markets of Taiwan [27], and for stocks traded on stock exchanges in Australia and New Zealand [20]. In general, the results indicate integration between markets, and that price discovery is predominantly performed in the local (stock) market. For the Brazilian market, [25] analyzed the price discovery process of cross-listed stocks of Brazilian companies simultaneously traded on the BM&FBOVESPA and NYSE stock exchanges. In short, the results indicated the existence of co-integration for most stock-ADR pairs and they concluded the existence of arbitrage opportunities between markets.

A well known and widely traded type of DR is the American Depositary Receipt (ADR), traded exclusively in the U.S. financial market. According to [13], the U.S. market accounts for 80% of the total value of DRs traded in the world, and of this percentage approximately 83% are negotiated on the NYSE. Of the six ADRs with the highest turnover on the NYSE in 2013, three are issued by Brazilian companies. These ADRs constitute depositary receipts issued by a U.S. bank, which are traded in the financial market of that country, with a guarantee in shares of Brazilian corporations [25].

For the companies, the issuance of these stocks is a means of capitalization, in the case of ADR backed by new issuances, as well as a means to increase the liquidity of the security. Furthermore, it is argued that the issuance of ADRs increases the transparency of a company, which concomitantly decreases the risk perception of the company by financial agents, and consequently reduces the cost of capital. From the perspective of U.S. investors, by conferring economic benefits equal to a share issued in the home market of the corporation, the ADRs enable the inclusion of securities from other countries in their investment portfolios, which allows them to avoid carrying out international operations to diversify their portfolios [14].

Given the existence of stocks cross-listed on stock exchanges in different countries, issues relating to the integration of these markets and the possibility of arbitrage in such securities have been the subject of research in finance since the late 1970s. Arbitrage can make two or more non-stationary economic time series to "move together" over time and have a long-run relationship.

In this context, the aim of this paper is to analyze the relationship of dually listed stocks of Brazilian companies, simultaneously traded on the BM&FBOVESPA and NYSE. In this sense, we evaluate which one of the markets has the greatest influence on the pricing of those assets. For this purpose, we extract knowledge in the form of linguistic summaries representing attribute co-variations, enriched by different types of additional information.

III. Proposed Approach

Among the variety of possible data mining approaches, we propose to use a method based on gradual itemsets for two reasons: they adapt to numerical data and are able to summarize and to reduce data to amounts of information understandable and interpretable, thus facilitating the understanding of their content by the expert.

This section presents the different forms of linguistic summaries we extract from the data, detailing for each of them its principle and its extraction approach: after recalling the principles of gradual itemsets as well as their quality criterion in Section III-A, we describe the two considered enrichment types, characterization in Section III-B and acceleration in Section III-C. Section IV presents the proposed application to time series.

A. Gradual Itemsets

1) Notation and definitions: To formally define gradual itemsets one needs first to define gradual items. Let $\mathcal{D}$ denote the data set constituted of objects described by a set of numerical attributes. A *gradual item* is defined as a pair of an attribute A and a variation $* \in \{\geq, \leq\}$ which represents a comparison operator: $A^{\geq}$ and $A^{\leq}$ represent the fact that the attribute values increase (in case of $\geq$) or decrease (in case of $\leq$).

A *gradual itemset* is then defined as a combination of several gradual items, semantically interpreted as their conjunction. For instance $I = A^{\geq}B^{\leq}$ is interpreted as the more A and the less B. It imposes a variation constraint on A and B attributes simultaneously.

It is important to underline the difference between these gradual items and fuzzy gradual rules [8], [11]: the latter are a fuzzy generalization of association rules and express whether the fuzzy presence of an item implies the fuzzy presence of another item, using a fuzzy implication operators: for fuzzy gradual rules, each object has an individual contribution to the quality criterion, based on its membership degree to the considered fuzzy items. Gradual items as considered in this paper consider the data set as a whole and rank all objects with respect to the considered item, examining global co-variation across the data set.

Several interpretations of gradual itemsets have been proposed as e.g. regression [12], correlation of induced order [4], [19] or identification of compatible object subsets [6], [7]. Each interpretation is associated with the definition of a support to quantify the validity of gradual itemsets and with methods for identifying itemsets that are frequent according to the support definition.

In this paper, we consider the interpretation of covariation constraint by identification of compatible subsets [6], [7]. This

interpretation is based on the order induced by the attribute values, not on the values themselves: for an itemset I, we denote $\preceq_I$ the pre-order induced by I. It is formally defined as

$$o \preceq_I o' \text{ iff } \forall\ j = 1..k\ A_j(o) *_j A_j(o')$$

where $A_j(o)$ represents the value of attribute A_j for object o and k the associated length of I, defined as the number of attributes it involves.

The compatible subset approach consists in identifying subsets D of $\mathcal{D}$, called *paths* supporting I, that can be ordered so that all data pairs of D satisfy the pre-order induced by the considered itemset. More formally, for an itemset $I = \{(A_i, *_i), i = 1 \dots k\}$, $D = \{o_1, \dots, o_m\} \subseteq \mathcal{D}$ is a path if and only if there exists a permutation π such that $\forall l \in [1, m-1], o_{\pi_l} \preceq_I o_{\pi_{l+1}}$.

Such a path is called *complete* if no object can be added to it without violating the order constraint imposed by I. Denoting $\mathcal{L}(I)$ the set of complete paths associated to I, the set of maximal complete paths, i.e. complete paths of maximal length, is defined by

$$\mathcal{L}^*(I) = \{D \in \mathcal{L}(I) / \forall D' \in \mathcal{L}(I)\ |D| \geq |D'|\}$$

2) Quality Criterion: The gradual support of I, $GS_{\mathcal{D}}(I)$, is defined as the length of its maximal complete paths divided by the total number of objects [6]:

$$GS_{\mathcal{D}}(I) = \frac{1}{|\mathcal{D}|} \max_{D \in \mathcal{L}(I)} |D| \tag{1}$$

This definition of gradual support has been implemented for extracting gradual itemsets as the GRITE algorithm [7]: it constitutes an efficient method for extracting valid gradual itemsets according to this support definition, based on a binary representation: this representation leads to efficient bitwise operators. It also allows the automatic extraction of the paths satisfying these gradual itemsets.

B. Characterized Gradual Itemsets

This section describes the principle of characterized gradual itemsets [21] as well as their formalization, illustrating them with an example.

1) Principle of Characterized Gradual Itemsets: The objective of the characterization of a gradual itemset I is to identify a set of attributes $J \subset I$ and a region R, leading to the linguistic expression "especially if $J \in R$" where validity of I must increased. This region must maximize two criteria: the number of objects it contains and the validity of the gradual itemset. This region is called interval of interest.

Figure 1 represents a data set described with two attributes, for which the gradual itemset *the more A, the more B* is supported by the points represented by •. Now it can be observed that the covariation between A and B especially holds in the central part of the graph, whereas more noisy data occur for low A values and high A values: if the data are restricted to objects for which A takes values in the interval $[32; 53]$, graphically delimited by the vertical lines on Figure 1, the support of the itemset increases. This motivates

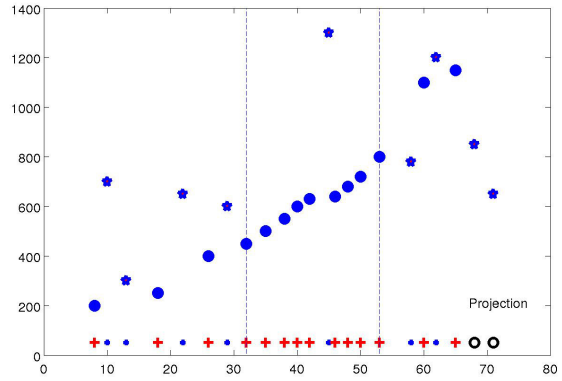

Fig. 1: Example of gradual itemset characterization, leading to the "the more A, the more B, especially if $A \in [32; 53]$".

the extraction of the characterized itemset *the more A, the more B, especially if* $A \in [32; 53]$.

The characterization of gradual itemsets is interpreted as an increased validity when the data are restricted to the objects satisfying the characterization clause.

Thus, the principle of characterized gradual itemsets is to find a trade-off between a high support and a high number of objects when restricting the data set to a subset.

2) Formalization: The principle illustrated above can be formalized as follows: for a gradual itemset I, a characterization is denoted as "I, *especially if* $J \in R$", where J is a set of attributes occurring in I and R is an associated set of intervals. The region R induces a restriction $\mathcal{D}'$ of the data set $\mathcal{D}$, considering only the data satisfying the value constraint expressed by R.

The previous principle consists in both maximizing the support of the considered itemset I on the restricted data and the number of objects satisfying the constraints, i.e.

$$\begin{cases} \max\limits_{R} |\mathcal{D}'| \\ \max\limits_{R} GS_{\mathcal{D}'}(I) \end{cases} \tag{2}$$

A trade-off must be found between these two objectives that can be contradictory: indeed, an increase of the size of the subset $\mathcal{D}'$ can lead to the decrease of the proportion of objects compatible with the order induced by the considered itemset.

3) Proposed Approach: We proposed to decompose the task of identifying relevant attributes and their associated intervals of interest by successively considering each attribute occurring in the considered gradual itemset I and further by successively considering each path supporting I, e.g. available when I is extracted by GRITE algorithm [7]: the computation of the restricted gradual support $GS_{\mathcal{D}'}(I)$ can be based on the restriction of these paths. We thus propose to consider the effect of candidate restriction for each path, later combining them to select the optimal bounds.

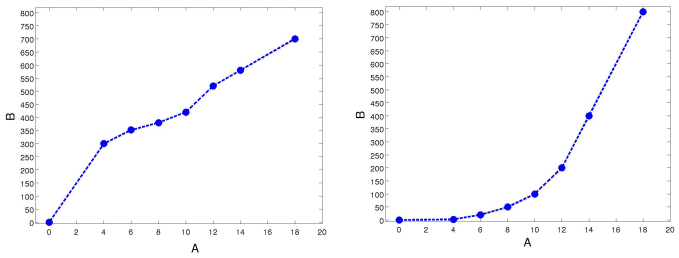

Fig. 2: Two data sets, leading to "*the more A, the more B*" where an acceleration effect is observed for the right data set and not for the left one.

To that aim, we proposed an approach based on mathematical morphology tools to automatically extract these characterized gradual itemsets [21]. Given a gradual itemset I, a maximal math D and an attribute A for which an interval of interest is searched, the information of the path is coded through a transcription process in a sequence of symbols $\{+, -\}$, where $+$ represents an object satisfying I and $-$ one which does not verify it. Thereby, we obtain at the end a succession of symbols of $+$ and $-$. One looks for the region where I is truer, which is equivalent to looking for the longest sequence of $+$. In the proposed approach, this sequence of $+$ is extended, by incorporating some $-$ symbols, so as to increase the size of the restricted data set represented by the longest sequence of $+$ without deteriorating the proportion of $+$ in the considered sequence. This is performed by a morphological filter, applied to the succession of symbols obtained after transcription of the data, as justified and discussed in [21].

The lower part of Figure 1 indicates the transcription result for the illustrative example. More details about the transcription process and morphological step are given in [21].

The quality of the characterized gradual itemsets is evaluated by the characterized gradual support, $GS_{\mathcal{D}'}$, computed in Equation 2.

C. Accelerated Gradual Itemsets

Accelerated gradual itemsets aim at extracting information in the form of correlations between attribute values by focusing on how fast the values of some attributes vary as compared to others [22]. This section briefly recalls their principle, formalization and the evaluation criterion of the acceleration effect.

1) Principle: Figure 2 illustrates the principle of accelerated gradual itemsets: for both the left and right cases, a co-variation constraint is satisfied, that justifies the extraction of the gradual itemset "*the more A, the more B*". However, on the right-hand example, the speed of B augmentation appears to increase, making it possible to enrich the gradual itemset to "*the more A increases, the more quickly B increases*".

The principle of acceleration is naturally understood as speed variation increase, which can be translated as a convexity constraint on the underlying function associating the considered attributes. This constraint can be modeled as an additional co-variation constraint, leading to the definition of a criterion called *accelerated support* to assess the validity of such accelerated gradual itemsets [22]. It is added to the evaluation of the gradual itemset through the classical gradual support defined in Equation 1.

2) Formalization: An accelerated gradual itemset is formalized as a triplet: $A^{*_1}B^{*_2}\ \left(\frac{\Delta B}{\Delta A}\right)^{*_3}$, where $A^{*_1}B^{*_2}$ represents a gradual itemset, and $\left(\frac{\Delta B}{\Delta A}\right)^{*_3}$ the acceleration clause that compares the variations of B with that of A. $*_1$ determines whether "the more A increases" ($*_1 = \geq$) or "the more A decreases" ($*_1 = \leq$). $*_2$ plays the same role for B. $*_3$ determines whether acceleration or deceleration is considered: $*_3 = \geq$ leads to "the more quickly" and $*_3 = \leq$ leads to "the less quickly" or equivalently "the more slowly" .

We focus on the acceleration effect, i.e. attributes for which values increase "quickly", i.e. the case $*_3 = \geq$. It corresponds to the convex curve case.

3) Evaluation Criterion of the Acceleration Effect: The quality of accelerated gradual itemsets is measured both by the classical gradual support as recalled in Equation (1) and an accelerated gradual support that measures the quality of the acceleration, as defined below.

The itemset I induces a pre-order on objects as defined in Section III-A; the acceleration clause $I_a = \left(\frac{\Delta B}{\Delta A}\right)^{*_3}$ induces a pre-order on pairs of objects denoted $\preceq_{I_a}$: for any o_1, o_2, o_3 and o_4

$$\begin{array}{c}(o_1, o_2) \preceq_{I_a} (o_3, o_4) \\ \Leftrightarrow \\ \dfrac{B(o_2) - B(o_1)}{A(o_2) - A(o_1)} *_3 \dfrac{B(o_4) - B(o_3)}{A(o_4) - A(o_3)}\end{array} \tag{3}$$

where $A(o)$ and $B(o)$ respectively represent the value of attributes A and B for object o.

The quality of the candidate accelerated gradual itemset II_a is then high if there exists a subset of data that simultaneously satisfies the order induced by I and that induced by I_a. Therefore, the acceleration quality first requires to identify a data subset that satisfies $\preceq_I$. To that goal, GRITE can be used to identify candidate gradual itemsets as well as their set of maximal complete support paths $\mathcal{L}^*(I)$.

For any $D \in \mathcal{L}^*(I)$, the computation of the accelerated support then consists in identifying subsets of D that simultaneously verify the constraint $\left(\frac{\Delta B}{\Delta A}\right)^{*_3}$.

We denote φ the function that identifies a maximal subset of objects from D such that $\forall o_1, o_2, o_3 \in \varphi(D)$, $(o_1 \preceq_I o_2 \preceq_I o_3 \Rightarrow (o_1, o_2) \preceq_{I_a} (o_2, o_3))$

The accelerated gradual support of II_a is then computed as:

$$GS_a = \frac{1}{|D| - 1} \max_{D \in \mathcal{L}^*(I)} |\varphi(D)| \tag{4}$$

where $|D|$ denotes the size of any maximal complete path in $\mathcal{L}^*(I)$, as, by definition of $\mathcal{L}^*(I)$, they all have the same size. $|D| - 1$ is then the maximal possible value of $\varphi(D)$ and thus the normalizing factor.

D. Proposed Extension to Time Series

This section describes the proposed extension to extract gradual itemsets from time series: each data point is not a vector of p numerical attribute $x = (A_j)_{j=1...p}$, but a vector of the attribute values at each date, which can be denoted $x = (t, A_1(t), A_p(t))$ for any $t \in [0, T]$.

Two types of gradual items can be proposed. The first one considers time as an attribute, leading to items of the form $I = \{(t, \leq)(A_{*i}, *i), i = 1...k\}$. A specific linguistic form is then proposed. Indeed, the expression "the more the time increases" is not relevant. Therefore we propose, to replace the linguistic summaries by the form "A_i tends to rise" (if $*i = \geq$) and "A_i tends to decrease" (if $*i = \leq$). The example "the risk tends to rise, especially if the closing prices vary in short time belongs to the interval $[1, 10]$ days" illustrate this case.

A second type does not consider t as an item, but considers a data point for each date independently: it applies classical gradual itemsets to the vector $(A_1(t), A_p(t)) \ \forall \ t \in [0, T]$ and all data x. The example "the more the risk in Brazil market increases, the more the potential return of ADR is expected, especially if risk belongs to $[0.0002; 0.0008]$" illustrate such a case.

IV. Application to Financial Data

A. Considered Data

In this paper, we consider the price time series of two Brazilian companies, Embraer[2] and Cemig[3], that currently have ADRs on the NYSE as well as stocks traded on the BM&FBOVESPA. The Embraer company has become one of the largest aircraft manufacturers in the world by focusing on specific market segments with high growth potential in commercial, defense and executive aviation. The Cemig company operations comprise the areas of electric energy generation, transmission, distribution and commercialization, as well as the distribution of natural gas, telecommunications and the efficient use of energy.

The data were selected from the observation of operations in 2013[4]. Each company has a series of stocks prices and ADR prices. Thus, there were a total of 4 series of daily closing prices in U.S. dollars and adjusted for the proceeds received[5]. The calendar of the home market was adjusted to the New York market; on days when there was no trading on one market, but there was trading on the other, prices were interpolated.

The Embraer data are over the period from June 5, 2006 through July, 5 2013, which corresponds to 1705 observations; the Cemig price series cover the period from September 4, 1997 through July 5, 2013, which corresponds to 3805 data points.

[2]Empresa Brasileira de Aeronáutica – Embraer

[3]Companhia Energética de Minas Gerais – Cemig

[4]Data were collected in the Economatica software.

[5]The prices of shares traded on the BM&FBOVESPA in Real (R$) have been converted into U.S. dollars (US$) at the exchange rate of the day, as announced by the Central Bank of Brazil.

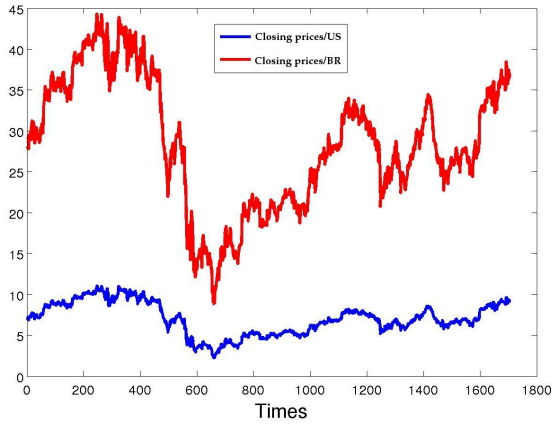

Fig. 3: Closing price time series from U.S. and Brazil – Embraer Company.

B. Considered Attributes

The observations are described by a date attribute and $p = 4$ numerical attributes as detailed in Table I.

In this paper, the volatility (risk), denoted h_t, was estimated by an univariate GARCH(1,1) process given by:

$$R_t = a + \sqrt{h_t \epsilon_t}$$

$$h_t = \alpha_0 + \alpha_1 R_{t-1}^2 + \beta_1 h_{t-1}$$

where $R_t = \ln(P_t/P_{t-1})$ is the daily logarithm returns series of price P_t of an asset at time index t; a is a constant; h_t represents the conditional variance of R_t; ϵ_t is error term with mean 0 and variance 1.0; $\alpha_0 > 0; \alpha_1 \geq 0, \beta_1 \geq 0$, and $(\alpha_1 + \beta_1) < 1$.

Another attribute obtained from volatility is the annual volatility σ which is computed by:

$$\sigma = \frac{\sqrt{h_t}}{\sqrt{T}}$$

where $T = 1/252$ because there are 252 trading days in any given year.

C. Illustrative Examples

Figure 3 shows the evolution of attribute prices in the U.S. and of attribute prices in Brazil markets to Embraer company, according to the time (Date) attribute. While the general pattern is similar, we can note that the stock prices in Brazil are higher than the prices in U.S. market.

To look for correlations between attribute values, we transform the data represented as time series to data that we can represent as a scatter plot, where each point represents the couple of price values for any given date, as shown in Figure 4. For this attribute pair, a clear increasing tendency can be observed, that can be described as "the more the US closing price increases, the more the Brazilian closing price increases". The aim is to enrich such tendencies with characterization and acceleration clauses.

TABLE I: Attribute description

Attribute	Description
Closing price	It is the final price at which a security is traded on a given trading day. Closing prices provide a useful marker for investors to assess changes in stock prices over time - the closing price of one day can be compared to the previous closing price in order to measure market sentiment for a given security over a trading day.
Returns	It is the gain or loss of a security in a particular period. The general rule is that the more risk you take, the greater the potential for higher return and loss.
Volatility/Risk	It is a measure of variation of the price of a financial instrument over time. The volatility refers to the amount of uncertainty or risk about the size of changes in a security's value. A higher volatility means that a security's value can potentially be spread out over a larger range of values.
Annual trading volume	It is the number of shares or contracts traded in a security or an entire market during a given period of time.

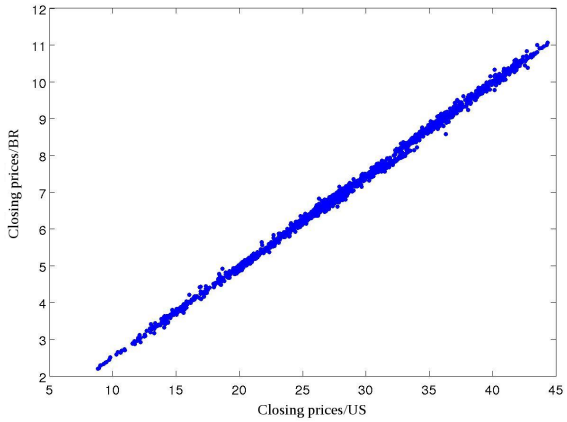

Fig. 4: Closing price data in the U.S. and Brazil as scatter plot – Embraer company.

V. Obtained Results

This section presents the obtained results, after describing the experimental protocol and successively focusing on characterized gradual itemsets and accelerated gradual itemsets.

A. Experimental Protocol

We collect gradual itemsets whose $GS_{\mathcal{D}}$, as defined in Equation (1), is higher than the threshold $s = 5\%$, characterized gradual itemsets whose $GS_{\mathcal{D}'}$ is higher than threshold $s_c = 40\%$ and accelerated gradual itemsets whose GS_a is higher than threshold $s_a = 20\%$.

It may be noted that setting optimal values for the thresholds of these criteria largely depends on the dynamics of the data, i.e. the change of variation values describing the time series.

B. Obtained Characterized Gradual Itemsets

We illustrate here some characterized gradual itemsets extracted from the Embraer dataset.

It is important to note that, when an itemset is characterized with several intervals, it means that each of the intervals is associated with the same support.

1) Gradual Itemsets with Support Almost 100%*:* It must first be underlined that among the results, two gradual itemsets have both gradual support an characterized gradual support almost 100%: They are characterized by the whole domain of the attributes that compose them:

- the more the risk increases, the more the annual volatility increases in Brazil, especially if the risk $\in [0.0002; 0.009]$:
  $GS_{\mathcal{D}} = 99.97\%$, $GS_{\mathcal{D}'} = 99.97\%$
- the more the risk increases, the more the annual volatility increases in USA, especially if the risk $\in [0.0018; 0.005]$:
  $GS_{\mathcal{D}} = 99.97\%$, $GS_{\mathcal{D}'} = 99.97\%$

These results are expected since one of the attribute is calculated from the other. This means that 100% of correlation between these attributes is established.

2) Gradual Itemsets with Non-temporal Attributes: In the case of gradual itemsets characterized by non-temporal attributes, we can cite the following example (the values of the risk in Brazil vary between 0.0002 and 0.0090).

- the more the risk in Brazil increases, the more the potential return of ADR is expected, especially if the risk belongs to $[0.0002; 0.0008]$:
  $GS_{\mathcal{D}} = 32\%$, $GS_{\mathcal{D}'} = 42\%$

This characterized gradual itemset has a medium quality: its characterized gradual support is equal to 42%, but it remains interesting because, on the one hand, its gradual support before characterization is significant (equal to 32%), and, on the other hand, it is characterized by a single interval, which means that about half of the data verifying this characterized gradual itemset is in this interval. The number of data this interval covers is 259.

Another characterized gradual itemsets are extracted with a high characterized gradual support, equal to 86% and significant gradual support (equals to 27.76% for the following example), but the size of their characteristic interval is lower, such as:

- the more the closing price in Brazil increases, the more the closing price in USA increases, especially if the closing price in USA belongs to $[8.83; 10.4]$ or $[17.4; 18.1]$: $GS_{\mathcal{D}} = 27.76\%$, $GS_{\mathcal{D}'} = 86\%$

The number of data they cover respectively equals 82 and 147. The values of the closing price in US vary between 8.8 and 44

3) Characterized Gradual Itemsets Involving Date: The gradual itemsets characterized with a date attribute using the data of this company are characterized by very short time intervals, but with high characterized gradual support, for example:

- the closing price of Brazilian stocks tends to rise, especially if the date belongs to [Dec-30-08; Jan-12-09] or [Nov-06-09; Nov-17-09] or [May-25-10; Jun-04-10]: $GS_{\mathcal{D}} = 7.28\%$, $GS_{\mathcal{D}'} = 89\%$

The date characterization of this characterized gradual itemset is precise, covering about 10 days over 7 years. However, they are highly relevant as they indeed allow to extract precise contexts. These date intervals correspond to the global financial crisis, when the evidence of the crisis were intense impacting growth of stock market index.

Other examples include:

- the annual trading volume in USA tends to rise, especially if the date belongs to [March-27-08; April-21-08]: $GS_{\mathcal{D}} = 5.69\%$, $GS_{\mathcal{D}'} = 86\%$
- the more the risk in USA increases, the more the annual trading volume in USA increases, especially if the date belongs to [Sep-24-08; Oct-03-08] or [Jan-21-09; Feb-16-09] or [May-27-09; Jun-12-09] or [Jul-06-09; Jul-16-09]: $GS_{\mathcal{D}} = 7.57\%$, $GS_{\mathcal{D}'} = 68\%$

C. Obtained Accelerated Gradual Itemsets

In this section, we show the results of accelerated gradual itemsets for the two companies: Cemig and Embaer.

Using the data available for the Cemig company and the thresholds s and s_a, 5 gradual itemsets are extracted as listed below:

- the more the risk in the Brazilian market increases, the more quickly the annual trading volume in Brazil increases: $GS_{\mathcal{D}} = 99.97\%$, $GS_a = 50\%$.
- the more the risk in the American market increases, the more quickly the annual trading volume in USA increases: $GS_{\mathcal{D}} = 99.97\%$, $GS_a = 53\%$.
- the more the risk in USA increases, the more quickly the returns from ADR increases: $GS_{\mathcal{D}} = 33\%$, $GS_a = 38\%$.
- the more the risk in Brazil increases, the more quickly the returns in the Brazilian market increase: $GS_{\mathcal{D}} = 18\%$, $GS_a = 69\%$.
- the more the closing prices in the Brazilian market increase, the more quickly the closing prices in the American market increase: $GS_{\mathcal{D}} = 21.24\%$, $GS_a = 74\%$.

Using the data available for the Embaer company and setting the gradual support threshold as $s = 5\%$, 5 gradual itemsets are extracted:

- the more the risk in the Brazilian market increases, the more quickly the annual trading volume in Brazil increases: $GS_{\mathcal{D}} = 99.94\%$, $GS_a = 88\%$.
- the more the risk in the American market increases, the more quickly the annual trading volume in USA increases: $GS_{\mathcal{D}} = 99.94\%$, $GS_a = 59\%$.
- the more the risk in the Brazilian market increases, the more quickly the returns from ADR increase: $GS_{\mathcal{D}} = 32\%$, $GS_a = 36\%$.
- the more the returns in the Brazilian market increase, the more quickly the risk in USA increases: $GS_{\mathcal{D}} = 16.3\%$, $GS_a = 67\%$.
- the more the closing prices in the Brazilian market increase, the more quickly the closing prices in the American market increase: $GS_{\mathcal{D}} = 27.76\%$, $GS_a = 64\%$.

It can be noted that gradual itemsets extracted from the Embraer and Cemig data are partially the same. This highlights a joint influence between the US and Brazilians markets. Similar results are also obtained but 6 other companies which are omitted here.

VI. Conclusion and Future Works

In this paper, we proposed to extend the data mining method that extracts gradual itemsets to time series so as to summarize their characteristics, considering an application to financial data. We used enriched gradual itemsets to extract more precise contexts and make the information extracted from these data understandable and easily interpretable. More precisely, we used characterized gradual itemsets and accelerated gradual itemsets, which respectively characterize the context and describe the covariation type.

The results highlight a joint influence between the USA and Brazilians markets. Actually, most enriched gradual itemsets obtained with the Embraer and Cemig data sets shown here are also obtained with other companies which we do not in this paper.

Future works include the definition of gradual itemsets apply to generalize the results and quantify the proportion of data: for example, “in most companies, the more the USA closing prices increase, the more quickly the Brazilian closing prices increase”.

Another prospect for this work focuses on the study of exceptions, i.e. itemsets which are not observed in the majority of the companies, but in one or few companies. This task is

related to the extraction of rare itemsets [26] for classic association rules and raises computational challenges motivated by the relevance of this type of extracted knowledge.

REFERENCES

[1] C. Aansotegui, A. Bassiouny, and E. Tooma. An investigation of intraday price discovery in cross-listed emerging market equities. *Investment Analysts Journal*, 77:55–67, 2013.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of the Int. Conf. on Very Large Data Sets*, pages 487–499, 1994.

[3] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proc. of the Int. Conf. on Data Engineering*, pages 3–14, 1995.

[4] F. Berzal, J. C. Cubero, D. Sanchez, M. A. Vila, and J. M. Serrano. An alternative approach to discover gradual dependencies. In *Fuzziness and Knowledge-Based Systems*, pages 559–570, 2007.

[5] K. C. Chen, G. Li, and L. Wu. Price discovery for segmented us-listed chinese stocks: Location or market quality? *Journal of Business Finance & Accounting*, 37(1–2):242–269, 2010.

[6] L. Di Jorio, A. Laurent, and M. Teisseire. Fast extraction of gradual association rules: a heuristic based method. In *Proc. of the Int. Conf. on Soft Computing as Transdisciplinary Science and Technology*, pages 205–210, 2008.

[7] L. Di Jorio, A. Laurent, and M. Teisseire. Mining frequent gradual itemsets from large databases. In *Advances in Intelligent Data Analysis*, pages 297–308, 2009.

[8] D. Dubois and H. Prade. Gradual inference rules in approximate reasoning. *Information Sciences*, 61(1–2):103–122, 1992.

[9] K. D. Garbade and W. L. Silber. Dominant and satellite markets: a study of dually-traded securities. *The Review of Economics and Statistics*, 61(3):455–460, 1979.

[10] J. Grammig, M. Melvin, and C. Schlag. Internationally cross-listed stock prices during overlapping trading hours: price discovery and exchange rate effects. *Journal of Empirical Finance*, 12(1):139–164, 2005.

[11] E. Hüllermeier. Implication-based fuzzy association rules. In *Principles of Data Mining and Knowledge Discovery*, pages 241–252, 2001.

[12] E. Hüllermeier. Association rules for expressing gradual dependencies. In *Proc. of the Int. Conf. on Principles of Data Mining and Knowledge Discovery*, pages 200–211, 2002.

[13] M. H. Iyeki. Adr volume lifts market. *NYSE Euronext*, July 2013.

[14] N. Jayaraman, K. Shastri, and K. Tandon. The impact of international cross listing on risk and return: the evidence from american depository receipts. *Journal of Banking and Finance*, 17(1):91–103, 2014.

[15] J. Kacprzyk and R.R. Yager. Linguistic summaries of data using fuzzy logic. *International Journal of General Systems*, 30:133–154, 2001.

[16] J. Kacprzyk, R.R. Yager, and S. Zadrony. A fuzzy logic based approach to linguistic summaries of databases. *International Journal of Applied Mathematics and Computer Science*, 10:813–834, 2000.

[17] K. Kehrle and F. J. Peter. Who moves first? an intensity-based measure for information flows across stock exchanges. *Journal of Banking & Finance*, 37(5):1629–1642, 2013.

[18] P. Korczak and K. Phylaktis. Related securities and price discovery: evidence from nyse-listed non-u.s. stocks. *Journal of Empirical Finance*, 17:566–584, 2010.

[19] A. Laurent, M.-J. Lesot, and M. Rifqi. Graank: Exploiting rank correlations for extracting gradual itemsets. In *Proc. of the Int. Conf. on FQAS*, pages 382–393, 2009.

[20] E. Lok and P. S. Kalev. The intraday price behavior of australian and new zealand cross-listed stocks. *International Review of Financial Analysis*, 15(4–5):377–397, 2006.

[21] A. Oudni, M.-J. Lesot, and M. Rifqi. Characterisation of gradual itemsets through especially if clauses based on mathematical morphology tools. In *Proc. of the Int. Conf. of the European Society for Fuzzy Logic and Technology*, pages 826–833, 2013.

[22] A. Oudni, M.-J. Lesot, and M. Rifqi. Accelerating effect of attribute variations: Accelerated gradual itemsets extraction. In *Proc. of the Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2014.

[23] R. Pascual, B. PascualL-Fuster, and F. Climent. Cross-listing, price discovery and the informativeness of the trading process. *Journal of Financial Markets*, 9(2):144–161, 2006.

[24] M. Qadan and J. Yagil. Main or satellite? testing causality-in-mean and variance for dually listed stocks. *International Journal of Finance and Economics*, 71:279–289, 2012.

[25] R. L. F. Silveira, L. Maciel, and R. Ballini. Cointegration and causality-in-mean and variance tests: evidence of price discovery for brazilian cross-listed stocks. In *Proceedings of the International Conference on Finance, Banking, and Regulation*, Brasilia, Brazil, July 2014.

[26] L. Szathmary, A. Napoli, and P. Valtchev. Towards rare itemset mining. In *Proc. of the 19th IEEE Int. Conf. on Tools with Artificial Intelligence*, pages 305–312, 2007.

[27] J.-Y. Wan and C.-W Kao. Price discovery in taiwan's foreign exchange market. *Journal of International Financial Markets, Institutions and Money*, 19(1):77–93, 2009.

[28] S. S. Wang, O. M. Rui, and M. Firth. Return and volatility behavior of dually-traded stocks: the case of hong kong. *Journal of International Money and Finance*, 21(2):265–293, 2002.

[29] R.R. Yager. A new approach to the summarization of data. *Information Sciences*, 28:69–86, 2001.

[30] L. A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, 9:149–184, 1983.