ANALYSIS OF INTERACTIONS BETWEEN VARIABLES USING TOPIC MODELING.

Afonso Valau de Lima Junior
Escola de Administração UFRGS
Rua Washington Luiz, 855 - Centro Histórico, Porto Alegre – RS
avljunior@yahoo.com.br

João Luiz Becker
FGV EAESP
Av. 9 de julho, 2029, Bela Vista, São Paulo – SP
joao.becker@fgv.br

RESUMO

Um modelo de tópico é baseado em um corpus de documentos, descobre os tópicos que permeiam o corpus e atribui documentos a esses tópicos, um dos mais citados na literatura é a *Latent Dirichlet Allocation (LDA).* O presente estudo tem como objetivo investigar o grau de relação entre o índice Ibovespa e os tópicos extraídos do *New York Times* (dados textuais), bem como avaliar os impactos a longo prazo quando uma variável recebe um choque. Foi possível evidenciar relação entre notícias e índice do mercado de ações, embora ainda haja melhorias e comparações a serem feitas.

PALAVRAS CHAVE. Topic model, textual data, LDA.

Tópicos (EST&MP, BDA, OA)

ABSTRACT

A topic model is based on a corpus of documents, it discovers the topics that permeate the corpus and assigns documents to those topics, one of the most cited in the literature is the Latent Dirichlet Allocation (LDA). The present study aims to investigate the degree of relationship between the Ibovespa index and the topics extracted from the New York Times (textual data), as well as to evaluate the long-term impacts when a variable receives a shock. It was possible to relate evidence of the relationship between news and the stock market index, although there are still improvements and comparisons to be made.

KEYWORDS. Topic model, textual data, LDA.

Paper topics (EST&MP, BDA, OA)

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

1. Introduction

With the continued growth of various information sources, where much data is being generated every minute, the need for data compression and analysis tools and management are becoming more remarkable. According to [Davenport 2014], in 2012, the world used over 2.8 trillion gigabytes of data. Companies that could take advantage of this large volume of data to convert them into ideas, innovations, and commercial value they are only scratching the surface of what's possible. The same study points out that only 0.5% of the 2.8 trillion gigabytes is analyzed. Developing hardware and software platforms, mainly for the web network, allowed the rapid creation of large amounts and different types of data, particularly for the case of text data.

For [Chen et al. 2012], computational tools are required to organize, search, and understand large amounts of data, especially as our collective knowledge continues to be digitized and stored in many types of media. Regarding the textual data, [Srivastava et al. 2013] highlight that they were originally available in an unstructured form. For useful representation of documents to obtain high quality information promotes the research and design of algorithms capable of discovering patterns and trends through Topic modeling.

The main purpose of topic modeling is to discover the main themes that permeate a collection of documents, discovering patterns in word usage and how to connect documents that share similar patterns. Pattern discovery reflects the underlying topics that come together to form the documents [Han et al. 2011]. The original resource representation plays a key role in defining topics and identifying which over are present in each document, resulting in an understandable representation of documents useful for analyzing subjects in documents [Miner 2012]. Extraction models of topics can be divided into non-probabilistic and probabilistic.

Probabilistic models have become popular, with the probabilistic Latent Semantic Analysis (pLSA) model introduced by [Hofmann 1999], the first to formalize the extraction of probabilistic topics. Although it provides a good basis for an analysis of texts, the pLSA model presents two problems. First, the topic generation process for each document is not defined, which requires determining a quantity of parameters that grows linearly with the number of documents and can lead to overfitting the estimated parameters. In addition, the pLSA model does not determine a natural way to calculate probabilities related to a document not in the training set [Blei et al. 2003], [Kim et al. 2012]. To avoid these problems, [Blei et al. 2003] proposed the Latent Dirichlet Allocation (LDA) model.

Numerous other extensions and applications with topic models can be found in the literature, not just for textual data; for example, [Fei-Fei and Perona 2005] and [Sivic et al. 2005] used topic templates to analyze images. There is intense research in developing new extensions of topic models that allow the incorporation of metadata and in several areas; for example, [Rosen-Zvi et al. 2004] proposes the Author-Topic Model that relates topics and authorship of documents; Words and voting patterns [Wang et al. 2005].

The analysis of numerical time series has a long history in statistics, and many methodologies are consolidated in the literature. The modeling of themes considering the time factor is another promising direction of research, since this factor can identify important information hidden in the documents. As time passes, topics in a document corpus evolve; most topic templates assume documents are interchangeable, meaning their probability is invariant to permutation [Wang et al. 2008].

In this sense, the present study aims to investigate the degree of relationship between the Ibovespa index and the topics extracted from the New York Times, as well as assess the long-term impacts when a variable receives a shock.

This study is organized as follows. Section 2 is based on the consistency of the study that explains the published scientific literature, addressing the Latent Dirichlet Allocation (LDA) model and a brief approach to the VAR/VEC methodology. Section 3 presents the results obtained. Finally we conclude in section 4.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

2. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a document generating probabilistic model. In this model, the observable variables are the terms of each document, and the non-observable variables are the topic distributions. The parameters of the topic distributions, known as hyperparameters, are given as priors in the model. The distribution used to sample the distribution of topics is the Dirichlet distribution. In the generative process, the Dirichlet sampling result is used to allocate the words of different topics, and they will fill in the documents. One can perceive the meaning of the name Latent Dirichlet Allocation, which expresses the intention of the model to allocate the latent topics distributed obeying the distribution of Dirichlet

The intuitive idea behind the LDA and illustrated by the authors [Blei et al. 2003], it is assumed that several topics, distributions over words, exist for the entire collection. Each document is assumed to be generated as follows: First, choose a distribution on the topics and then, for each word, choose a topic assignment and choose the corresponding topic word. In the seminal work, the authors exemplify the idea with a scientific paper on the use of data analysis to determine the number of genes an organism needs to survive (in an evolutionary sense) by analyzing the article and different words used in the article as Computer, forecast, life, organism, genes, sequenced. In the example, if we take the time to highlight each word in the article, you would see this article combines genetics, data analysis, and evolutionary biology with different proportions. LDA is a statistical model of document collections that attempts to capture this intuition [Blei 2012].

The distribution of Dirichlet has some important properties [Bishop 2006], and for this reason, they are commonly used in Bayesian statistics; the distribution of Dirichlet is *a priori* conjugated of the multinomial distribution.

The LDA generator process is an imaginary process and, conversely to what is proposed in a computational task of extracting information, it is assumed the topics are specified before any data is generated. Here, topics are defined as probability distributions over a fixed vocabulary (words), while documents, nothing more than bags of words, arise from the probabilistic choice of words belonging to topic distributions. The whole generative process can be represented graphically with a Bayesian network. This network is illustrated in the figure 1.

As an introduction, the Bayesian model of LDA is a hierarchical model with three levels, where the first level represents the distribution of topics throughout the collection of documents. At the second level, one has the distribution of the topics for each document. And in the third level, the distribution of the topics internally for words in a document and with the last level is repeated, it is possible to represent a document as a mixture of topics.

To represent the distributions, two variables are used; the variable φ is an n-dimensional variable, where n is the number of words in the vocabulary. The variable θ is a K-dimensional variable, where the value of K is the number of topics. These two variables are generated by the Dirichlet distribution with their respective β and α hyperparameters.

Then, with the distributions φ and θ_d, the document d_j is generated. In the LDA model, it is considered that a document is simply a bag of words, with nd terms in a document d. The terms of a bag of words are vocabulary words, and occasionally, repetitions of the same word may occur. For each position i, of the nd terms positions of a bag of words, a word of the distribution of topics is chosen. To do this, one must choose a topic k of the existing K topics and associate this topic with the position i of the document d.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

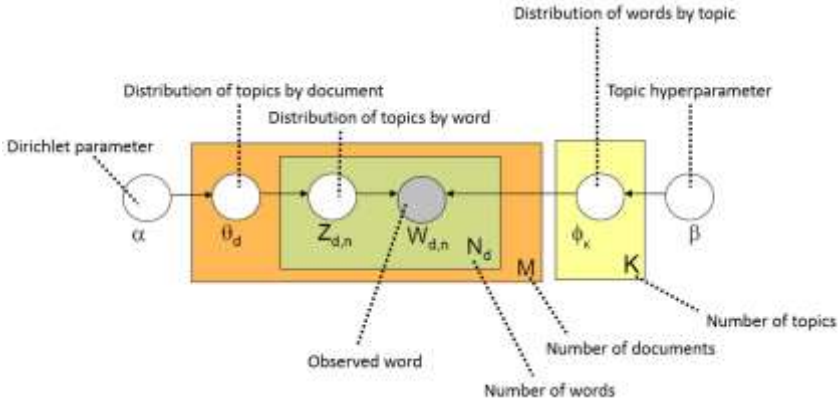


Figure 1 – Graphic model of the LDA.
Source: Adapted [Blei et al. 2003].

The topic is chosen by obeying the θ_d distribution, which informs the participation of the topics in the *d* document; the variable z_{dn} will store the chosen topic. Then, the φ distribution is chosen as the word that will fill the position *i*. The variable φ are K n-dimensional distributions, where each distribution *k*, φ_k, corresponds to proportions of words that semantically describe the subject of which the topic *k* treats. The term wdn should be chosen from the topic zdn, obeying the word distribution φ_{zdn} .

In the LDA, the same document can be related to several topics with different proportions of relevance, since each document has its own distribution of topics θ_d. One can see this in the generative model by choosing the topic assigned to variable z_{dn}, where occasionally there will be a chance of choosing different topics according to θ_d distribution.

The central computational problem can be solved by inferring the a posteriori probability of the whole model. This can be thought of as the inverse of the generative process. Theoretically, this inference calculation can be done by summing the joint distribution of all possible values assigned to the unobserved variables (all words in the collection). The number of possible assignments is exponentially large, making this calculation computationally intractable [Blei 2012]. Despite this, several methods approximate a posterior distribution and among the methods most used in the literature for inference of the LDA model is Gibbs Sampling [Griffiths and Steyvers 2004].

The Gibbs Sampling is most popular mainly for the ease of implementation and its application in several problems. It is a special case of Monte Carlo simulation in Markov Chain. Markov Chain Monte Carlo methods can emulate high-dimensional probability distributions through the Markov chain stationary behavior [Geman and Geman 1984].

The process performed by the Gibbs Sampling is based on sampling each dimension alternately, one at a time, conditioned to the value of all other dimensions. Suppose there is an unobserved K-dimensional variable, $z = \{z_1,z_2,...,z_K\}$, where zi corresponds to the value of the ith vector dimension *z* and $z_{-i} = \{z_1,z_2,...,z_{i-1},z_{i+1},...,z_K\}$; also suppose the evidence given by the observed variable, *w*. In this example, the distribution to be inferred is $p(z|w)$. Instead of making samples throughout the *z* distribution to infer $p(z|w)$, the Gibbs Sampling makes separate choices for each *i* dimension of *z*, where sampling z_i depends on the other dimensions in z_{-i} sampled so far [Russell and Norvig 2003].

With LDA, the Gibbs Sampling must sample three hidden variables, *z*, θ and φ. For simplicity, the Gibbs method applied in the LDA is collapsed to sample only the variable *z*, and from *z* find the values of the variables θ and φ. The sampling process is performed by statistics obtained by counting word assignments for topics and topics for documents made after sampling [Blei 2012].

3. VAR and VEC models

[Sims 1980] proposed, through the VAR (Vector autoregression), a model capable of evaluating, with few restrictions, the relationships existing between all the variables of a study; that is, all variables are treated as endogenous. Each endogenous variable in the model is explained by its lagged values and the lagged values of all other endogenous variables of the model. In this way, the VAR allows to evaluate the dynamic impact of random perturbations on the system of variables, making them particularly useful and efficient in predicting the future behavior of interrelated time series [Hamilton 1994]. It can be mathematically represented by equation 1:

$$X_t = A_0 + \sum_{i=1}^{p} A_i X_{t-i} + \epsilon_t \tag{1}$$

In that X_t is a vector of endogenous variables under study, A_0 is a vector of intercepts, A_i are matrices, and e_t is white noise and a vector of innovation or shock not expected in the variable under study. This shock on the variable *i-th* does not only directly affect the variable *i-th* of the model but is transmitted to the other endogenous variables of the model.

This first stage is of great importance because, according to [Engle and Granger 1987], the elements X_t , *nx1*, are said to be cointegrated of order *(d, b)*, denoted by $X_t \approx CI(d, b)$, if two conditions are met: the first condition states that all the variables under study must have the same integration order (*d*) to be cointegrated; that is, all the elements of X_t are integrated of order *d*, we have, *I (d)*; the second condition refers to the existence of a nonzero vector, β, such that $u_t = X_t\beta \approx I(d - b)$, $b > 0$, the variables hold a long-term relationship because they are not stationary, therefore, have a stochastic tendency, and this stochastic tendency justifies the long-term relationship, according to [Hamilton 1994].

Estimation of the long-term relationship, identifying the variables that have common trajectories, is the second methodological step. In this way, a VAR model is estimated, allowing the determination of the residuals that will structure the vector error correction model VEC) and the equations to interpret the response impulses [Stoffer 2017].

If the variables are cointegrated, the next methodological step is the estimation of the error correction model. According to [Enders 2014], considering the variables X_t and Y_t , we have (equation 2):

$$\Delta Y_t = \alpha_1 + \alpha_y(Y_{t-1} - \beta_1 Z_{t-1}) + \sum_{i=1} \alpha_{11}(i)_{t-i} + \sum_{i=1} \alpha_{12}(i)_{t-i} + \varepsilon_{yt} \tag{2}$$

If the error correction model is a VAR model in the first differences and if the model has the same explanatory variables and the same number of lags, it is possible to use the Ordinary Least Squares estimation. Otherwise, other estimation methods, such as the near VAR, are used [Stoffer 2017].

According to [Lütkepohl and Krätzig 2004], the number of lags needed to be incorporated in the error correction model is determined in this step. The Akaike (AIC) and Schwarz (SBC) information criteria, calculated based on the estimated variance, can be used. These criteria, called penalty criteria, take into account the number of parameters used in the modeling, algebraically. These criteria are defined as (equations 3 and 4):

$$AIC = -2\ln(L)/T + 2n\ /\ T \tag{3}$$

$$SBC = -2\ln(L)/T + n\ \ln(T)\ /\ T \tag{4}$$

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

where T is the sample size, l is the value of the *log* likelihood function, and k is the number of estimated parameters.

Another important consideration is the number of cointegration vectors that can exist in the system. When $n = 2$, no matter what the dependent variable, there is only one cointegrating vector. Similar to the regression case of X versus Y, as in the case of being estimated by the regression of Y against **X,** there is a single integration vector. When there is more than one integration vector, the method developed by Engle-Granger may suffer some bias. For the determination of the number of integration vectors, two other tests can be used: trace statistics and maximum eigenvalue [Weigend 2018].

According to [Johansen and Juselius 1990], considering the rank (r) of a matrix, that is, its eigenvalues λi not null, the test of the trait has the following hypotheses: H_0 (the number of different cointegration vectors is less than or equal to) and H1 (the number of cointegration vectors is greater than r) (equations 5 and 6):

$$\lambda_{trace}(r) = -T \sum_{i=r+1}^{n} ln(1 - \hat{\lambda}_i) \tag{5}$$

$$\lambda_{max}(r, r+1) = -Tln(1 - \hat{\lambda}_{r+1}) \tag{6}$$

Where λ_i are the characteristic roots obtained from the matrix π, which represents the error correction vectors, for details see [Enders 2014], and T corresponds to the number of observations.

In the adjustment phase of the adjusted model, the diagnosis of the model is performed through the residual analysis, which searches for residues with white noise characteristics. The impulse response of one variable in the other is checked later, using Choloeski's decomposition method, so the impulses are orthogonal.

To verify the direction of causality, in this last step, the Granger Causality Test, proposed by [Granger 1969], is used. It is not a causal test in the sense of a cause and effect relationship, but a temporal provenance test; that is, if the variable X_t causes the variable Y_t , then changes in X_t will cause temporary changes Y_t . The specification of this test can be expressed as follows (equation 7):

$$Y_t = \sum_{i=1}^{p} \alpha_i Y_{t-1} + \sum_{i=1}^{p} \beta_i X_{t-1} + \mu_t \tag{7}$$

Where α and β are the coefficients of the regressions to be estimated, X_t and Y_t are the variables and μ_t indicates the random error term.

3. Instance

The textual database used was composed of news from The New York Times (https://www.nytimes.com/), an American newspaper based in New York City with worldwide influence and readership.

The LDA model is conditioned by three parameters: the Dirichlet hyperparameters (α and β) and the number of topics (k). To determine the number of topics (k) to be used in the LDA model, the 280381 NYT news articles were used, totaling 137660 words.

The first step started with filtering and cleaning the data by removing special characters and accents besides the removal of stopwords. The topics were extracted with the LDA model using the Gibbs Sampling technique [Griffiths and Steyvers 2004], using $k = 7$ (number of topics). To consider a study over time, the topic extraction model was applied in sets of news of the day, and the performance of each topic was analyzed over the years - at this stage, only the news published between 01.01.2014 and 01.31.2017 were considered.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

Table 1: Most likely word from 7 topics.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
may	clinton	attack	art	brother	first	friend
say	donald	day	book	game	monday	husband
state	like	latest	met	point	week	love
year	new	need	schoool	score	won	school
york	trump	time	work	team	world	wife

Topic 1 (may, say, state, year, york) was classified as relations of the city of New York, but it is not very clear and evident as the others; topic 2 (clinton, donald, how, new, trump) presidential elections/campaign that began in 2015/2016; in topic 3 (attack, day, last, necessity, time), more issues emerge related to security issues and international relations; in topic 4 (art, book, meeting, school, work) are the words related to the question of education and art; topic 5 (brother, game, point, score, team) was classified as sportsrelated issues; topic 6 (first, Monday, week, won, world) as topic 1 was not very clear but was classified as international relations according to the word "world"; topic 7 (friend, husband, love, school, wife) was classified as a family because they refer to the traditional family composition. Topics over time is illustrated in the figure 2.

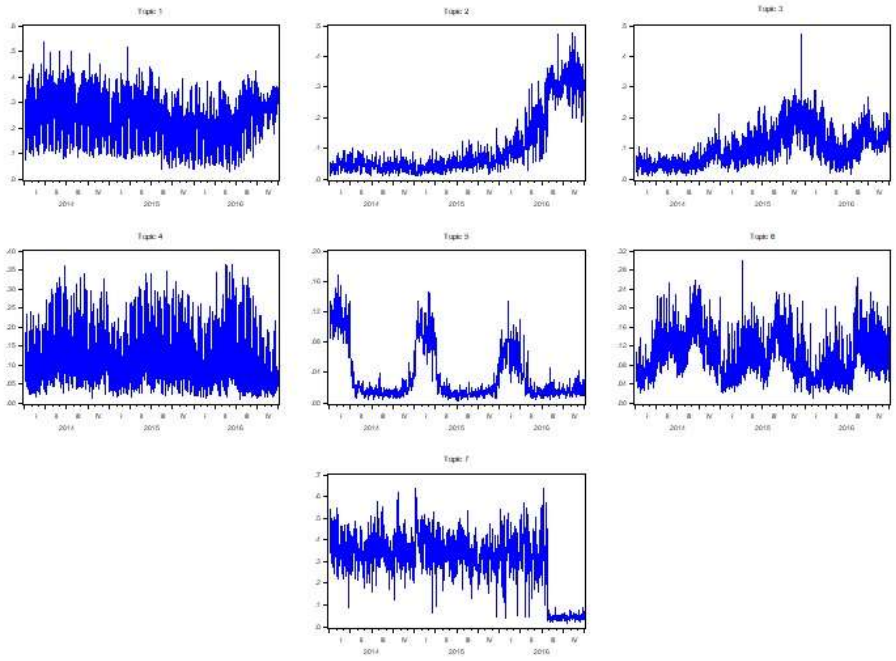


Figure 2: Topics over time (2014 - 2016).

There is a lot of change over time in relation to the participation of each theme, such as Topic 2. In 2016, the increasing in function of the electoral period is clear; topics 5, 6 and 7 present a very significant break in structure; topic 3 is growing at the end of 2015; topics 1 and 4 are the ones that present the most constancy in the analysed period. This shows an interesting evolution and variability of topics over time.

After the goal is to determine the degree of relationship between the Bovespa Index and the New York Times, as well as assess the long-term impacts when a variable receives an external shock. The financial database used was composed by The Bovespa Index (Ibovespa - http://www.bmfbovespa.com.br), the reference index of about 60 shares traded on Bovespa from

the volume traded in recent months. The present value represents an amount, in currency, of a theoretical portfolio of shares, from a hypothetical application.

To fulfill this objective, only a few topics were selected (of the 7 presented in previous step). Only topics that have some relevance to the economy (such as state issues, international relations, and elections) were considered – Topic 1, Topic 2, Topic 3 e Topic 6. In the figure 3, the evolution of the participation of the selected themes and Bovespa Index between 2014 and 2016 is presented.

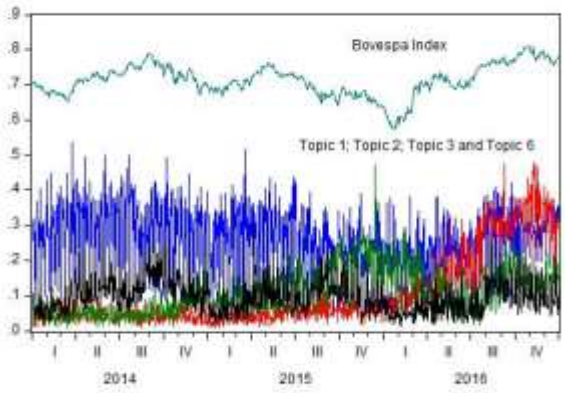


Figure 3: Topics over time and Bovespa Index (2014 - 2016).

First by visual analysis, it is possible to verify the non-stationarity of the series in level and to corroborate the visual inspection and verify the stationarity of the series, the tests (ADF) and (KPSS), and verified when applied in the differentiated series are stationary.

Since the series in level are non-stationary and after the first difference they become stationary, there must be a long-term equilibrium that deserves to be investigated through the incorporation of an error correction mechanism.

Initially, the VAR model was estimated. At this stage, the variables to be composed of the model must be stationary; therefore, they will be differentiated. The number of lags that will be part of the model will also be defined.

In the table 2, the lagging criteria to compose the VAR model are presented; are indicates lag order selected by the criterion; LR: sequential modified LR test statistic (each test at 5% level); AIC: Akaike information criterion; SC: Schwarz information criterion; HQ: Hannan-Quinn information criterion.

Table 2: Determination of the number of lags to composse the autoregressivevector model using the selection criteria described.

Lag	LogL	LR	AIC	SC	HQ
0	7363.560	NA	-13.48957	-13.46668	-13.48091
1	11246.19	7722.552	-20.56130	-20.42395	-20.50932
2	11609.60	719.4966	-21.18167	-20.92987	-21.08637
3	11791.94	359.33675	-21.47011	-21.10385	-21.33149
4	11880.61	173.9207	-21.58682	-21.10611	-21.40489
5	11977.50	189.1555*	-21.71860*	-21.12343*	-21.49335*

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

Table 3: Trace Test and Maximum Eigenvalue test to verify the existence of the cointegration vector number based on the eigenvalue.

Hypothesized nº of CE(s)	Eigenvalue	Trace statistic	0.05 Critical Value	Max-eigen statistic	0.05 Critical Value	p-value
None	0.289461	517.9086	69.818889	373.1711	33.87687	<0.05
At most 1	0.082229	144.7376	47.85613	93.70150	27.58434	<0.05
At most 2	0.036561	51.03606	29.79707	40.67312	21.13162	<0.05
At most 3	0.007996	10.36293	15.49471	8.766357	14.26460	<0.05

The criteria for the selection of the number of lags point to a VAR model with five lags ane is observed that there is an interdependence between variables, which is useful for determining long-term relationships, examining the impulse of response, and causal relations.

If the VAR model is estimated, it is necessary to confirm the number of cointegrating vectors by means of the statistical tests of the Trace and the maximum eigenvalue, according to table 3.

In table 3, the result of the Trace test and Maximum eigenvalue test, where: the null hypothesis is that the number of distinct cointegration vectors is less or equal to *r*, in this case, the null hypothesis is rejected, pointing to the existence of, at most, 3 vectors of cointegration.

The error correction model shows the variables are interconnected, and there is a significant combination, representing a longterm equilibrium between the variables, where all are significant. The variable Ibovespa was taken as standard and the variables Topic 1, Topic 2, Topic 3 and Topic 6 all with lags and with orders of magnitude of the linear combination of -79.961757; -4.20195748; -40.419986; 54.8890895 respectively.

From the estimation of the VAR model, it becomes possible to analyze the impulseresponse functions, which allow to verify the effects of shocks (variations) that a variable receives and is transmitted throughout the system.

In figure 4, the behavior of the response impulse is verified. A shock or a change in a variable does not only affect the variable at that instant, but it is transmitted to the other variables and in subsequent periods of time through the lag structure of the VAR model. For the analysis, the Cholesky decomposition was used so the pulses were orthogonal.

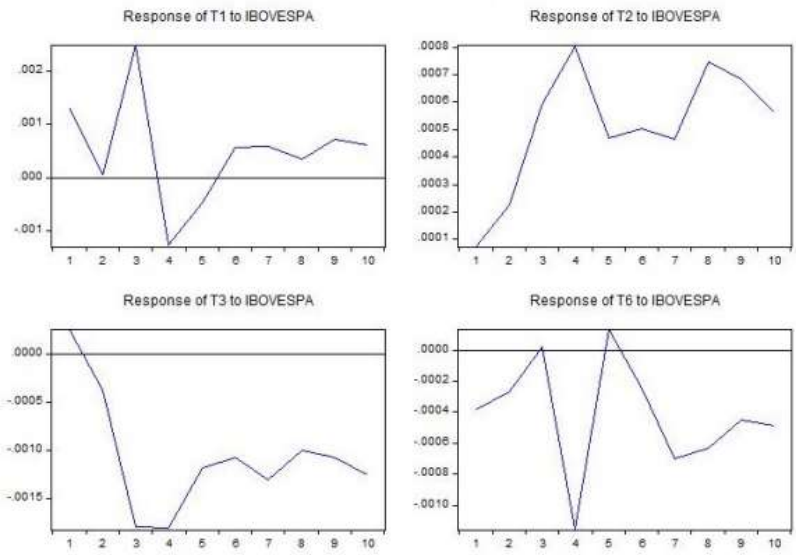


Figure 4: Response to Cholesky One S.D. (d.f. adjusted) Innovations.

When a change in the system is introduced through the variables T1, T2, T3 and T6, two similar behaviors of the variable Ibovespa can be perceived. Initially, a change introduced

into the system by means of the variable T1 causes the Ibovespa variable to fall at the first instant and rise at the second instant and continue with this oscillation until the sixth instant, where the oscillations are smaller. Very similar when the change is introduced in variable T6.

When introduced into the system, a change through variable T2 causes the Ibovespa variable to increase substantially to the fourth instant and decay, but in the seventh period, it returns to rise; a similar behavior occurs, but only inversely, when a change is introduced through the variable T3.

The Granger causality test was used (table 4) to verify if there is a causal orientation between the Topics and the Ibovespa variable.

Table 4: Causal relationship for the series according to Granger Causality.

Null hypothesis	Statistics F	Probability
T1 does not Granger Cause IBOVESPA	1.43088	>0.10
IBOVESPA does not Granger Cause T1	**13.6556**	**<0.10**
T2 does not Granger Cause IBOVESPA	**2.01236**	**<0.10**
IBOVESPA does not Granger Cause T2	0.52464	>0.10
T3 does not Granger Cause IBOVESPA	0.88430	>0.10
IBOVESPA does not Granger Cause T3	1.72703	>0.10
T6 does not Granger Cause IBOVESPA	0.57165	>0.10
IBOVESPA does not Granger Cause T6	**4.23290**	**<0.10**

Considering the results organized in table 4, it is possible to observe the absence of two-dimensional relations between the variables. The Ibovespa causes one-dimensional variations in the T1 and T6 variables. The Ibovespa variable suffers one-dimensional variations of the variable T2. A significance of 10. Causal relationship is illustrated in the figure 5.

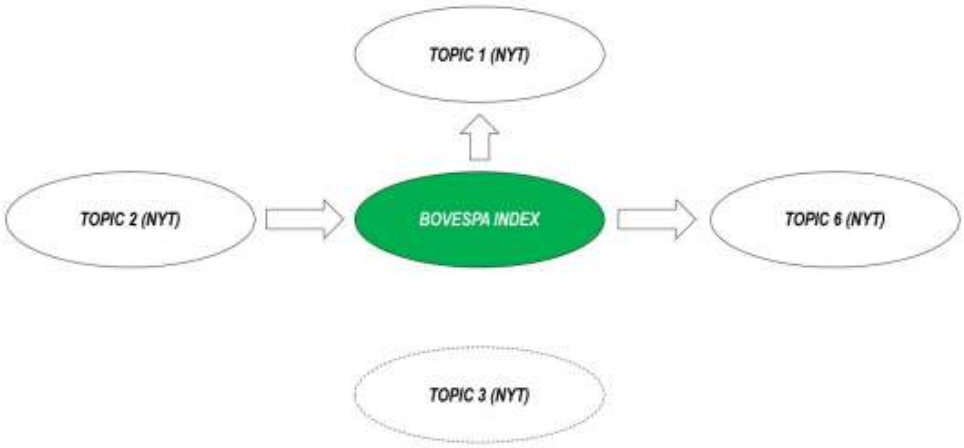


Figure 5: Causal relationship for the series according to Granger Causality.

According to the results of the Granger causality test, the T3 variable has no causal relationship with any other variable in the study. As shown, the relationship between the Ibovespa index and the New York Times represented by topic 2 (clinton, donald, how, new, trump) that was classified with presidential elections/campaign is evident. This result is very relevant because, in studies of economic indicators, most of the data used are from other economic indicators, that is, numerical data with numerical data.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

4. Conclusion

Probabilistic topic modeling in text mining is a promising direction of research, considered mainly the important time factor. Traditional probabilistic mixing models, such as LDA, imply that data records (such as documents) are fully interchangeable. However, the data are collected naturally over time, following a sequence.

As shown, the relationship between the Ibovespa index and the New York Times represented by topic 2 (clinton, donald, how, new, trump) that was classified with presidential elections/campaign is evident. This result is very relevant because, in studies of economic indicators, most of the data used are from other economic indicators, that is, numerical data with numerical data. It is necessary to search a database of a Brazilian newspaper to show some relationship between economic indexes and news in the country. Another future contribution is the evidence of the relationship between news and stock market index, although there are still improvements and comparisons to be made.

Referências

BISHOP, Christopher M. Pattern recognition and machine learning. [S.l.]: springer, 2006.

BLEI, David M; NG, Andrew Y; JORDAN, Michael I. Latent dirichlet allocation. Journal of machine Learning research, v. 3, n. Jan, p. 993–1022, 2003.

BLEI, David M. Probabilistic topic models. Commun. ACM, ACM, New York, NY, USA, v. 55, n. 4, p. 77–84, abr. 2012. ISSN 0001-0782. Disponível em: https://dl.acm.org/doi/10.1145/2133806.2133826.

CHEN, Hsinchun; CHIANG, Roger HL; STOREY, Veda C. Business intelligence and analytics: From big data to big impact. MIS quarterly, v. 36, n. 4, 2012.

DAVENPORT, Thomas. Big data at work: dispelling the myths, uncovering the opportunities. [S.l.]: Harvard Business Review Press, 2014

ENDERS, W. Applied Econometric Time Series, 4th Edition. Wiley, 2014. (Wiley Series in Probability and Statistics). ISBN 9781118918661. Disponível em: https://books.google.com.br/books?id=GP1EBQAAQBAJ .

ENGLE, Robert F.; GRANGER, C. W. J. Co-integration and error correction: Representation, estimation, and testing. Econometrica, [Wiley, Econometric Society], v. 55, n. 2, p. 251–276, 1987. ISSN 00129682, 14680262. Disponível em: https://www.jstor.org/stable/1913236?seq=1 .

FEI-FEI, Li; PERONA, Pietro. A bayesian hierarchical model for learning natural scene categories. In: IEEE. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. [S.l.], 2005. v. 2, p. 524–531.

GEMAN, Stuart; GEMAN, Donald. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence, IEEE, n. 6, p. 721–741, 1984.

GRANGER, Clive WJ. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, JSTOR, p. 424–438, 1969.

GRIFFITHS, Thomas L; STEYVERS, Mark. Finding scientific topics. Proceedings of the National academy of Sciences, National Acad Sciences, v. 101, n. suppl 1, p. 5228–5235, 2004.

KIM, Hyun Duk; PARK, Dae Hoon; LU, Yue; ZHAI, ChengXiang. Enriching text representation with frequent pattern mining for probabilistic topic modeling. Proceedings of the American Society for Information Science and Technology, Wiley Online Library, v. 49, n. 1, p. 1–10, 2012.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

HAMILTON, James D. Time Series Analysis. [S.l.]: Princeton University Press Usa, 1994.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. Data mining: concepts and techniques. [S.l.]: Elsevier, 2011.

HOFMANN, Thomas. Probabilistic latent semantic indexing. In: ACM. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. [S.l.], 1999. p. 50–57.

JOHANSEN, Søren; JUSELIUS, Katarina. Maximum likelihood estimation and inference on cointegration—with applications to the demand for money. Oxford Bulletin of Economics and statistics, Wiley Online Library, v. 52, n. 2, p. 169–210, 1990.

LÜTKEPOHL, Helmut; KRÄTZIG, Markus. Applied time series econometrics. [S.l.]: Cambridge university press, 2004.

MINER, Gary. Practical text mining and statistical analysis for non-structured text data applications. [S.l.]: Academic Press, 2012.

ROSEN-ZVI, Michal; GRIFFITHS, Thomas; STEYVERS, Mark; SMYTH, Padhraic. The author-topic model for authors and documents. In: AUAI PRESS. Proceedings of the 20th conference on Uncertainty in artificial intelligence. [S.l.], 2004. p. 487–494.

RUSSELL, Stuart J.; NORVIG, Peter. Artificial Intelligence: A Modern Approach. 2. ed. [S.l.]: Pearson Education, 2003. ISBN 0137903952.

SIMS, Christopher A. Macroeconomics and reality. Econometrica, v. 48, n. 1, p. 1–48, 1980.

SIVIC, Josef; RUSSELL, Bryan C; EFROS, Alexei A; ZISSERMAN, Andrew; FREEMAN, William T. Discovering objects and their location in images. In: IEEE. Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. [S.l.], 2005. v. 1, p. 370–377.

STOFFER, Robert H. Shumway; David S. Time Series Analysis and Its Applications. [S.l.]: Springer, 2017.

SRIVASTAVA, Nitish; SALAKHUTDINOV, Ruslan; HINTON, Geoffrey E. Modeling documents with deep boltzmann machines. CoRR, abs/1309.6865, 2013. Disponível em: https://arxiv.org/abs/1309.6865.

WANG, Xuerui; MOHANTY, Natasha; MCCALLUM, Andrew. Group and topic discovery from relations and text. In: ACM. Proceedings of the 3rd international workshop on Link discovery. [S.l.], 2005. p. 28–35.

WANG, Chong; BLEI, David M.; HECKERMAN, David. Continuous time dynamic topic models. 2008. Disponível em: http://uai2008.cs.helsinki.fi/UAI_camera_ready/wang.pdf.

WEIGEND, Andreas S. Time series prediction: forecasting the future and understanding the past. [S.l.]: Routledge, 2018.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings