Using High-Frequency Data to Price Financial Assets: A Systematic Review

Abstract

This study conducts a bibliometric analysis and systematic review of the empirical literature on using high-frequency data to price financial assets, examining 93 articles. The bibliometric analysis was developed by counting frequency and co-citations, while the systematic review encompassed qualitative analysis to establish a correlation between relevant themes still little explored. These articles were retrieved from the Scopus and Web of Science databases, and the software Biblioshiny and Rank Words were adopted to conduct the analysis and the review. In addition, the study verified the main laws of bibliometric analysis, such as Zipf (1949), Bradford (1934), and Lotka (1926). The research contributed to delving into modeling and predictability models to price financial assets based on high-frequency data and identify the main gaps to increase knowledge in the area.

Keywords: volatility; high-frequency data; forecast

Introduction

The current capital market is developing sharply in Brazil, enabling a competitive environment where agents seek to optimize capital management. Agents willing to operate efficiently must have a theoretical background and practical tools to increase their understanding of the uncertainties influencing prices.

Recent econometric studies have shown great advances concerning practical tools, developing more assertive models to predict market variations. The proposed models – heterogeneous autoregression realized volatility (HAR-RV) and high-frequency-based volatility (HEAVY) – use several measures of realized volatility containing different frequencies and methods to transform frequencies. Agents willing to operate as market traders can use this type of tool to define strategies and obtain support from daily volatility forecasts.

Numerous factors can generate instability in the capital market, which requires constant evolution in support mechanisms. The majority of the existing theoretical and empirical studies on realized volatility show the possibility of obtaining accurate forecasts through models containing ultra-high frequency time series (Andersen & Bollerslev, 1998; Andersen, Bollerslev, Diebold, & Labys, 2001; Andersen, Bollerslev, & Meddahi, 2005; Ghysels & Sinko, 2006; Koopman, Jungbacker, & Hol, 2005).

This study carries out a bibliometric analysis and systematic review of the literature on using high-frequency data to price financial assets. The research examines the methods adopted in the literature and seeks to identify possible gaps to be addressed in future studies.

Literature Review

Although the financial market works continuously, practically all available data sets on its activity are based on discrete sampling. High-frequency datasets eliminated this disparity, allowing for virtually continuous observations of price, volume, trade size, and even depth. Against this backdrop, a series of key questions arise to increase understanding of the foundations guiding the price behavior of financial assets.

The emergence of new possibilities due to high-frequency datasets accelerated the transactions, improving market efficiency. However, this acceleration favored adverse events, such as the 2010 Flash Crash, which led to the creation of new laws to increase market regulation.

The efficient market hypothesis implies that incorrect pricing and the associated arbitrage opportunities between related markets must be eliminated quickly. In addition, such a hypothesis states that the speed through which a transaction is completed influences the success of a financial asset’s purchase or sale.

For agents who want to work in this segment, making significant investments becomes a mandatory condition for success. Both the physical location and the use of computers with fast connections to the stock exchange are factors capable of reducing response time in negotiations. Since not everyone has the same ideal resources to operate their strategies, reflections on justice and equality emerge for debate.

In general, subjects are willing to give up rewards to punish those who treat them unfairly (Kahneman, Knetsch, & Thaler, 1986a, b). In a way, the human brain is programmed to prefer fair outcomes. Tabibnia, Satpute, and Lieberman (2008) report brain imaging studies demonstrating the existence of pleasure activations when a fair outcome is achieved, while other brain regions are activated for unfair results. This notion of fairness is enshrined in our securities law and regulatory apparatus, which seek to create a level playing field by forcing the disclosure of relevant information.

Various tools to improve transactions have been used in daily market activities, although academics are still debating their real effectiveness. Many technical trading systems are readily automated with computers. Today some practitioners of the reaction to news strategy use computers to examine news feeds for relevant information and to make and implement trading decisions. Most strategies are low-frequency, but some of them consider the use of high-frequency technology.

The definition of realized volatility is based on the theory of quadratic variation (Andersen et al. 2001b). It is considered a consistent estimator of the integrated volatility RV_t, defined as the sum of squared returns observed in very small time intervals. Its concept is widely applied to empirical finance (Andersen, Bollerslev, & Diebold, 2002). Among its advantages, it provides asymptotically unbiased measurements, making measurement errors (approximately) serially uncorrelated. The sampling frequency required for the realized volatility tests should be as high as the characteristics of the market microstructure without creating a biased estimator of realized volatility (Alexander, 2008). Also, the optimal sampling frequency is chosen as the highest frequency for which the term autocovariance polarization minimizes.

Among the tests carried out, the autoregressive fractionally integrated moving average (ARFIMA) model was highly effective in capturing realized volatility (Andersen & Bollerslev, 1997, 1998; Andersen et al., 2003, 2005; Thomakos & Wang, 2003; Angelidis & Degiannakis, 2008; Giot & Laurent, 2004; Koopman et al. 2005). Later, Corsi (2009) suggested another efficient alternative: the heterogeneous autoregression (HAR) model, conceived as an autoregressive structure of realized volatilities in intervals of different sizes. Its economic interpretation derives from the heterogeneous market hypothesis presented by Müller et al. (1993). The basic idea is that market participants have a different perspective on their investment horizon, generating volatility. This type of modeling proved to be very effective in other countries. However, to the best of our knowledge, it is not yet representatively documented in the literature produced with data from emerging countries.

Methodology

This study implemented seven steps to answer the question – what are the main research topics related to using high-frequency data for modeling and predicting the price of financial assets? Steps 1 to 5 use bibliometric analysis and systematic review techniques, while steps 6 and 7 refer exclusively to systematic review.

Step 1 – Choosing the database. The articles forming the sample were collected from Scopus and Web of Science (WoS) databases, two of the main databases used as a source of citations. They offer articles published in scientific journals with a high impact factor classified through the Journal Citation Reports (JCR).

Step 2 – The researchers used initial research parameters to browse the articles in the two databases selected. The research considered the period from January 1, 2015, to August 1, 2022, and 369 articles were identified. The keywords adopted to filter the publications were variations of the terms “high-frequency data,” “forecast,” and “volatility.” Subsequently, other filters were applied to refine the search, reaching 78 articles from the Scopus database and 89 articles from the WoS database, gathering 167 articles as an intermediate sample, as shown in Table 1.

Table 1

Evolution of the Sample Using the Filters From the Scopus and Web of Science Databases

	Description	**Scopus**	**WOS**	**Total**
(+)	Search: “Article title,” “Abstract,” “Keywords,” equal to: “high-frequency data,” “forecast,” and “volatility”	196	173	369
(=)	**Initial sample**	**196**	**173**	**369**
(-)	Date: before 2015	72	63	135
(-)	Language: different from “English”	3	0	3
(-)	Topic: different from “economics,” “econometrics,” and “business finance”	43	21	64
(=)	**Intermediary sample**	**78**	**89**	**167**

Step 3 – At this stage, the duplicate articles were deleted. After identifying the articles that met the criteria observed in steps 1 and 2, we used the R system to consolidate the results from the two databases, excluding duplicate articles. This process showed 49 articles repeated in the two databases. When consolidating this number, the sample was reduced to 118 articles.

In this step, the 118 articles were verified to observe if they were available in full so it would be possible to apply the bibliometric procedures. This verification led to the exclusion of 25 articles, reaching a final sample of 93 (Table 2).

Table 2

Total of the Final Sample

	Description	**Number of articles**
(+)	Web of Science + Scopus	167
(-)	Articles listed in the two databases	49
(=)	**Total of the sample**	**118**
(-)	Articles not located	25
(=)	**Final sample**	**93**

Step 4 – Research database creation and collection of the selected articles. The 93 articles that formed the final sample were downloaded from Scopus and WoS databases and analyzed. The following information was collected: author, keywords, research topics, origin, citations, subject, abstract, corresponding author, type of publication, link of the article, co-authors, funding agency, journal of publication, language, publisher, the focus of the study, year of publication, university the first author is affiliated with, complete reference, simplified reference. In addition to these data, the impact factor h-index was added, indicating the article's relevance for academia.

Step 5 – This step refers to the bibliometric analysis. The articles' objective data – such as countries, authors, keywords, and institutions – were analyzed through the Biblioshiny software, offering subsidies to elaborate tables and relationship maps. The analyses carried out were complemented with the verification of the main laws of bibliometrics: a) Zipf's Law (1949) – categorization and estimation of the frequency of keywords, using the Rank Words software to calculate Goffman's transition point or the point of transition of words from low to high frequency (concentrating words with high semantic load); b) Bradford's Law (1934) – quantifying the productivity of journals in relation to the subject studied; and c) Lotka's Law (1926) – identifying researchers that have a higher frequency of production in a given area of knowledge.

Step 6 – The articles were read and coded in this step, identifying their objectives, sample, methods, and contributions. Also, they were classified and categorized in a structured way, according to Table 3. Subcategories were defined for each of the six categories, and their sum, when counting the frequency, totals 100%.

Table 3

Matrix of Categories and Subcategories

Categories	Subcategories	Definition
1. Main theme/focus of the study	A – Use of traditional predictive models	Test effectiveness of existing traditional predictive models applied to high-frequency data
	B – Use of a risk management model	Test effectiveness of existing traditional risk models applied to high-frequency data
	C – Proposal of new predictive models	Improvement of traditional predictive models applied to high-frequency data
	D – Macroeconomic variables influencing the volatility of the asset	Based on high-frequency data, investigate the effects of relevant macroeconomic events on asset volatility
	E – High-frequency data vs. low-frequency data	Test the effectiveness of using high-frequency data to improve predictive models
	F – Others	Other topics unrelated to subcategories 1A, 1B, 1C, 1D, and 1E
	A – Study of volatility	Price variation presented by a financial asset

2. Theories related to the hypotheses	B – Risk management	Management of critical variables that can impact the price of financial assets
	C – Study of events	Relevant macroeconomic variables directly affect the price of financial assets
	D – Portfolio management	Tools aimed at the management of financial assets that make up the investor's portfolio
	E – Others	Other theories unrelated to subcategories 2A, 2B, 2C, and 2D
3. Method	A – Autoregressive model	Model that presents the same dependent and explanatory variable, and the dependent variable will be at a later time (t) than the independent variable (t-1)
	B – Risk model	Model built to develop and test different scenarios of strategies, supporting decision making
	C – Analysis of jump	Identify the presence of jump for volatility analysis
	D – Others	Other methods unrelated to subcategories 3A, 3B, and 3C
4. Data sources	A – Global	Data from companies that cover two or more groups (B, C, D, E, F)
	B – The United States and Canada	Data from companies operating in the United States and Canada
	C – Europe	Data from companies operating in Europe
	D – Asia/Oceania	Data from companies operating in Asia and Oceania
	E – Latin America	Data from companies operating in Latin America
	F – Africa	Data from companies operating in Africa

5. Results	A – Confirmation of the efficacy of existing models and hypotheses	Validation of models and hypotheses recognized in previous studies
	B – Confirmation of the efficacy of new models and hypotheses	Validation of models and hypotheses developed by the research
	C – Inconclusive results	Results that do not validate the existing and proposed hypotheses in the tested aspects
	D – Others	Other results unrelated to subcategories 5A, 5B, and 5C
6. Avenues for future studies	A – Alternative to improve the model	Include new variables identified as promising within the proposed model
	B – Expanding the use of the model for new purposes	Use the model from new possible angles, helping to solve new research problems
	C – Using the models in new samples	Testing the proposed model in other markets and asset classes

Step 7 – This step corresponds to the systematic review. After coding the matrix of categories and subcategories (Table 3) for the final sample, a frequency count of the subcategories was performed to identify knowledge gaps. These gaps were then compared with the subcategories of the sixth category, offering indications for future studies on using high-frequency data to price financial assets.

Analysis of Results

Bibliometric analysis

The final sample consisted of 93 articles from the Scopus and WoS databases, published from 2015 to 2022. When disregarding 2022, we obtained an average of 12 articles related to the topic per year, as shown in Figure 1. Among the main frequencies of keywords, the term "high-frequency data" appeared 36 times (24%), and "volatility" and "forecasting" 26 times (17%) (Figure 2). Figure 3 corresponds to the relationship between keywords and the country of the first author's university, and Figure 4 to the keyword's relevance represented by its size in the word cloud.

Figure 1

Annual Distribution of the Articles

Year	Number of articles
2015	12
2016	9
2017	9
2018	13
2019	13
2020	11
2021	17
Total	**84**
Mean	**12**

Figure2

Keywords and Their Frequency in Publications

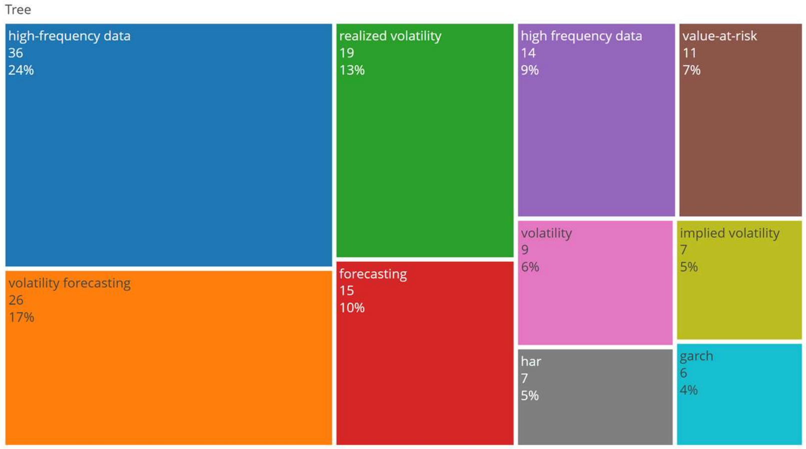


Figure 3

Relation Between Keywords and the Country of the First Author's University

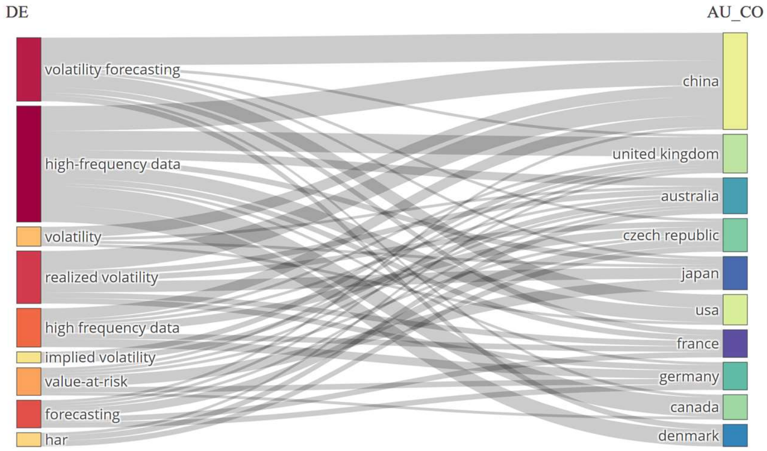
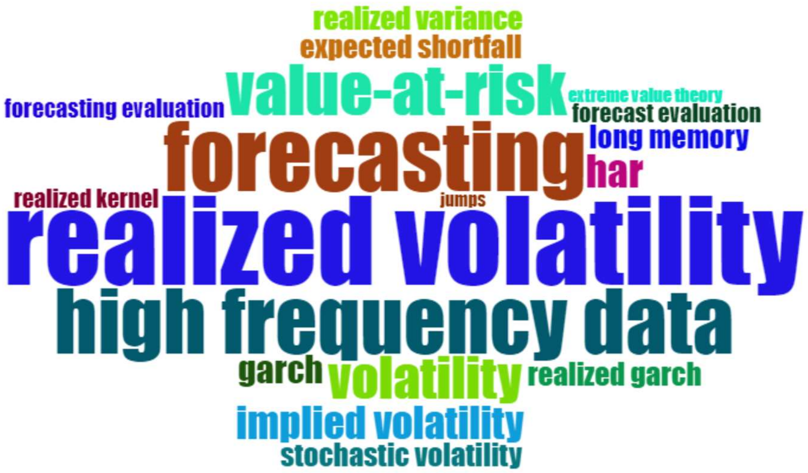


Figure 4

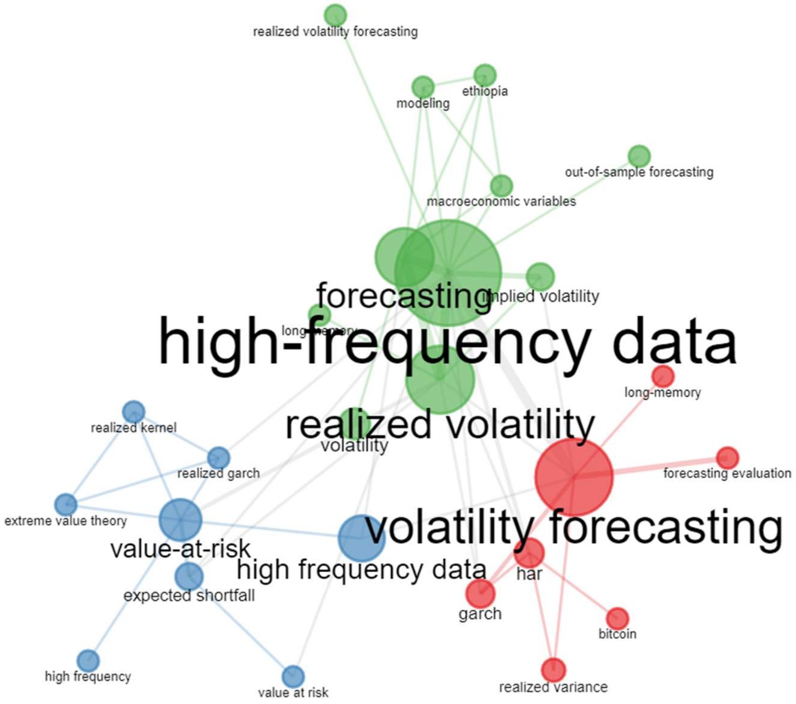
Relevance of Keywords (Represented by Font Size)



In turn, Figure 5 presents the map of co-occurrences of the most used keywords in the articles. For this observation, we highlight the words “HIGH-FREQUENCY DATA,” “VOLATILITY FORECASTING,” “FORECASTING,” and “REALIZED VOLATILITY.”

Figure 5

Keywords and the Interactions Among Them



Note: Colors represent clusters of relationships among keywords. The thickness of the lines indicates the strength of the connection between keywords.

Regarding the keywords, Zipf's (1949) law refers to optimization when using them, so they do not disperse. On the contrary, it is possible to observe a tendency for minimal use of keywords and a high frequency of occurrence. Zipf's first law states that the series (r) of a word multiplied by its frequency of occurrence (f) will be approximately constant (C), resulting in Equation 1:

$$r \text{ x } f = C \qquad (1)$$

For words with low frequency, Zipf (1949) proposed a second law, which was revised and modified by Booth (1967). As demonstrated in Equation 2, Booth (1967) suggests that in a given text, several words with low occurrence have the same frequency:

$$In = 2I1 / n\,(n+1) \qquad (2)$$

Where,

$I1$ = number of words that have frequency 1

In = number of words that have frequency n

n = Goffman's transition point from low to high frequency words

Zipf's laws (1949) define the ends of a word distribution list. Between these extreme points is a transition area from high to low-frequency words. According to Goffman (1971), this area has the words with the highest semantic content, the most suitable for thematic indexing of a given text. Pao (1978) presents the equation of Goffman's transition point (Equation 3).

$$T= (-1+\sqrt{}1+8I1) / 2 \qquad (3)$$

Where,

T = Goffman's transition point

I_1 = number of words that have frequency 1

The Rank Words software identified Goffman's transition point by sorting the keywords in descending order. Subsequently, those repeated only once were identified to calculate Goffman's transition point, locating those positioned above the classification indicated by this point. Table 4 indicates that for the 93 articles in the final sample, this point varied between 31.75 (Zevallos, 2019) and 76.35 (Sun, Chen, & Yu, 2015), with an average of 45.53.

Next, the region where words were most adherent to the text's main theme was analyzed. Words irrelevant to the study were excluded – e.g., prepositions, articles, pronouns, adverbs, and numerals. Then, the words with the highest frequency were classified (Table 4). According to this table, the term volatility was more frequent in 34 articles, especially in Lyócsa & Todorova (2020), when it is repeated 347 times in a text with a total of 15,534 words.

Table 4

Goffman's Transition Point and Classification of Words (Zipf Law)

Author	**Number of words**	**Number of different words**	**Number of words with frequency 1**	**Most recurrent word**	**Frequency of the most recurrent word**	**T = Goffman's transition point**
BOLLERSLEV T (2020)	*18,631*	*3,010*	*1,394*	*Semicovariance*	*72*	*51.80388243*
BOLLERSLEV T (2018)	*23,389*	*3,058*	*1,431*	*Risk*	*270*	*52.5*
SUN EW (2015)	*21,467*	*4,838*	*2,991*	*Financial*	*82*	*76.3450063*
PATTON AJ (2015)	*15,002*	*2,197*	*1,027*	*Volatility*	*131*	*44.32383479*
SYMITSI E (2018)	*20,632*	*3,149*	*1,664*	*Models*	*140*	*56.69098716*
TIAN S (2015)	*11,700*	*1,878*	*923*	*Volatility*	*117*	*41.96801136*

NING C (2015)	*13,620*	*2,195*	*1,102*	*Volatility*	*202*	*45.94944089*
LI J (2018)	*20,555*	*2,724*	*1,258*	*Forecast*	*116*	*49.16223679*
ASAI M (2017)	*10,516*	*2,044*	*1,103*	*Volatility*	*111*	*45.97073557*
BAUM CF (2021)	*11,817*	*2,418*	*1,421*	*Model*	*119*	*52.31275645*
WANG J (2020)	*15,439*	*2,648*	*1,427*	*Volatility*	*167*	*52.42518133*
ZHANG YJ (2019)	*11,660*	*2,332*	*1,478*	*Oil*	*233*	*53.37140793*
AVDULAJ K (2015)	*13,875*	*2,710*	*1,456*	*Oil*	*100*	*52.96526661*
DIMITRIADIS T (2022)	*16,376*	*2,617*	*1,292*	*Time*	*108*	*49.83551908*
LUNDE A (2016)	*12,251*	*2,348*	*1,176*	*Covariance*	*70*	*47.5*
LYOCSA S (2021)	*16,526*	*2,458*	*1,184*	*Volatility*	*301*	*47.66466891*
HOGA Y (2021)	*11,348*	*2,076*	*996*	*Risk*	*145*	*43.63462781*
LYOCSA S (2020)	*15,534*	*2,023*	*863*	*Volatility*	*347*	*40.54816482*
LI X (2020)	*12,638*	*1,811*	*753*	*Volatility*	*226*	*37.81043674*
TAKAHASHI M (2016)	*16,313*	*2,342*	*1,085*	*Model*	*155*	*45.58594209*
HORTA E (2018)	*9,814*	*1,995*	*960*	*Forecasting*	*47*	*42.82065723*
FENG Y (2015)	*10,001*	*1,557*	*649*	*Memory*	*65*	*35.03123645*
DUARTE C (2017)	*9,696*	*1,841*	*916*	*Midas*	*119*	*41.80478945*
LIU M (2021)	*20,981*	*2,411*	*959*	*HAR*	*420*	*42.797831*
EL OUADGHIRI I (2016)	*17,461*	*2,936*	*1,476*	*Jumps*	*126*	*53.33461144*
RESCHENHOFER E (2020)	*12,517*	*2,031*	*970*	*Loss*	*109*	*43.04826898*

MA F (2019)	*11,398*	*1,970*	*982*	*Model*	*264*	*43.31986011*
LI X (2020)	*13,344*	*1,943*	*860*	*Volatility*	*242*	*40.47589661*
IZZELDIN M (2019)	*8,417*	*2,052*	*1,197*	*Forecasting*	*90*	*47.93107397*
SIRIGNANO J (2019)	*7,745*	*1,467*	*749*	*Model*	*116*	*37.70723447*
GATHERAL J (2018)	*12,138*	*2,044*	*954*	*Volatility*	*211*	*42.68352092*
LIU F (2018)	*10,933*	*2,311*	*1,304*	*Model*	*126*	*50.07102897*
BAYER C (2016)	*9,204*	*1,400*	*634*	*Volatility*	*104*	*34.61249781*
ZHU X (2017)	*12,250*	*2,217*	*1,247*	*Volatility*	*248*	*48.9424669*
ZEVALLOS M (2019)	*3,193*	*962*	*536*	*Volatility*	*65*	*31.74522866*
WANG J (2021)	*13,279*	*1,876*	*913*	*Model*	*245*	*41.73464637*
CAI G (2021)	*16,258*	*3,866*	*2,646*	*Model*	*152*	*71.7478522*
ZHANG Y (2019)	*8,488*	*1,595*	*797*	*Model*	*103*	*38.92806031*
BANULESCU D (2016)	*16,738*	*2,395*	*1,113*	*VaR*	*174*	*46.18315377*
TIAN F (2017)	*7,791*	*1,702*	*900*	*Volatility*	*111*	*41.42935305*
YANG K (2015)	*20,963*	*2,502*	*1,230*	*Forecast*	*188*	*48.60090725*
HUANG X (2022)	*7,980*	*1,669*	*948*	*Volatility*	*144*	*42.54595274*
LIU M (2021)	*17,724*	*2,516*	*1,279*	*GARCH*	*225*	*49.5791459*
MA F (2018)	*12,488*	*2,283*	*1,110*	*Volatility*	*148*	*46.11952886*
BEN OMRANE W (2020)	*12,753*	*2,018*	*880*	*News*	*219*	*40.95533339*
KONG A (2021)	*13,435*	*2,214*	*1,008*	*Jumps*	*115*	*43.90267253*
PAPADAMOU S (2018)	*9,868*	*1,844*	*877*	*Monetary*	*68*	*40.88376774*

BUGGE SA (2016)	*7,602*	*1,574*	*801*	*Volatility*	*113*	*39.02811512*
QU H (2019)	*10,306*	*1,793*	*956*	*Futures*	*98*	*42.72928081*
PHAM BT (2022)	*9,926*	*2,015*	*1,015*	*Connectedness*	*100*	*44.05829557*
MA F (2018)	*11,444*	*2,247*	*1,190*	*Volatility*	*212*	*47.78780585*
PROKOPCZUK M (2016)	*37,783*	*2,504*	*831*	*HAR–RV*	*1,155*	*39.77070026*
CARTEA (2016)	*8,589*	*1,538*	*729*	*Volatility*	*136*	*37.18703969*
PLÍHAL T (2021)	*14,739*	*2,393*	*1,141*	*Volatility*	*346*	*46.77290027*
CHEN W (2020)	*9,797*	*1,837*	*952*	*RV*	*209*	*42.63771305*
CHAN JSK (2019)	*11,409*	*2,316*	*1,304*	*Model*	*110*	*50.07102897*
WANG C (2015)	*9,770*	*2,007*	*1,113*	*Volatility*	*77*	*46.18315377*
CAPORIN M (2015)	*14,181*	*2,311*	*1,090*	*Volatility*	*155*	*45.69314725*
HAMMOUDEH (2015)	*7,173*	*1,660*	*924*	*Risk*	*79*	*41.99127819*
SONG S (2021)	*16,195*	*2,235*	*1,082*	*VaR*	*148*	*45.52150041*
BERGSLI L (2022)	*15,178*	*2,292*	*1,182*	*HAR*	*270*	*47.62355396*
SHEPPARD K (2019)	*16,680*	*2,269*	*1,052*	*Factor*	*211*	*44.87210481*
SMETANINA E (2017)	*20,181*	*2,975*	*1,631*	*GARCH*	*421*	*56.11610981*
HALBLEIB R (2016)	*16,789*	*2,581*	*1,304*	*ARFIMA*	*177*	*50.07102897*
BAILEY G (2019)	*14,977*	*2,441*	*1,225*	*Volatility*	*242*	*48.5*
LYÓCSA L (2020)	*10,071*	*2,048*	*1,020*	*Volatility*	*173*	*44.16912662*
HORPESTAD JB (2019)	*13,587*	*2,374*	*1,095*	*Volatility*	*250*	*45.80010684*

TAN SK (2019)	*11,638*	*2,085*	*1,154*	*Volatility*	*82*	*47.04425044*
MA F (2018)	*13,903*	*2,177*	*1,030*	*Volatility*	*157*	*44.38997687*
RODRÍGUEZ G (2017)	*23,172*	*3,640*	*1,957*	*ARFIMA*	*258*	*61.56396727*
BERNARDI M (2017)	*24,227*	*2,683*	*1,298*	*VaR*	*85*	*49.95341009*
BOTHA B (2021)	*11,910*	*2,163*	*1,023*	*Model*	*133*	*44.23549491*
AVDULAJ K (2017)	*10,031*	*1,879*	*1,022*	*Quantile*	*103*	*44.21338297*
GROSSMASS L (2015)	*18,690*	*2,468*	*1,181*	*Intraday*	*141*	*47.60298345*
GHYSELS E (2019)	*12,969*	*1,610*	*766*	*Returns*	*221*	*38.14396505*
ABEBE TH (2020)	*8,869*	*1,680*	*849*	*Volatility*	*148*	*40.20982892*
PAUL S (2018)	*14,674*	*2,501*	*1,112*	*GARCH*	*131*	*46.16195501*
CAMPANI CH (2018)	*11,679*	*2,238*	*1,130*	*Volatility*	*172*	*46.54208662*
ROKICKA A (2021)	*14,739*	*2,393*	*1,141*	*Volatility*	*346*	*46.77290027*
CHAROENWONG B (2017)	*7,244*	*1,462*	*778*	*Model*	*94*	*38.4493346*
ASLAM F (2021)	*9,056*	*2,110*	*1,161*	*Markets*	*141*	*47.1897292*
JHA KK (2020)	*6,326*	*1,265*	*654*	*Volatility*	*242*	*35.16973873*
CHUONG LUONG CL (2018)	*7,990*	*1,680*	*859*	*Volatility*	*145*	*40.45177921*
ADASCALITEI S (2015)	*3,212*	*980*	*586*	*Volatility*	*45*	*33.23813663*
SANTOS DG (2022)	*24,150*	*2,662*	*1,070*	*Forecasts*	*212*	*45.26283606*
WANG D (2022)	*8,289*	*1,887*	*1,054*	*Volatility*	*147*	*44.9156836*

MAO X (2021)	*5,760*	*1,283*	*671*	*Sentiment*	*99*	*35.6367302*
ANDERSEN TG (2021)	*21,488*	*2,850*	*1,323*	*Cointegration*	*115*	*50.44171459*
BOLLERSLEV T (2021)	*9,552*	*1,866*	*964*	*Model*	*96*	*42.9118435*
LAI YS (2022)	*10,010*	*2,198*	*1,293*	*Models*	*91*	*49.85518656*
CAPORIN M (2021)	*11,505*	*1,703*	*760*	*Jumps*	*150*	*37.99038343*
WANG M (2022)	*8,537*	*1,637*	*814*	*Model*	*86*	*39.35157989*
JIN X (2021)	*13,988*	*2,574*	*1,450*	*Covariance*	*93*	*52.85396921*
					Mean	*45.53162926*

Regarding the articles' authorship, five authors stand out with the highest number of published journals: Ma F, Molnr P, Zhang Y, Chen L, and Lycsa L (Figure 6). Regarding relevance based on the h-index, another five authors are pointed out, especially Ma F and Zhang Y (Figure 7). When observing the number of citations, the most cited authors were Bollerslev and Lunde, with 274 and 108 citations, respectively (Figure 8).

Figure 6

Authors With the Most Articles Published About the Theme

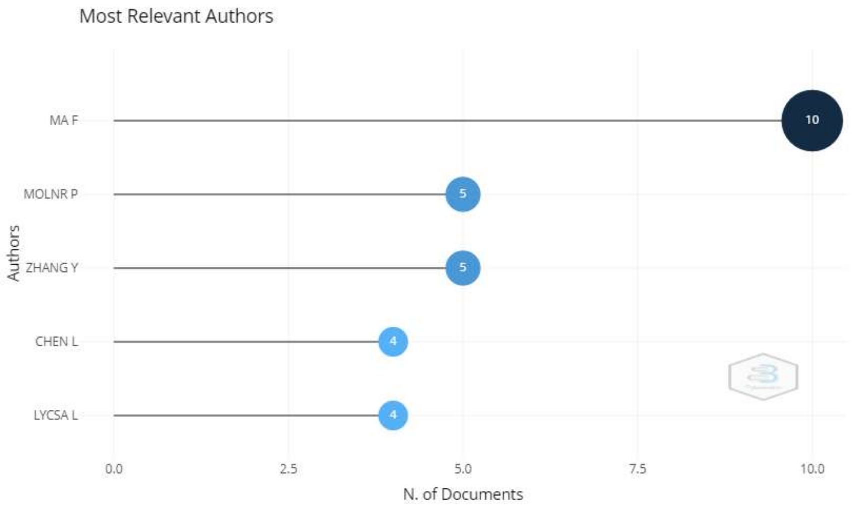


Figure 7

Authors With a Higher Impact on the Theme (Based on the H-Index)

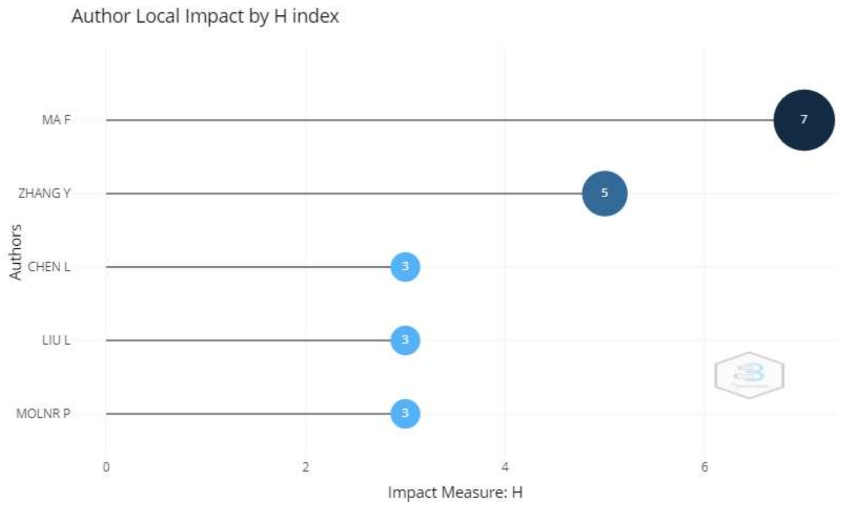
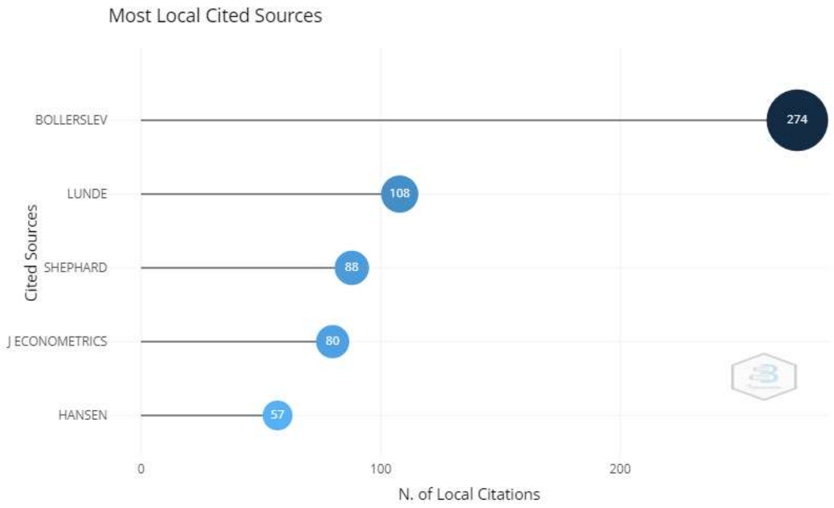


Figure 8

Most Cited Authors Working on the Theme



For Lotka (1926), a small number of authors produce many articles, and their production is equivalent to the production of all other authors together. This law is called the inverse square law (Equation 4).

$$an = a1 / n^2, n = 1, 2, 3 \ldots \qquad (4)$$

Where,

an = number of authors who published n articles a1 = number of authors who published an article

n = number of articles published per author

The numbers provided by Equation 4 cannot be validated for the final sample of the 93 articles. However, a logical relationship indicated low frequency among authors with a more significant number of publications. The author Ma F had the highest number of publications on the subject (ten in total), representing a frequency of 0.40%. The same occurred for Molnr P, who published five articles.

Table 5

Lotka's Law

Written documents	Number of articles	%
1	197	81.70%
2	31	12.90%
3	7	2.90%
4	4	1.70%
5	1	0.40%
10	1	0.40%
Total	**241**	**100.00%**

Table 6 refers to the application of Bradford's Law (1934), indicating the journals with the highest publication incidence. The law states that few journals produce many articles, and many journals produce few articles on a given topic. For Brookes (1969), Bradford's law estimates the degree of relevance of certain academic journals that work in specific areas of knowledge. If the journals are sorted in descending order of productivity, they can be distributed into zones. For this study, these zones were divided into two parts: Zone A, identified as the most essential to the subject, constituted by journals with more than one reference. Zone B presents journals with only one publication.

The most relevant journals based on the h-index were the International Journal of Forecasting and the Journal of Forecasting (with an impact of 6 and 5, respectively) (Figure 9).

Table 6

Bradford's law

Zone	**Journal**	**Individual number**	**Accumulated number**	**Accumulation (%)**
A	JOURNAL OF FORECASTING	14	14	**15%**
	INTERNATIONAL JOURNAL OF FORECASTING	8	22	**24%**
	INTERNATIONAL REVIEW OF ECONOMICS AND FINANCE	6	28	**30%**
	QUANTITATIVE FINANCE	6	34	**37%**
	NORTH AMERICAN JOURNAL OF ECONOMICS AND FINANCE	5	39	**42%**
	ENERGY ECONOMICS	4	43	**46%**
	JOURNAL OF ECONOMETRICS	4	47	**51%**
	JOURNAL OF FINANCIAL ECONOMETRICS	4	51	**55%**
	JOURNAL OF BANKING AND FINANCE	3	54	**58%**
	EMPIRICAL ECONOMICS	2	56	**60%**
	JOURNAL OF EMPIRICAL FINANCE	2	58	**62%**

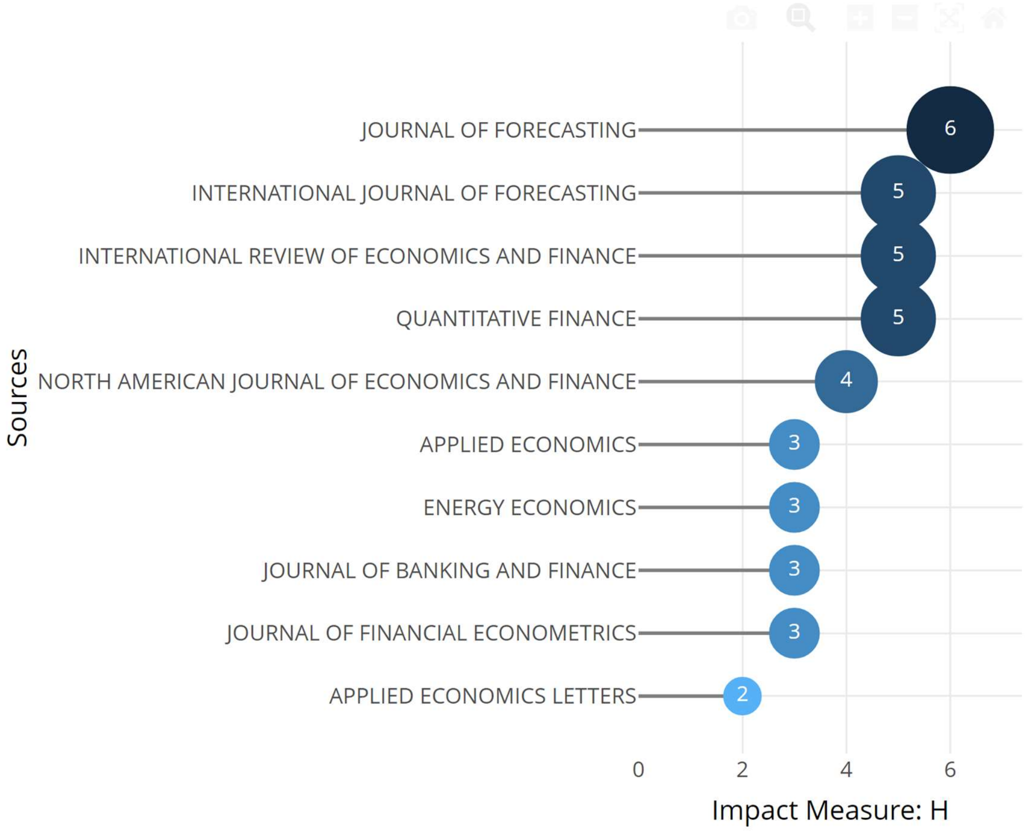
	JOURNAL OF RISK	2	60	**65%**
	JOURNAL OF RISK AND FINANCIAL MANAGEMENT	2	62	**67%**
	STUDIES IN NONLINEAR DYNAMICS AND ECONOMETRICS	2	64	**69%**
B	ANNUAL REVIEW OF FINANCIAL ECONOMICS	1	65	**70%**
	APPLIED ECONOMICS	1	66	**71%**
	ECONOMETRIC REVIEWS	1	67	**72%**
	ECONOMETRICA	1	68	**73%**
	ECONOMETRICS	1	69	**74%**
	ECONOMIC MODELLING	1	70	**75%**
	ECONOMIC RESEARCH-EKONOMSKA ISTRAZIVANJA	1	71	**76%**
	ESTUDOS ECONOMICOS	1	72	**77%**
	EUROPEAN JOURNAL OF FINANCE	1	73	**78%**
	FINANCE RESEARCH LETTERS	1	74	**80%**
	GLOBALIZATION AND HIGHER EDUCATION IN ECONOMICS AND BUSINESS ADMINISTRATION - GEBA 2013	1	75	**81%**
	INTERNATIONAL JOURNAL OF FINANCE AND ECONOMICS	1	76	**82%**

	INTERNATIONAL JOURNAL OF FINANCIAL STUDIES	1	77	**83%**
	INTERNATIONAL JOURNAL OF PRODUCTION ECONOMICS	1	78	**84%**
	INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS	1	79	**85%**
	JOURNAL OF APPLIED ECONOMICS	1	80	**86%**
	JOURNAL OF ASIAN ECONOMICS	1	81	**87%**
	JOURNAL OF BUSINESS \& ECONOMIC STATISTICS	1	82	**88%**
	JOURNAL OF BUSINESS AND ECONOMIC STATISTICS	1	83	**89%**
	JOURNAL OF FUTURES MARKETS	1	84	**90%**
	PACIFIC BASIN FINANCE JOURNAL	1	85	**91%**
	PROCEEDINGS - 2021 INTERNATIONAL CONFERENCE ON COMPUTER, BLOCKCHAIN AND FINANCIAL DEVELOPMENT, CBFD 2021	1	86	**92%**
	RESEARCH IN INTERNATIONAL BUSINESS AND FINANCE	1	87	**94%**

	RESOURCES POLICY	1	88	**95%**
	REVIEW OF ECONOMICS AND STATISTICS	1	89	**96%**
	REVIEW OF FINANCIAL STUDIES	1	90	**97%**
	REVISTA ECONOMIA	1	91	**98%**
	SOUTH AFRICAN JOURNAL OF ECONOMICS	1	92	**99%**
	STUDIES IN ECONOMICS AND FINANCE	1	93	**100%**

Figure 9

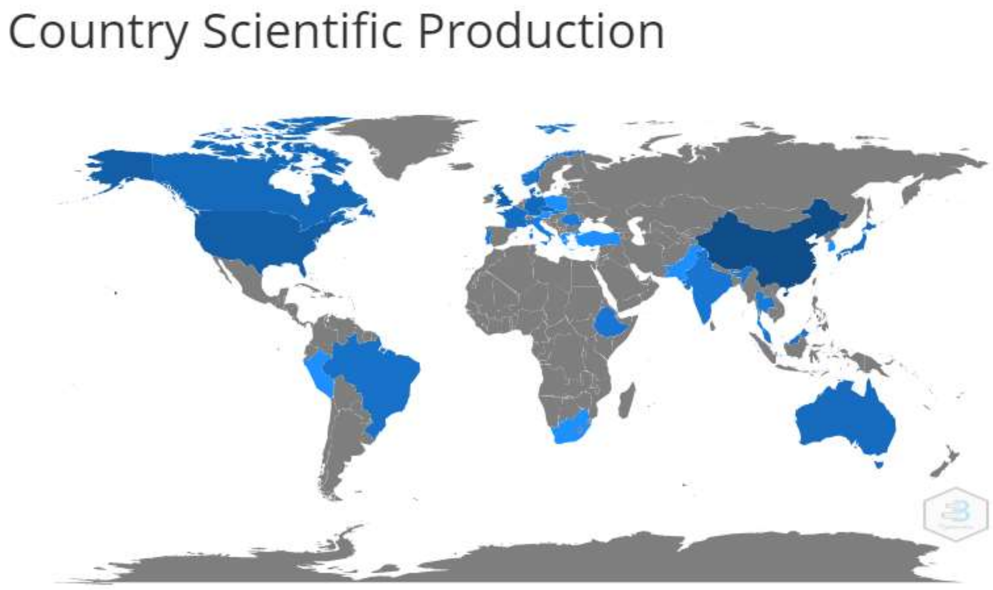
Higher Impact Journals



Finally, we analyzed the number of articles considering the country of the first author's university. The United States and China had the most published articles, while Latin American countries and African countries had only a few studies on the subject. This scenario is similar regarding the number of citations, with a prominence of the US, China, and the UK.

Figure 10

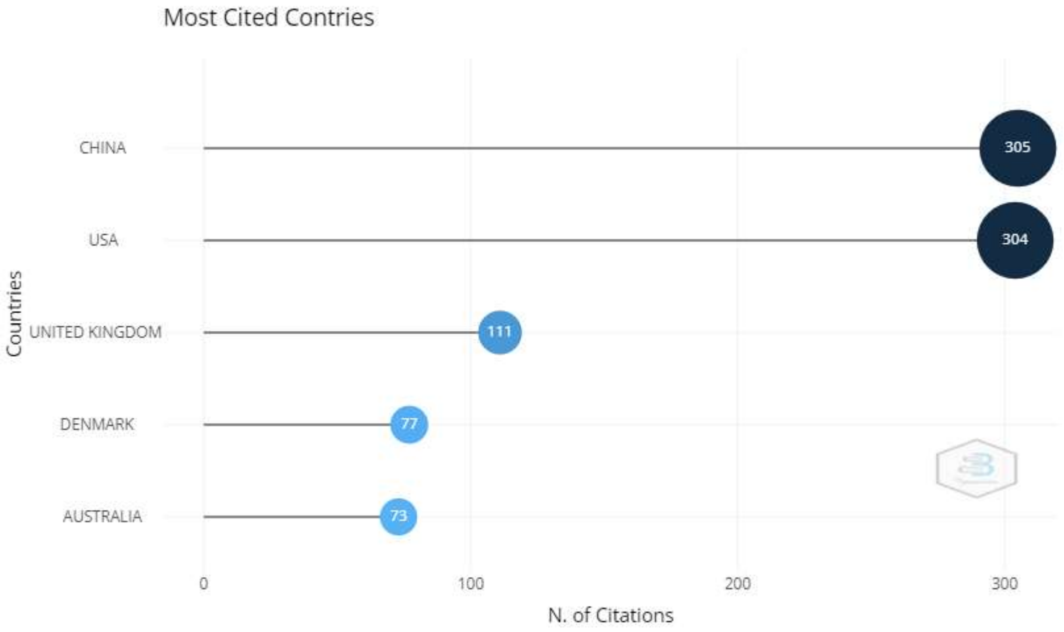
Articles per Country of the First Author's University



Note: Darker colors represent more articles published.

Figure 11

Countries with more citations



Systematic review

The systematic review seeks to identify gaps in knowledge related to the topic addressed in this study. Therefore, in Step 6 of Item 3, a matrix of categories and subcategorization was presented (Table 3) with definitions. For each article of the sample, we identified up to five subcategories per category. The frequency was determined based on the total of subcategories and not the total of articles.

Figure 12 highlights the most and least frequent subcategories to be prioritized in future research. For category 1 – "main theme/focus of the study," what stood out the most in the final sample was the "use of traditional predictive models" (40%), with an opportunity for "Macroeconomic variables influencing the volatility of the asset" (12%). As for "theories related to the hypotheses," the "study of volatility" had the highest recurrence (53%), while the "study of events" was promising and little explored (4%).

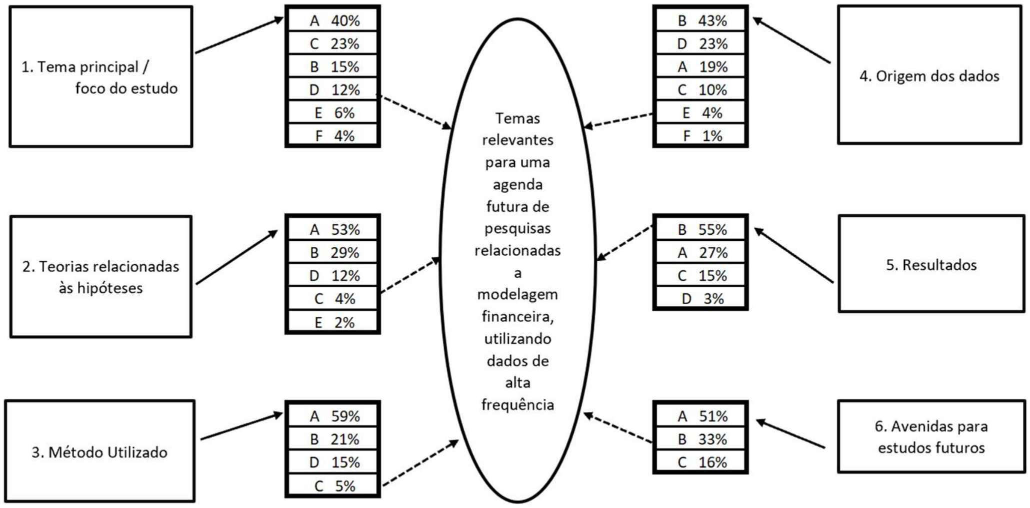
Regarding the category "method," the "autoregressive model" was the most approached by publications (59%), and for the subcategory "analysis of jump," there was a lower occurrence (5%). Given its theoretical basis, it is relevant to consider exploring further the topic "analysis of jump" within the theme addressed in this research. For the "data sources" category, "The United

States and Canada" was the subcategory that stood out (43%), with an opportunity to deepen research for "Latin America" (4%).

Within the "results" category, most of the new (55%) or existing (27%) models and hypotheses were confirmed. Finally, regarding the perspective for future studies, the results showed the relevance of the subcategories in the following order: "alternative to improve the model" (51%); "expanding the use of the model for new purposes" (33%); and "Using the models in new samples" (16%).

Figure 12

Analysis of Categories and Subcategories to Identify a Lack of Knowledge


1. Tema principal / foco do estudo
A 40%
C 23%
B 15%
D 12%
E 6%
F 4%
2. Teorias relacionadas às hipóteses
A 53%
B 29%
D 12%
C 4%
E 2%
3. Método Utilizado
A 59%
B 21%
D 15%
C 5%
Temas relevantes para uma agenda futura de pesquisas relacionadas a modelagem financeira, utilizando dados de alta frequência
B 43%
D 23%
A 19%
C 10%
E 4%
F 1%
4. Origem dos dados
B 55%
A 27%
C 15%
D 3%
5. Resultados
A 51%
B 33%
C 16%
6. Avenidas para estudos futuros

Note:

Subcategory with the higher frequency in each category: →

Subcategory to be prioritized in a future research agenda: ⇢

Conclusion

The development of the financial market favors the emergence of tools focused on managing new information, and models assimilating high-frequency data are important to improve economic agents' decision-making. This study conducted bibliometric analysis and systematic review to identify the journals that published articles addressing the topic, examining studies that used high-frequency data. The research collected articles from Scopus and WoS databases, published between January 2015 and August 2022.

The research results validate Zipt's law regarding the use of keywords, Lotka's law referring to authorship, and Bradford's law about the leading journals publishing articles on the theme. Also, the software Biblioshiny was used to identify the main keywords and level of relationship among these terms within the journals, noting that in many cases, these terms are repeated in the different publications, highlighting their relevance within the subject. As for authorship, it was possible to observe the prominent scholars producing on the theme, considering the frequency they were cited in the articles of the sample.

The analysis of the journals pointed out the relevance of these publications in developing the theme. Most of the articles were produced by authors affiliated with universities in North America and Western Europe, which also have more data available, facilitating this type of analysis.

After the bibliometric analysis, a systematic review of the articles selected was carried out, defining their main themes, theories, methods, data origin, results, and avenues for further research. This type of analysis is important in determining the main gaps in the literature. The method applied suggests a reflection on the influence of the macroeconomic variables on the volatility of assets, as well as deepening the analysis of jump, supported by the availability of high-frequency data.

As for the countries of the first authors' universities, the study demonstrated that countries in Latin America could potentially increase the number of studies using high-frequency data to price financial assets, which are only a few currently. This research seeks to encourage and support new research related to the use of high-frequency data to improve the predictability of financial assets. The analysis offered in this research aims to clarify the main gaps and topics that must be addressed in a future research agenda.

References

Alexander, C.O. (2008). Market Risk Analysis: Quantitative Methods in Finance, vol. 1. New York, John Wiley and Sons.

Andersen, T., & Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. Journal of Empirical Finance 4, 115–158.

Andersen, T., & Bollerslev, T. (1998). DM-dollar volatility: intraday activity patterns, macroeconomic announcements and longer-run dependencies. *Journal of Finance,* 53, 219–265.

Andersen, T., Bollerslev, T., & Lange, S. (1999). Forecasting financial market volatility: sample frequency vis-à-vis forecast horizon. *Journal of Empirical Finance*, 6, 457–477.

Andersen, T., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61, 43–76.

Andersen, T., Bollerslev, T., Diebold, F. X., & Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96, 42–55.

Andersen, T., Bollerslev, T., & Diebold, F. X. (2002). Parametric and nonparametric volatility measurement. In: Hansen, L.P., Aït-Sahalia, Y. (Eds.), Handbook of Financial Econometrics. Amsterdam, North Holland.

Andersen, T., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71, 529–626.

Andersen, T., Bollerslev, T., & Meddahi, N. (2005). Correcting the errors: volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica*, 73, 279–296.

Andersen, T., Bollerslev, T., Christoffersen, P., & Diebold, F. X. (2006). Volatility and correlation forecasting. In: Elliott, G., Granger, C.W.J., & Timmermann, A. (Eds.), Handbook of Economic Forecasting. North Holland Press, Amsterdam.

Angelidis, T., & Degiannakis, S. (2008). Volatility forecasting: intra-day vs. inter-day models. *Journal of International Financial Markets, Institutions and Money*, 18, 449–465.

Black, F. (1975). Fact and fantasy in the use of options. *Financial Analysts Journal*, 31, 36-72.

Boatright, J. R. (2010). In John R. Boatright (Ed.). Ethics in finance in finance ethics: critical issues in theory and practice (Robert W. Kolb Series) (Chapt. 1, pp. 3-20). Chichester: Wiley.

Bollerslev, T., Engle, R. F., & Nelson, D. (1994). ARCH models. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. Elsevier Science, Amsterdam, pp. 2959–3038.

Brookes, B. C. (1969). Bradford’s law and the bibliography of science. *Nature*, 222, 953-956

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7 (2), 174–196.

Corsi, F., Mittnik, S., Pigorsch, C., & Pigorsch, U. (2008). The volatility of realised volatility. *Econometric Reviews*, 27 (1–3), 46–78.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 1143(3), 817-868.

Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative value arbitrage rule. *Review of Financial Studies*, 19, 797-827.

Ghysels, E., & Sinko, A. (2006). Comment. *Journal of Business and Economic Statistics*, 24, 192–194.

Giot, P., & Laurent, S. (2004). Modelling daily value-at-risk using realized volatility and ARCH type models. *Journal of Empirical Finance*, 11, 379–398.

Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics*, 14, 227–238.

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1, 15–39.

Hansen, P.R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics,* 23, 365–380.

Hansen, P. R., & Lunde, A. (2005). A realized variance for the whole day based on intermittent high-frequency data. *Journal of Financial Econometrics*, 3 (4), 525–554.

Hansen, P. R., & Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, 131, 97–121.

Heath, E. (2010). Fairness in financial markets. In John R. Boatright (Eds.), Finance ethics: Critical issues in theory and practice (Robert W. Kolb Series) (Chapt. 9, pp. 163-178). Hoboken: Wiley.

Kahneman, D., Knetsch, J. L., & Thaler, R. (1986a). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, 76, 728-741.

Kahneman, D., Knetsch, J. L., & Thaler, R. (1986b). Fairness and assumptions of economics. *Journal of Business*, 59, S285-S300.

Koopman, S., Jungbacker, B., & Hol, E. (2005). Forecasting daily variability of the S&P100 stock index using historical, realised and implied volatility measurements. *Journal of Empirical Finance,* 12, 445–475.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 317-323.

Lyócsa, S., & Todorova, N. (2020). Trading and non-trading period realized market volatility: Does it matter for forecasting the volatility of US stocks? *International Journal of Forecasting*, 36 (2), 628-645.

McAleer, M., & Medeiros, M. C. (2008a). A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. *Journal of Econometrics,* 147, 104–109.

McAleer, M., & Medeiros, M. C. (2008b). Realized volatility: a review. *Econometric Reviews*, 27 (1), 10–45.

Müller, U. A., Dacorogna, M.M., Davé, R.D., Pictet, O.V., Olsen, R.B., & Ward, J.R. (1993). Fractals and intrinsic time – a challenge to econometricians. In: International AEA Conference on Real Time Econometrics, 14–15 October 1993, Luxembourg.

Saez, M. (1997). Option pricing under stochastic volatility and stochastic interest rate in the Spanish case. *Applied Financial Economics*, 7, 379–394.

Sun, E. W., Chen, Y-T., & Yu, M-T. (2015). Generalized optimal wavelet decomposing algorithm for big financial data. *International Journal of Production Economics*, 165, 194-214.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness. *Psychological Science*, 19(4), 339-347.

Thomakos, D. D., & Wang, T. (2003). Realized volatility in the futures markets. *Journal of Empirical Finance*, 10, 321–353.

Walsh, D. M., & Tsou, G.Y-G. (1998). Forecasting index volatility: sampling interval and non-trading effects. *Applied Financial Economics*, 8, 477, 485.

Zevallos, M. (2019). A Note on Forecasting Daily Peruvian Stock Market Volatility Risk Using Intraday Returns. *Economia*, 42(84), 94-101.