Resources Policy 73 (2021) 102162

Contents lists available at ScienceDirect

Resources Policy

journal homepage: www.elsevier.com/locate/resourpol

Clustering commodity markets in space and time: Clarifying returns, volatility, and trading regimes through unsupervised machine learning

James Ming Chen [a], Mobeen Ur Rehman [b,c,*], Xuan Vinh Vo [d]

[a] Michigan State University, USA
[b] Institute of Business Research, University of Economics Ho Chi Minh City, Viet Nam
[c] South Ural State University, 76, Lenin Prospekt, Chelyabinsk, Russian Federation
[d] Institute of Business Research and CFVG, University of Economics Ho Chi Minh City, Viet Nam

ARTICLE INFO

Keywords:
Commodity markets
Precious metals
Energy markets
Agricultural markets
Machine learning
t-SNE

ABSTRACT

Unsupervised machine learning can interpret logarithmic returns and conditional volatility in commodity markets. This article applies machine learning in order to visualize and interpret log returns and conditional volatility in commodities trading. We emphasize two classes of unsupervised learning methods: clustering and manifold learning for the reduction of dimensionality. We source daily prices from September 18, 2000 through July 31, 2020, for precious metals, base metals), energy commodities and agricultural commodities. Our results highlight that at the very least, returns-based clusters conform more closely to traditional boundaries between precious metals, base metals, fuels, temperate-climate agricultural commodities, and tropical agricultural commodities. On the other hand, volatility-based clustering succeeds in identifying periods of extreme market distress, such as the global financial crisis of 2008–09 and the Covid-19 pandemic.

1. Introduction

Commodity markets represent a quarter of global trade in goods and the most important source of income for some of the world's poorest countries (Cashin and McDermott, 2002; Cashin et al., 2004; Nazlioglu et al., 2013). The increasing financialization of commodity markets (Falkowski, 2011; Silvennoinen and Thorp, 2013) highlights comovement among different commodity markets. Thanks to their hedging properties, commodity futures have become popular tools for diversification against systematic risk (Al-Yahyaee et al., 2020; Rehman et al., 2019). The value of commodities in hedging and diversifying risk, and in providing a safe haven during downturns, extends beyond equities to other asset classes, such as conventional currencies (Liu et al., 2020). The sharp increase in the financialization of commodity markets, however, has substantially escalated the integration of returns. Comovement can reduce the diversification benefits of commodities and increase their vulnerability to external shocks.

Clustering and manifold learning enable us to evaluate commodity markets in space and time. Patterns in price data enable unsupervised machine learning to distinguish among markets for fuels, precious metals, base metals, and agricultural commodities. Logarithmic returns and conditional volatility independently reveal spatial clusters of commodities corresponding closely to human ontologies of these markets. Where machine-generated ontologies diverge from those based on human expertise, however, the mathematical basis for that divergence may signal failures in subjective judgment.

Both spatial and temporal clustering bolster and extend conventional interpretations of commodity markets. Spatial clustering is the simpler task; it might seem trivial to sort a few dozen commodities among the traditional categories of fuels, metals, and agricultural commodities. But it does demonstrate the power of unsupervised learning if a machine, relying solely on returns or volatility, can distinguish base from precious metals, or the tropical crops of coffee, cocoa, and palm oil from temperate-climate crops such as wheat, corn, and soybeans. The resulting spatial ontology of commodity markets can inform efforts to assess and manage risk from comovement or volatility spillover. Temporal clustering is at once more difficult and more far-reaching in its practical implications. Conventional analysis of regime shifts throughout economics often relies upon human evaluations, whose quantitative rigor often subsists in nothing more rigorous than arbitrary benchmarks. If stock prices fall more than 20 percent from the most recent peak, for instance, many financial analysts are prepared to

* Corresponding author. Institute of Business Research, University of Economics Ho Chi Minh City, 59C Nguyen Dinh Chieu Street, District 3, Ho Chi Minh City, Viet Nam.
E-mail addresses: chenjame@law.msu.edu (J.M. Chen), Rehman@ueh.edu.vn (M.U. Rehman), vinhvx@ueh.edu.vn (X.V. Vo).

https://doi.org/10.1016/j.resourpol.2021.102162
Received 27 February 2021; Received in revised form 19 May 2021; Accepted 24 May 2021
Available online 16 June 2021

declare the onset of a bear market. A more mathematically cogent alternative to quibbling over the significance of a 20 percent bear market and a mere 10 percent "correction" may subsist in identifying contiguous or nearly contiguous clusters of trading days where conditional volatility among commodities matches a common pattern.

This paper makes the following contributions towards the mathematical understanding of commodity markets. We demonstrate the application of unsupervised machine learning using clustering and multidimensional scaling, often regarded as an anticipatory step before forecasting, supplies valuable information in its own right. We also distinguish between clusters drawn from logarithmic returns and those drawn from conditional volatility forecasts. The clustering-based ontology of precious metals reflects the structure periodic table, so that gold may be clustered with either silver or platinum, based either on shared membership in a group or period of elements, but never alone with palladium to the exclusion of silver and platinum. The clustering of trading days within a matrix transpose of conditional volatility forecasts reveals distinctive and critical periods corresponding to the global financial crisis of 2008–09 and the recent Covid-19 pandemic. Finally, we apply k-means clustering, hierarchical clustering, MDS followed by *k*-means clustering and *t-SNE* to a matrix of trading dates and their corresponding GJR-GARCH (1, 1, 1) volatility forecasts. The resulting clusters, though ineffective in comprehensively defining all temporally distinct states across two decades of commodities trading, identified the financial crisis of 2008–09 and the moments of greatest panic during the Covid-19 pandemic.

Our results show that clustering based on logarithmic returns outperforms the clustering of conditional volatility forecasts in distinguishing commodity markets from one another. At the very least, returns-based clusters conform more closely to traditional boundaries between precious metals, base metals, fuels, temperate-climate agricultural commodities,[1] and tropical agricultural commodities. On the other hand, volatility-based clustering succeeds in identifying periods of extreme market distress, such as the global financial crisis of 2008–09 and the Covid-19 pandemic. Because these techniques exhibit different strengths, commodity investors should consider subjecting financial data from these markets to both types of unsupervised machine learning.

The rest of this paper is structured as follows: Section 2 reviews the relevant literature. Section 3 presents applied estimations techniques. Section 4 provides detailed analysis along with the interpretation. Finally, section 5 concludes and presents implications of our work.

2. Literature review

2.1. Comovement and volatility spillovers in commodity markets

Comovement affects individual commodities as well as submarkets consisting of several commodities traditionally thought to be related. Volatilities among gold, silver, palladium, and platinum, for instance, have decoupled to such an extent that precious metals can no longer be considered as a single asset class (Batten et al., 2010). Nevertheless, commodities may react differently according to the integration of their returns with other assets. For example, energy markets are more sensitive to external shocks due to their volatile nature and international supply and demand mechanism. Such inconsistencies in the pricing of energy commodities typically induce investors to mix other assets with energy commodities in their portfolios. Among earlier observers of volatility in energy commodity markets, Plourde and Watkins (1998) have suggested that the oil market exhibits more volatility than gold and silver. Zhang and Wei (2010) isolated a strong unidirectional causal relationship running from oil to gold. Precious metal returns may be more sensitive to disaggregated structural oil shocks, which were more pronounced during the global financial crisis of 2008–09 (Rehman et al., 2018). Interdependence among precious metal commodities has generated dynamic asymmetric spillover across returns as well as volatilities (Todorova et al., 2014; Uddin et al., 2019). Such asymmetric spillover arises from negative and positive shocks and intensifies further during financial turbulence.

Though the level of integration has increased across different commodities, portfolio management opportunities persist. During financially turbulent periods, precious metals (particularly gold, platinum, and silver) exhibit hedging and safe haven properties not only against other energy commodities (Baur and McDermott, 2010; Mensi et al., 2015; Rehman and Apergis, 2019), but also vis-à-vis conventional asset classes (Reboredo and Ugolini, 2016). But increasing coherence among different commodity markets remains dynamic and varies across time (see Tang and Xiong, 2012). This is especially true during financial distress.

Such dynamic behavior carries important implications for investors. Comovement among commodities counsels changes in hedging and portfolio rebalancing (Uddin et al., 2019). Furthermore, the increasing popularity of commodity markets among international investors has caused even more comovement among commodities, which ultimately causes more volatility (Cagli et al., 2019; Rehman, 2020). One sector of the commodities market can transmit volatility to another. Recent literature has identified many instances of volatility spillover (Chen, 2010; Demiralay and Ulusoy, 2014; Naeem et al., 2020). These instances include the transmission of volatility from precious metals to equity (Haque and Kouki, 2009; Tang and Xiong, 2010), from oil to precious metals (Hammoudeh et al., 2013; Broadstock and Filis, 2014; Rehman et al., 2018), and precious metals to currency (Sakemoto, 2018). In light of the prominence of oil and export crops in many developing economies, volatility spillovers from energy to agricultural commodities command especially close attention (Du et al., 2011; Nazlioglu et al., 2013; Serra, 2011).

The traditional role of gold as a store of value and as macroeconomic ballast casts a spotlight on the internal dynamics of markets for precious metals. Gold and silver transmit volatility, whereas platinum and palladium receive it (Uddin et al., 2019). Although the volatility of platinum and palladium rose during the global financial crisis, the volatility of gold and silver did not change significantly (Sensoy, 2013). Silver has a downside and upside price spillover on gold (Reboredo and Ugolini, 2015). However, gold exhibits the highest efficiency relative to silver and platinum (Charles et al., 2015). Close integration in returns on precious metals minimizes their diversification benefits (Rehman and Vo, 2020). Observations of the cobalt, copper, and nickel markets have revealed the possibility of spillovers among base metals (Martino and Parson, 2013).

In sum, the literature on comovement and volatility spillovers sends mixed messages. On one hand, commodity markets offer diversification benefits to investors. On the other hand, increasing levels of integration within commodity markets heightens the need to understand those markets' structural relationships and dynamic interdependencies.

2.2. Unsupervised machine learning

In the extensive literature applying unsupervised learning to financial markets and mathematically cognate fields of economics and natural language processing, two works bear an especially close relationship to this study. Fernández-Avilés et al. (2020) directly inspired this article. Their application of multidimensional scaling to expected shortfall as a measure of extreme downside risk in commodity markets motivated us to apply clustering methods as well as manifold learning. Among other things, Fernández-Avilés et al. (2020) examined comovement among commodity markets during known periods of crisis.

Within the clustering literature, Münnix et al. (2012) undertook the most explicit step toward identification of *temporal* as well as spatial relationships in financial data. The Münnix study used hierarchical clustering to identify patterns of correlation structures similar enough to

[1] See Spencer et al. (2018); Cabrera and Schulz, (2016).

comprise distinct market states. It approaches the Fernández-Avilés study in its impact on our thinking.

Among the many clustering algorithms used in economics and finance (D'Urso et al., 2016; Kou et al., 2014; Musmeci et al., 2015; Pattarin et al., 2004), we focus on *k*-means and hierarchical clustering. One of the oldest clustering algorithms (MacQueen, 1967), *k*-means clustering remains one of the most popular (Soni and Patel, 2017, p. 901). *k*-means clustering is often used to forecast returns and manage investment risk (Cai et al., 2016; Kou et al., 2014; Nanda et al., 2010; Xu et al., 2020). Designed to partition mathematical space, *k*-means clustering is particularly useful for detecting fraud and other outliers (Deng and Mei, 2009), such as firms at risk of default or failure (Tsai, 2014).

k-means clustering does suffer from certain drawbacks. In particular, the ideal value of *k*, or the optimal number of clusters, is not known in advance. Finding the optimal value of *k* is a special instance of the broader problem of finding the ideal number of clusters in unsupervised learning (Jain et al., 1999; Xu et al., 2016). Three methods for finding optimal *k* are especially popular: elbow (Bholowalia & Kumar, 2014; Kodinarwa and Makwana, 2013), silhouette (Lengyel and Botta-Dukát, 2019; Rousseeuw, 1987), and gap (Tibshirani et al., 2001).

Hierarchical clustering methods decompose and arrange sets of mathematical objects according to dendrograms, or trees expressing phylogenetic relationships (Bouguettaya et al., 2015; Davidson and Ravi, 2005; Day, 1984; Gil-Garcia et al., 2006; Manning et al., 2008, pp. 346–368). The agglomerative method begins from the "bottom" of a dataset and combines instances into clusters until all data has been assigned to a single, overarching cluster (Murtagh, 1983). Agglomeration is less computationally demanding than the divisive or "top-down" approach (Ishikaza, Lokman & Tasiou, 2020; Roux, 2018). Four methods for computing distances and similarities in hierarchical clustering are widely used: Ward's method and single-, average-, and complete-linkage (Blashfield, 1976; Kuiper and Fisher, 1975; Milligan, 1980; Saraçli, Doğan, Doğan, 2013; Vijaya et al., 2019).

Because hierarchal clustering yields determinative answers without requiring stochastic instantiation or the often imprecise (and always inconvenient) exercise of finding the optimal number of clusters, this class of clustering methods is widely used in economics and finance (Marti et al., 2020). Hierarchical clustering has evaluated everything from ordinary stock markets (Micciche et al., 2005; Puerto et al., 2020; Tumminello et al., 2010) to buildings and real estate (Hepsen and Vatansever, 2012; Li et al., 2019), broader financial indicators (Kumar and Deo, 2012), and the relationship between financial markets and the real economy (Musmeci et al., 2015). Cognate fields of data science also apply unsupervised learning. In natural language processing. *k*-means clustering can spot semantic similarity and summarize texts (Agrawal and Gupta, 2014; Jain et al., 2012), Lin and Wu (2009); Wazarkar and Mahjrekar (2014). Hierarchical clustering is an especially popular method for topic modeling (Duch et al., 2008; Forster, 2006; Fung et al., 2009; Huang et al., 2011).

3. Data and methodology

3.1. Data

We use a wide array of assets sampled from the group of precious metals, base metals, energy and agricultural based commodities. We extracted daily prices from September 18, 2000 through July 31, 2020, for gold, silver, platinum, palladium (precious metal commodities), copper, zinc, tin, lead, nickel, aluminum (base metal commodities), Brent, West Texas intermediate crude (WTI), Gasoil, Gasoline (energy commodities) and palm oil, wheat, corn, soybeans, coffee, cocoa, cotton, lumber (agricultural commodities). Transforming daily prices into logarithmic returns shortened all series by a single day at the start. Our log return data (as well as the conditional volatility data derived from log returns) therefore covered the period from September 19, 2000, to July 31, 2020. Pricing data for all of our sampled commodities is extracted from Thomson Reuters DataStream.

3.2. Methodology

This study uses two forms of unsupervised machine learning: clustering and manifold learning. Clustering can reveal relationships within financial time series. To evaluate these relationships' most salient attributes and to reduction their dimensionality so that vital financial data can be visually interpreted, we deploy two forms of manifold learning, each aimed at reducing high-dimension objects to exactly three dimensions.

3.2.1. Unsupervised machine learning

Despite superficial differences between financial analysis and natural language processing, both financial and linguistic data are both amenable to unsupervised learning. If anything, financial data such as logarithmic returns and conditional volatility forecasts in commodity markets are more easily clustered and subjected to manifold learning. Financial data are much denser than their linguistic counterparts and can be gathered so that each tradable asset (whether a stock or a commodity) has exactly the same number of observations covering the same stretch of trading days. Moreover, financial data inherently take numeric form. Their conversion to quantities such as log returns or conditional volatility is at relatively easy. By contrast, the vectorization of verbal documents, never a trivial task, generates sparse matrices that do not readily reveal their semantic content. Discretionary decisions over stop words and the substitution of *n*-grams for single words never arise in finance. For these reasons, clustering and manifold methods that work in natural language processing are almost assuredly feasible in finance. In principle, a single study of financial markets could apply clustering methods not only to economic data, but also to news related to specific companies, sectors, or topics (Dai et al., 2010). At a minimum, the literature germane to unsupervised learning in economics and finance should be deemed to include applications of clustering and manifold learning in natural language processing.

3.2.2. k-means clustering

k-means clustering is highly sensitive to outliers. This trait, which makes it attractive in applications such as fraud detection, may undermine other uses of the algorithm. The use of sampling methods in variants such as minibatch *k*-means exposes the algorithm's dependence on randomized instantiation (Capó et al., 2017). Replicability of results requires special care. Finally, the *k*-means algorithm is not suitable for finding objects unless they assume a hyper-ellipsoidal shape (Kaushik and Mathur, 2014, pp. 95–96).

3.2.3. Hierarchical clustering

The application of hierarchical clustering to cryptocurrency markets (Song et al., 2019) warrants close attention. Investor behavior, if not actual returns, may give rise to hedging, diversification, and safe-haven effects in cryptocurrency and precious metals as asset classes (Andriano and Diputra, 2017; Corbet et al., 2020; Rehman and Vo, 2020; Selmi et al., 2018). Whether returns on cryptocurrency actually exacerbate rather than hedge downside risk during market turbulence is a different question. Evidence that Bitcoin has declined in lockstep with the S&P 500 during the Covid-19 pandemic heightens the urgency of the research question (Conlon and McGee, 2020). In light of these phenomena, clustering analysis of cryptocurrency should be treated as analytical next-of-kin to this study's evaluation of precious metals and other commodities.

These techniques depend on the arrangement of n time series expressing either log returns or conditional volatility into a two-dimensional, $m \times n$ matrix whose rows represent trading dates and whose columns designate individual commodities. It is trivially easy to transpose such a matrix so that it consists of n rows representing commodities and m columns representing log returns or conditional

volatility on individual trading days. Applying clustering and manifold learning to this transposed, $n \times m$ matrix reveals *temporal* clusters corresponding to mathematically distinctive regimes in commodity trading.

3.2.4. Multidimensional scaling

Finally, we use two different manifold methods for dimensionality reduction and visualization. First, multidimensional scaling (MDS) preserves distances among observations, even as it reduces commodity return or conditional volatility series as high-dimensional objects into three dimensions for visualization (Cox and Cox, 2008; Hout et al., 2013).

Second, *t*-distributed stochastic neighbor embedding (*t*-SNE) (van der Maaten, 2009, 2014; van der Maaten & Hinton, 2008, 2012) enables us to visualize trading days as clusters. *t*-SNE attempts to reduce distances between similar instances and to maintain distances between dissimilar instances. It is especially popular in natural language processing (Chan et al., 2018; Wang et al., 2018; Zech et al., 2018). Although manifold learning is often combined with clustering, it is a valuable form of unsupervised learning in its own right. At least one application of supervised learning using convolutional neural networks to classify images has deployed *t*-SNE during preprocessing to detect and remove outliers (Perez and Tah, 2020). Such an application of *t*-SNE is conceptually equivalent to the reduction of dimensionality through principal component analysis (Abdi and Williams, 2010; Wold et al., 1987).

3.2.5. GJR-GARCH modelling

We elected to represent conditional volatility as a GJR-GARCH(1, 1, 1) process using Student's *t* distribution so that our results could be compared more readily with Fernández-Avíles, Montero & Sanchis-Marco (2020), pp. 1212–1213, which constitutes the most extensive application of unsupervised machine learning to commodity markets in the literature. The GJR-GARCH(1, 1, 1) model with a constant mean has been a familiar, longstanding model for at least three decades (Bollerlev and Woolridge, 1992). Its mathematical assumptions are well understood and exhaustively documented (Alexander et al., 2021). GJR-GARCH models outperform alternatives such as GARCH-M, GARCH, and log-GARCH in fitting empirical market data (Nugroho et al., 2019). Finally, reliance upon Student's *t* distribution generates an estimate of the degrees of freedom. Our result of $\nu \approx 6.8175$ degrees of freedom indicates a mild but manageable amount of kurtosis in conditional volatility forecasts for nickel.

4. Analysis and discussion

4.1. Preliminary analysis

The raw, unscaled price data in Fig. 1 for our sampled commodities highlights the nature of our sampled commodities. Fig. 2 plots cumulative log returns in a uniform three dimensional view for each commodity. These series of logarithmic returns easily lend themselves to conventional descriptive statistics.

Fig. 3 reports kurtosis on a logarithmic scale so that all four of the first central moments of the distribution of log returns can be depicted on a comparable scale. The very high values for kurtosis associated with certain commodities are more readily seen when kurtosis is reported in raw terms, without the logarithmic transformation.

In Fig. 4, only four groups of commodities exhibit correlations approaching or exceeding 0.500. The four precious metals: gold, silver, platinum, and palladium; six base metals: copper, zinc, tin, lead, nickel, and aluminum; four fuels: Brent, WTI, gasoil, and gasoline; and three agricultural commodities — wheat, corn, and soybeans — all of which are temperate-climate crops, as opposed to crops associated with subtropical or tropical climates, such as palm oil, coffee, and cocoa.

In Table 1, we calculated the conditional, time-variant volatility for all 22 commodities according to a GJR-GARCH(1, 1, 1) process using Student's *t* distribution (Fernández-Avíles, Montero & Sanchis-Marco, 2020, pp. 1212–1213).

In some salient instances, differences among these 22 conditional volatility series during critical trading periods are visible. Volatility spikes for precious metals, fuels, and two agricultural commodities that contribute to fuel markets (corn and palm oil) are especially pronounced during the outbreak of the initial spread of the Covid-19 pandemic beyond China and Italy. We devote much of this paper to more mathematically exacting and precise evaluation of time-varying differences in volatility. Moreover, the application of clustering and manifold learning to the transpose of that matrix isolates the Covid-19 pandemic as one of the two most distinctive periods of commodities trading during the past two decades.

In Fig. 5, by analogy to the corresponding overview of logarithmic returns, we present this three-dimensional plot depicting each volatility series in a different color:

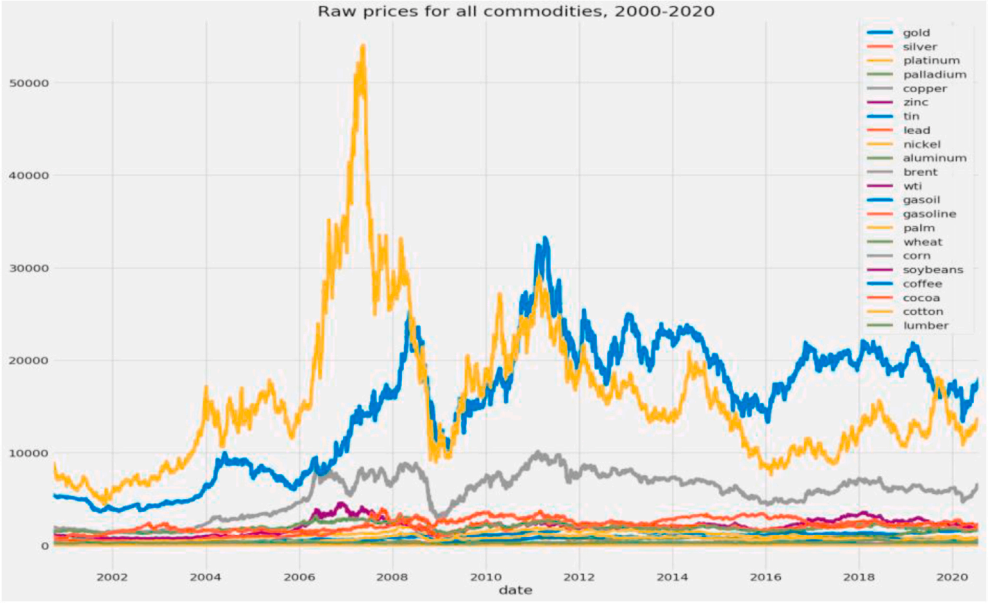

Fig. 1. Commodities pricing.

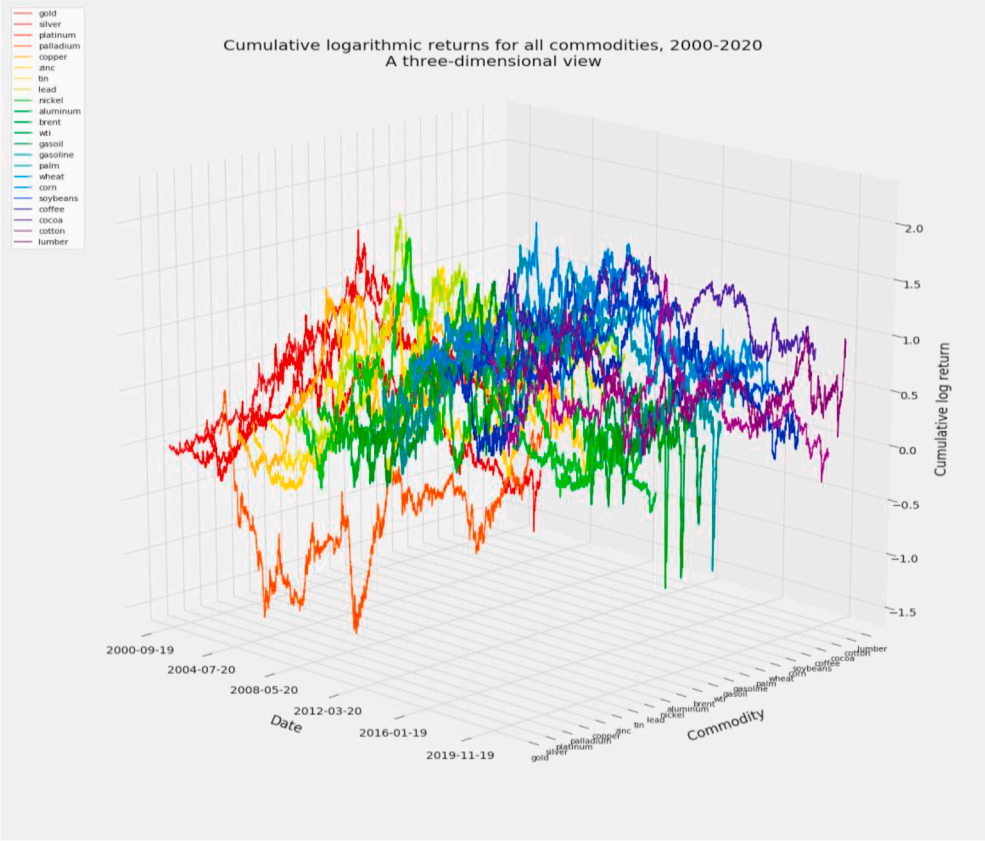

Fig. 2. Cumulative log returns.

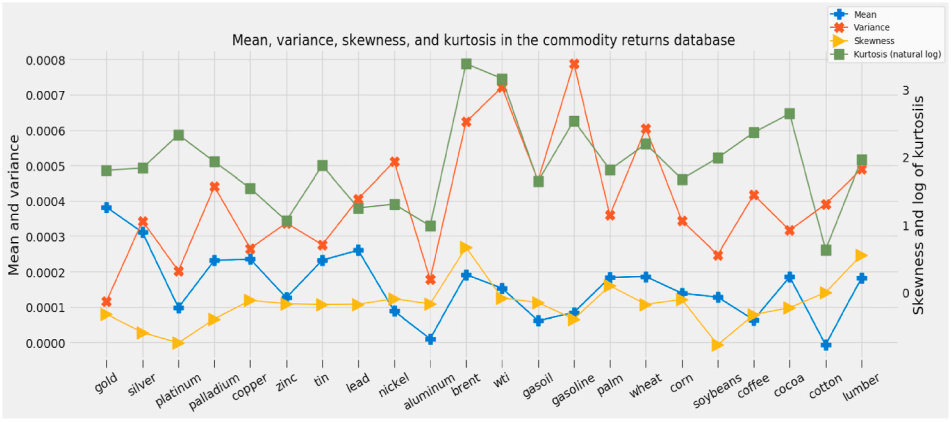

Fig. 3. Descriptive statistics.

Fig. 6 shows that correlations among conditional volatility for each of the commodities are higher, as a group, than correlations among logarithmic returns:

Because expected shortfall does add descriptive value, we completed this step (see Fig. 7) in accordance with Fernández-Avíles, Montero, and Sanchis-Marco (2020) as a component of exploratory data analysis. Expected shortfall indicates each commodity's vulnerability to extreme downside risk. The corresponding degrees of freedom, or ν, in the t-distribution best fitting each commodity are roughly inverse to expected shortfall. Therefore, the ν parameter is closely related to kurtosis (Rozga and Arnerić, 2013, p. 34).

The foregoing summary of expected shortfall and degrees of freedom reveals the tail risk associated with each commodity. Aluminum, for instance, is associated with the lowest expected shortfall— and with the highest number of degrees of freedom for the fitted t-distribution. The relative thinness of tail risk for aluminum follows intuitively from the insight that ν increases as a distribution approaches the normal distribution. Low values of ν are associated with commodities such as silver, tin, and palm oil, all of which have relatively high amounts of expected shortfall.

At just under 2.5, the lowest number of degrees of freedom in this study (tin) remains safely above the threshold of $\nu = 1$, at which point Student's t-distribution is equivalent to the Cauchy distribution. That pathological distribution lacks finite moments of any order greater than or equal to one or, for that matter, a moment-generating function (Balakrishnan and Nevrozov, 2003, p. 305; Feller, 1971, p. 704; Mahdizadeh and Zamanzade, 2019; Riley et al., 2006, p. 1333) — at least one obtained without truncation of the distribution or the dire expedient of

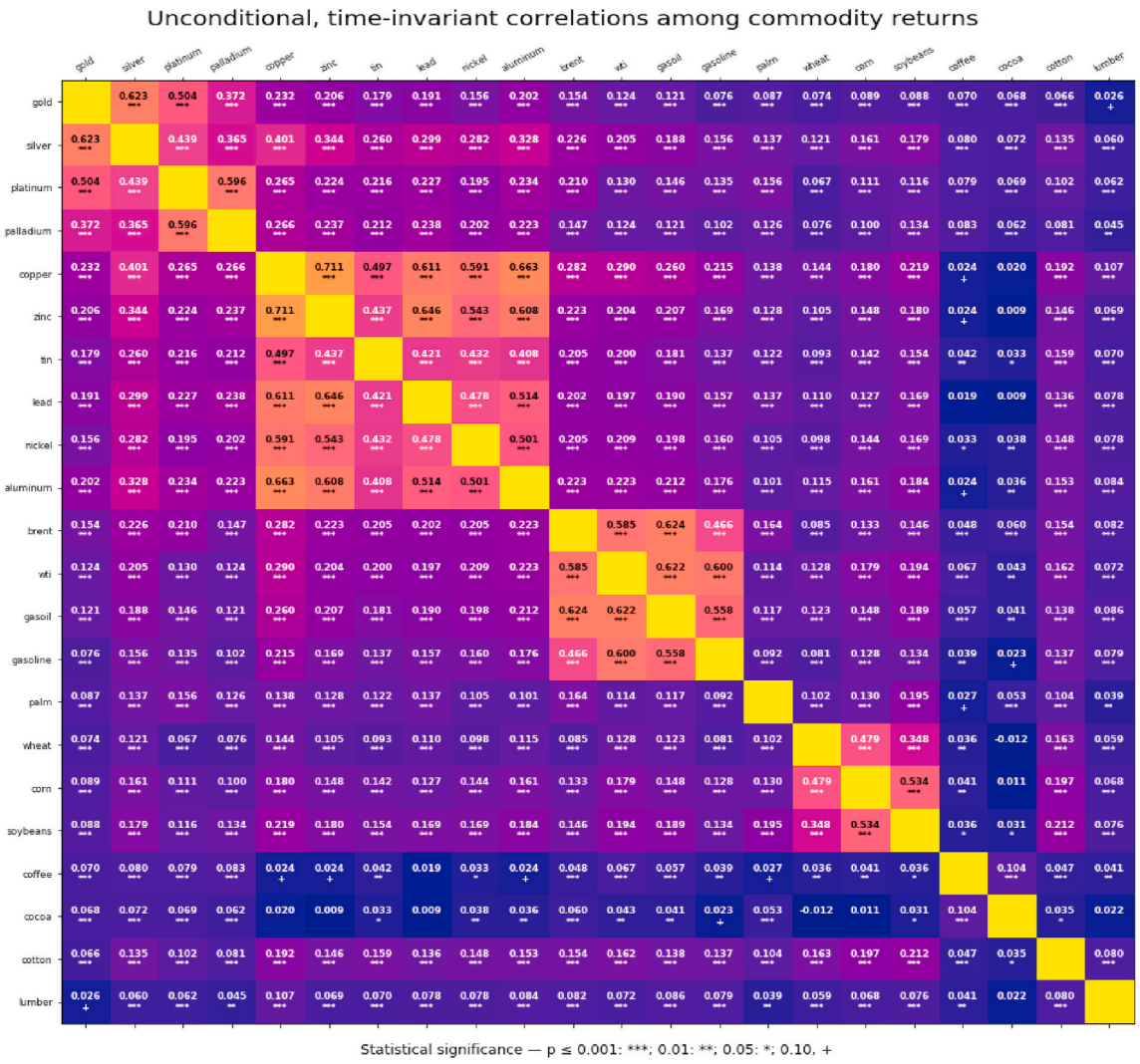

Fig. 4. Unconditional time-invariant correlation.

transforming sample values from a Cauchy distribution so that they are distributed on the interval [-1, 1] (Ohakwe and Osu, 2011).

In other words, commodities trading data, falls far short of the frontier at which conventional statistics and unsupervised machine learning might fail to decipher the economic secrets of these closely scrutinized markets.

4.2. *Spatial clustering by commodity: k-means*

The use of a GJR-GARCH(1, 1, 1) process to forecast conditional volatility for all commodities creates two distinct mathematical objects. For each commodity, the series of log returns or conditional volatility comprises a single vector from the first trading date in our study to the last. Let each trading date represent a row in a two-dimensional matrix, while each commodity represents a column. We now have two $m \times n$ matrices that permit clustering analysis. One covers conditional volatility; the other, logarithmic returns.

This section will subject each of these matrices to two different types of clustering. First, we will apply *k*-means clustering to conditional volatility and log returns. That clustering algorithm depends on distances among volatility or log return series for different commodities. *k*-means clustering partitions commodities into different sectors of financial space. Multidimensional scaling (MDS) then projects these clusters and their centroids into three dimensions capable of human perception and interpretation.

4.2.1. *k-means clustering plus multidimensional scaling of conditional volatility*

In Fig. 8, among cluster centroids, clusters 1 and 2 are striking opposites. Cluster 1 attained a single, distinctive peak in nearly two decades of trading. Volatility peaked during the financial crisis of 2008–09. By contrast, although volatility for cluster 2 also rose during that crisis, that cluster reached a truly impressive peak during the first wave of the Covid-19 pandemic in 2020. Volatility for cluster 2 remained subdued throughout 2020 (see Fig. 9).

Application of MDS places these *k*-means clusters into an easily visualized three-dimensional space and shows the relationship of each commodity to the seven *k*-means centroids: Among the seven clusters, three are singletons representing only one commodity: gasoline (4), wheat (5), and palm oil (6). The four other clusters, which conveniently came first in the sequence from 0 through 6, form an intriguing quartet:

0: cocoa, coffee, lumber, cotton; palladium.

1: corn, soybeans; tin, copper, zinc, lead, nickel; silver.

2: Brent, gasoil, WTI.

3: gold, platinum; aluminum.

Only cluster 2 corresponds to a traditional category of commodities. Even then, Brent, gasoil, and WTI are the only fuels in cluster 2; conditional volatility for gasoline is different enough from that of its fellow fuels to warrant a distinct cluster. Cluster 0 captures the tropical agricultural commodities of cocoa and coffee. Meanwhile, the temperate-climate crops of corn and soybeans (but not wheat) belong to cluster

Table 1
GJR-GARCH model results.

Dep. Variable:	nickel		
		R-squared:	0.000
Mean Model:	Zero Mean	Adj. R-squared:	0.000
Vol Model:		GJR-GARCH	
		Log-Likelihood:	−11084.2
Distribution			
Standardized Student's t		22178.5	
AIC:			
Method:			
Maximum Likelihood		BIC:	22211.2
No. Observations:		5182	
Df Residuals: 5177			
Df Model:		5	
Volatility Model			
Coefficient	*Std. err*	*t*	*P>|t|*	*95.0% Conf. Int.*
Omega	0.0474	1.890e-02	2.507	1.216e-02 [1.035e-02,8.444e-02]
alpha [1]	0.0412	9.343e-03	4.413	1.019e-05 [2.292e-02,5.954e-02]
gamma [1]	1.9531e-03	8.061e-03	0.242	0.809 [-1.385e-02,1.775e-02]
beta [1]	0.9485	1.172e-02	80.943	0.000 [0.925, 0.971]
Distribution				
coefficient	*Std. err*	*t*	*P>|t|*	*95.0% Conf. Int.*
Nu	6.8175	0.602	11.327	9.640e-30 [5.638, 7.997]

1, alongside four base metals and silver.

Arguably the greatest surprise involves the traditional quartet of precious metals: gold, silver, platinum, and palladium. The lighter period 5 metals of palladium and silver could be expected to be separated from period 6 platinum and gold. But palladium and silver are each closer to a different centroid. Less surprising, perhaps, is the alignment of most of the base metals — tin, copper, zinc, lead, and nickel — with silver, while the more energy-intensive aluminum is assigned to a cluster alongside gold and platinum.

4.2.2. *k-means clustering plus multidimensional scaling of logarithmic returns*

We project onto three dimensions through MDS, those cluster centroids appear to exert a financial form of gravitational force on familiar groupings of commodities. These eight clusters are divided into four singletons and four multimember categories:

0: gasoline.
1: gold, silver, platinum, palladium, cocoa.
2: lumber.
3: copper, zinc, tin, lead, nickel, aluminum.
4: Brent, WTI, gasoil.
5: wheat, corn, soybeans, cotton.
6: palm oil.
7: coffee.

The log return clusters make an immediate impression upon anyone familiar with commodity markets. They seem much more intuitively correct than the *k*-means clusters generated by conditional volatility. The four multimember clusters each correspond to a traditional category: (1) precious metals, (3) base metals, (4) fuels, and (5) temperate-climate commodities. The only "intruder" is the inclusion of cocoa with precious metals in cluster 1. The mathematical isolation of gasoline, lumber, palm oil, and coffee is generally consistent with the treatment of these commodities by *k*-means clustering using conditional volatility.

4.3. *Spatial clustering by commodity: hierarchical clustering*

We apply hierarchical agglomerative clustering (HAC) to the conditional volatility and log return matrices. HAC generates dendrograms whose branches represent distinctive clusters of commodities. The

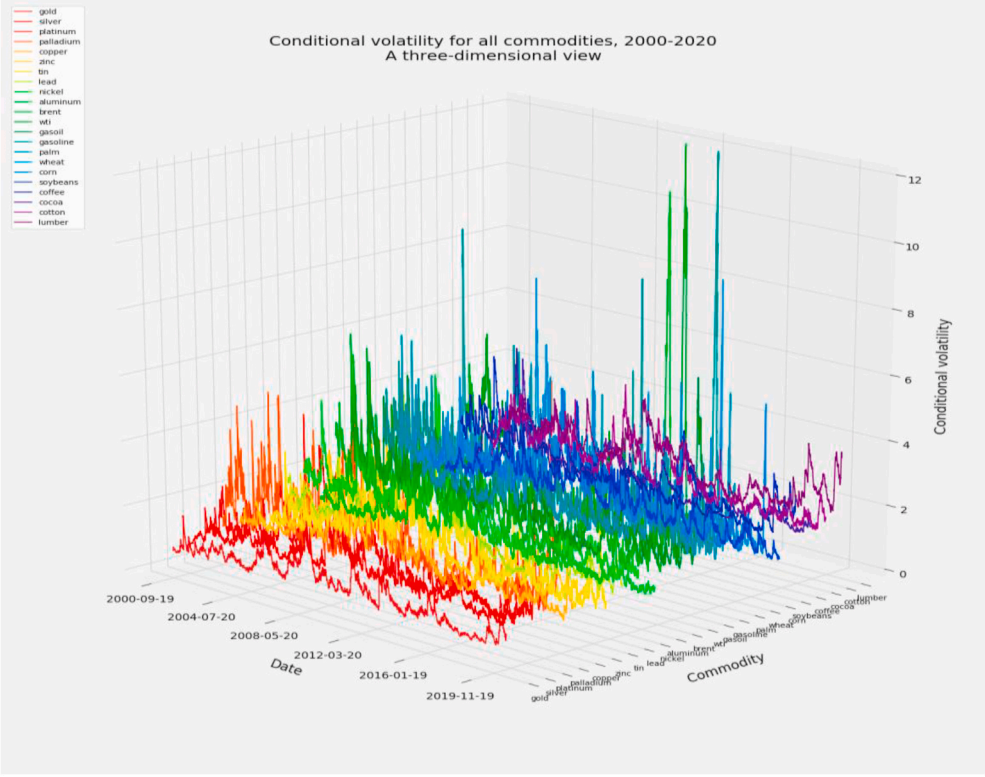

Fig. 5. Conditional volatility.

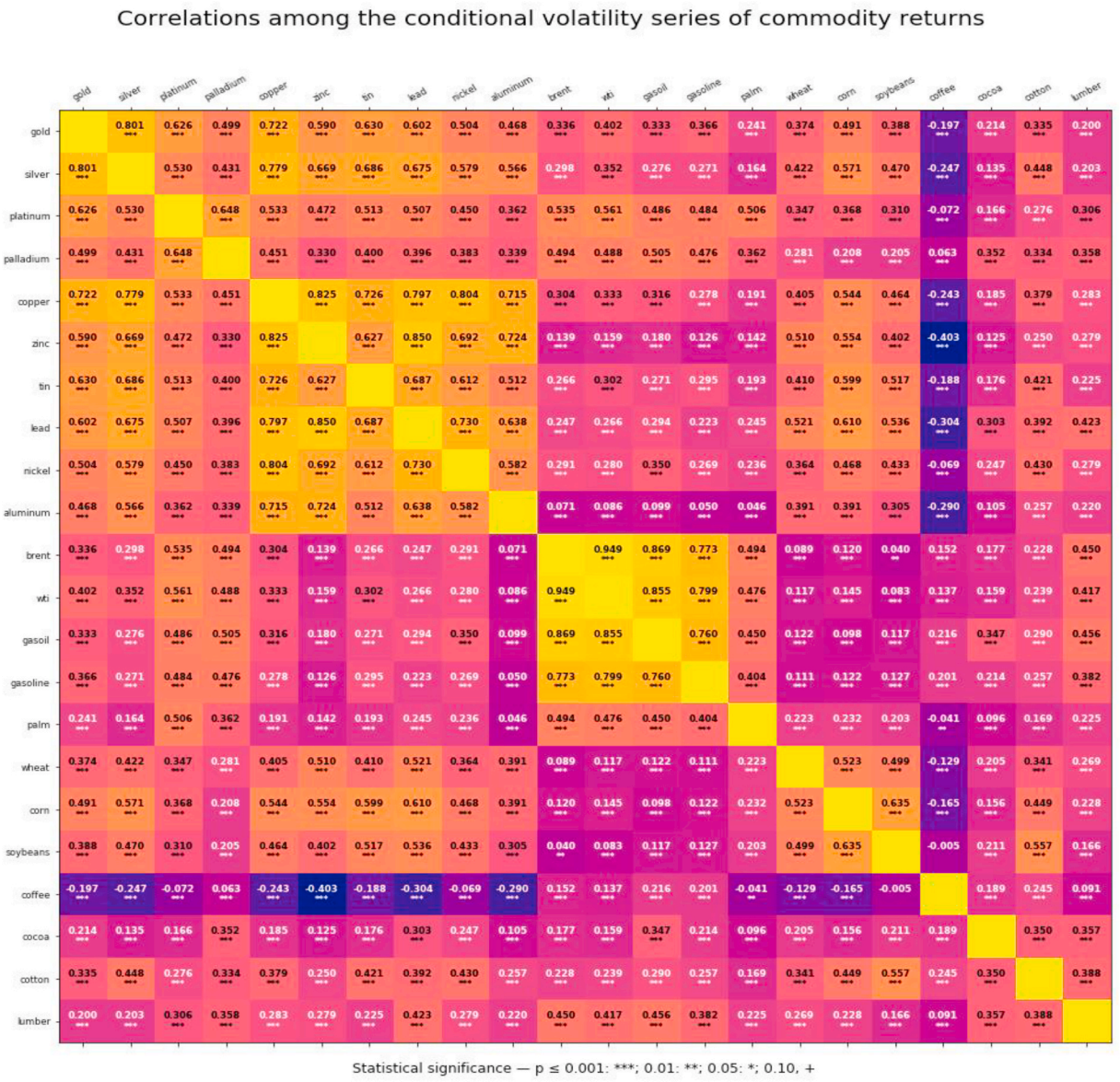

Fig. 6. Correlation-conditional volatility.

length of each branch conveys information on the distances among volatility or log return series. Since dendrograms report distances between clusters and subclusters along a single dimension, further reduction of dimensionality through MDS or any other manifold learning method is not only unnecessary, but also impossible.

Hierarchical clustering boasts at least two advantages over *k*-means. First, it requires no preliminary inquiry into the optimal number of clusters. Second, unlike the inescapably random nature of the *k*-means algorithm, hierarchical clustering generates determinate answers. Reproducible results do not require the setting of a random number seed. Python's SciPy library implements hierarchical agglomerative clustering and its visualization through dendrograms.

We begin with HAC dendrograms based on conditional volatility. This illustration of Ward's method demonstrates how a dendrogram reveals ways to generate a different number of clusters.

4.3.1. *Hierarchical agglomerative clustering of conditional volatility*

In Figs. 10–13, a dendrogram's vertical axis reports distances among clusters. That value can be interpreted in a fashion akin to genetic distance among related species within a phylogenetic tree in evolutionary biology (Nakhleh, 2013; Woese, 2000). We begin with HAC dendrograms based on conditional volatility. This illustration of Ward's method demonstrates how a dendrogram reveals ways to generate a different number of clusters. A horizonal line drawn so that it crosses the dendrogram at a cluster distance of 75 crosses five vertical lines. Consequently, this interpretation of the dendrogram for Ward's method of hierarchical agglomerative clustering using conditional volatility data generates five clusters. From left to right:

0: gasoline, gasoil, Brent, WTI.
1: wheat.
2: nickel, coffee, cocoa, cotton, lumber/palladium, palm oil.
3: platinum, gold, aluminum.
4: corn, soybeans/zinc, lead/silver, copper, tin.

The slashes within each of those cluster descriptions indicates further sub-clusters that may be inferred within clusters 2 and 4. In the case of cluster 4, the sub-cluster consisting of corn and soybeans stands apart from the (mostly) base metals comprising the rest of the clusters.

As tempting as it may be to separate corn and soybeans from the metals in cluster 6, the dendrogram shows how such a move, among other things, would also split gasoil from Brent and WTI in cluster 1 and

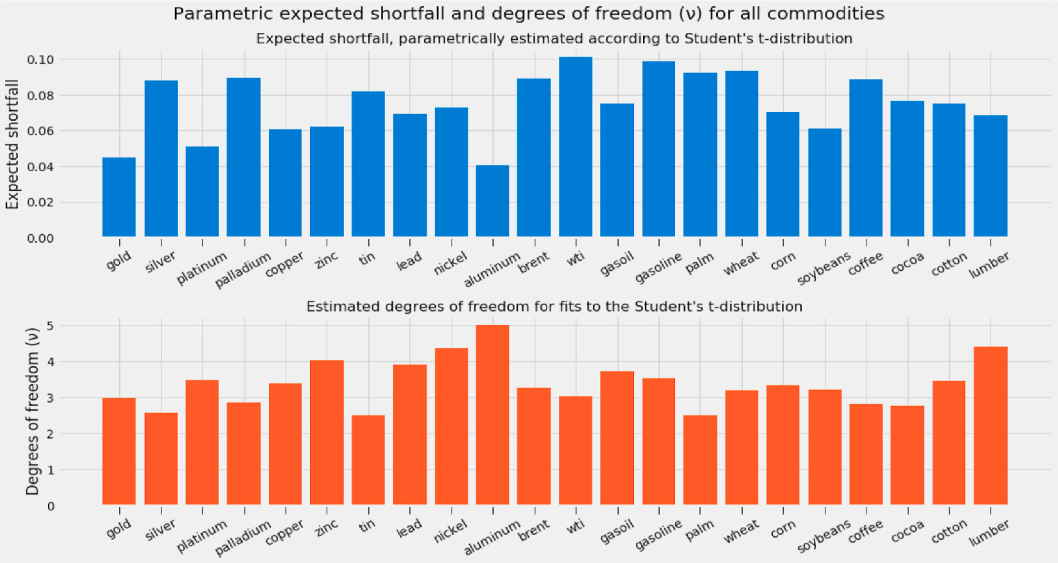

Fig. 7. Expected shortfall with degrees of freedom.

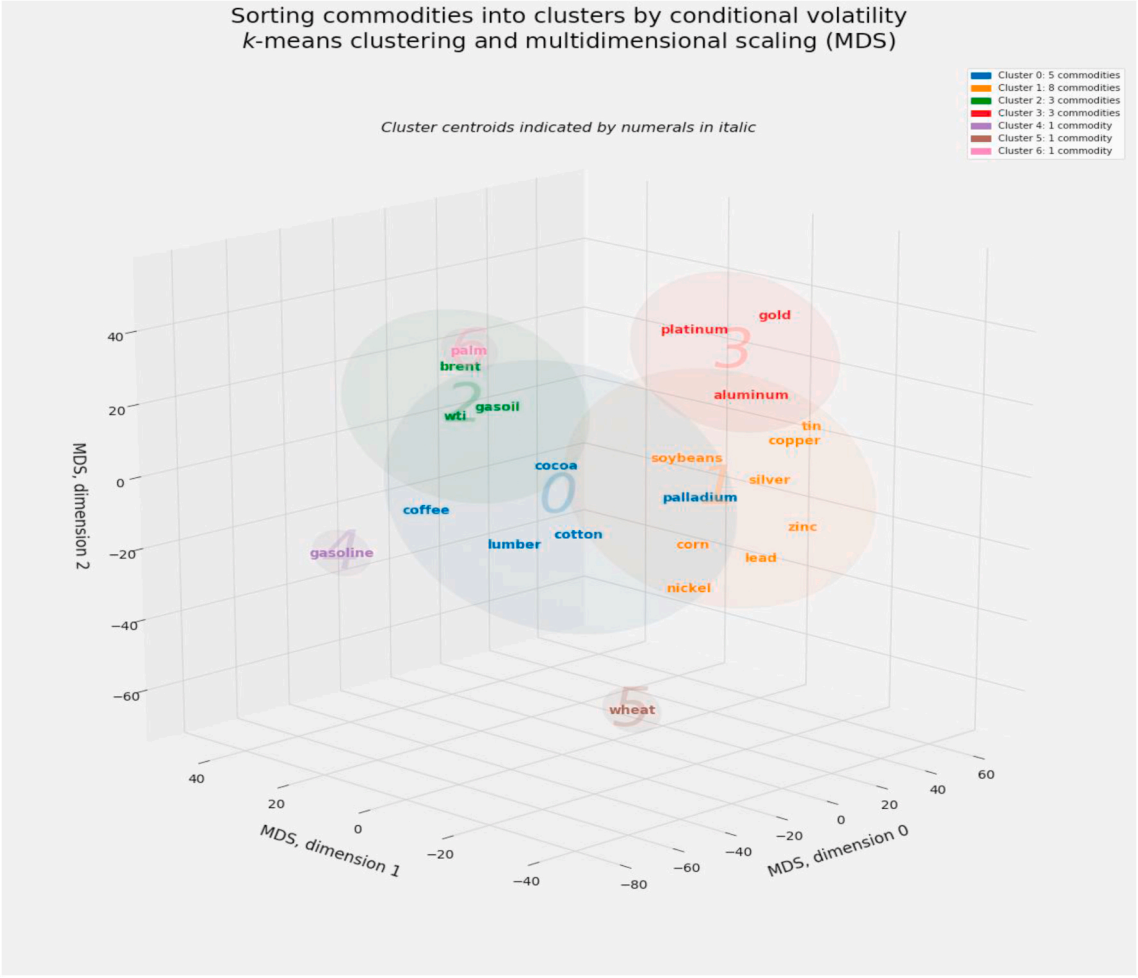

Fig. 8. *K*-means clustering and MDS- conditional volatility.

split palladium from palm oil in cluster 4. The latter move may make intuitive sense, even if it generates two singleton clusters. But further atomization of the fuels hardly appeals to human understandings of commodity markets. Figs. 10 and 11 illustrate Ward's method and the complete-link method of hierarchical clustering from conditional volatility data.

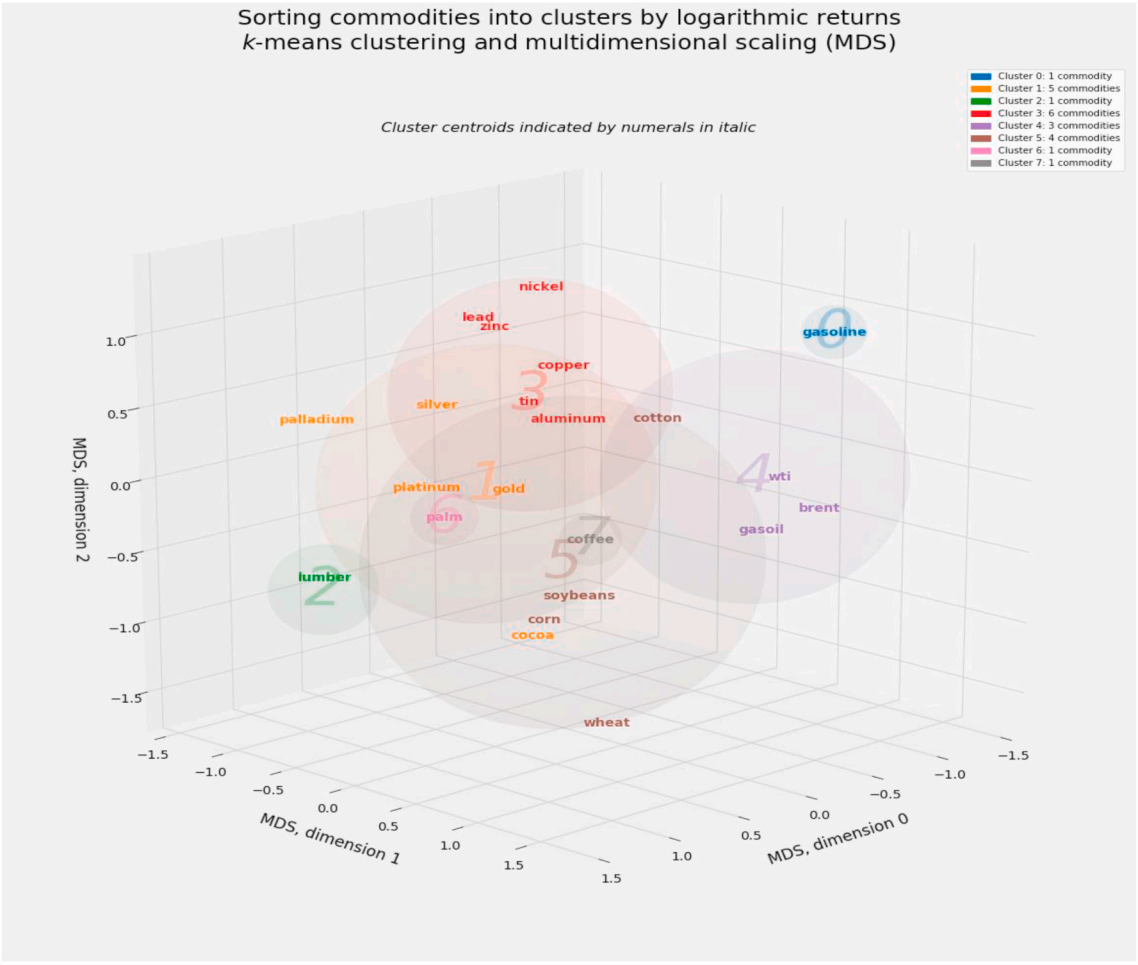

Fig. 9. *K*-means clustering and MDS- log returns.

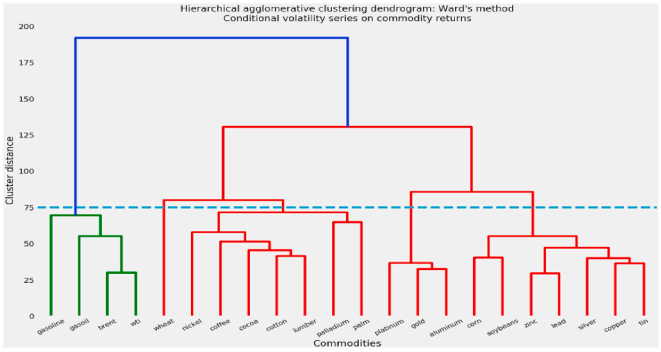

Fig. 10. Hierarchical clustering dendograms- Conditional volatility- Ward's method.

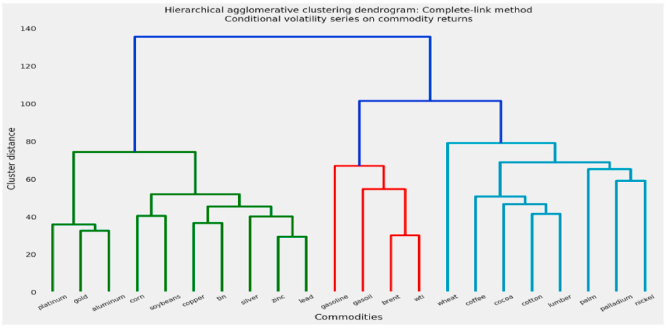

Fig. 11. Hierarchical clustering dendograms- Conditional volatility- Complete-link method.

4.3.2. Hierarchical agglomerative clustering of logarithmic returns

Hierarchical clustering generates equally if not even more compelling taxonomies of commodity trading data based on logarithmic returns. We confine our analysis to results from Ward's method and the complete-link methods. Stripped of differences in order, hierarchical agglomerative clustering of logarithmic returns using Ward's method and the complete-link method assigns the 22 commodities in this study to five broad categories. 1: Fuels; 2: Base metals; 3: Precious metals, with a distinct gap between these subgroups: The lighter period 5 metals of palladium and silver and, the heavier period 6 metals of platinum and gold; 4: The temperate-climate crops of wheat, corn, and soybeans; 5: The tropical "softs" of coffee, palm oil, and cocoa, plus lumber and cotton. These results can be discerned in Figs. 12 and 13.

Relative to hierarchical clustering by conditional volatility, hierarchical clustering by logarithmic returns appears to map more closely to intuitive human understandings of these commodity markets. HAC therefore duplicates the apparent effectiveness of *k*-means clustering on each of these measures of financial space.

One of the more remarkable successes of hierarchical clustering is the persistent separation of the period 5 metals of palladium and silver from the heavier period 6 metals of platinum and gold (Table 2). Given the strength of chemical similarities among elements in a single group, one could imagine an application of clustering that would pair

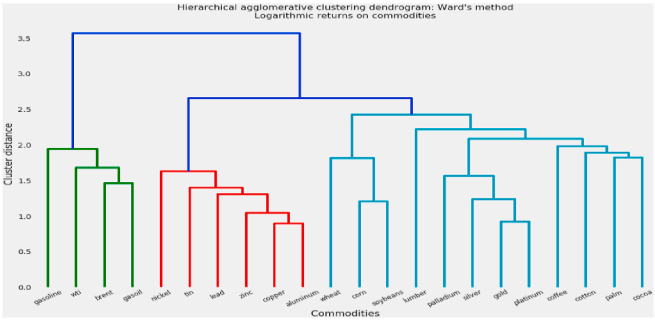

Fig. 12. Hierarchical clustering dendograms-log returns- Ward's method.

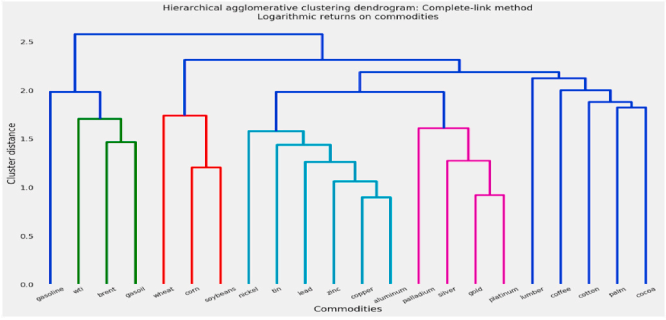

Fig. 13. Hierarchical clustering dendograms-log returns- Complete-link method.

Table 2
Matrix for precious metals in periods and groups.

	Group 10	Group 11
Period 5	46 — Palladium (Pd)	47 — Silver (Ag)
Period 6	78 — Platinum (Pt)	79 — Gold (Au)

palladium with its fellow group 10 metal platinum and silver with its fellow group 11 metal gold. But even the less intuitively cogent clusters generated from the conditional volatility matrix *never* clustered silver with platinum and palladium with gold. That combination of precious metals, unlike the others, would fail to cluster either by group or by period. The clustering of precious metals therefore obeys the periodic table.

Conditional volatility forecasts will nevertheless prove essential in evaluation of the time domain. Although log returns data appear to outperform conditional volatility in clustering commodities in financial space, the advantage of the volatility data will become vividly apparent in the clustering of the matrix transpose. The transpose of the conditional volatility matrix, properly clustered, can reveal critical periods in financial history.

4.4. *Temporal clustering of trading days*

This section applies clustering and manifold learning to the matrix transpose of conditional volatility forecasts. Unsupervised machine learning provides a more objective mathematical basis identifying crisis periods than the qualitative definition in Fernández-Avilés et al. (2020). Clustering has the potential to identify critical periods in many applications of financial economics. For now, this article seeks simply to identify unusual periods in commodities trading.

This temporal analysis requires the transposition of the same conditional volatility matrix previous studied. We now transpose the two-dimensional, $m \times n$ matrix consisting of m rows representing trading dates and n columns representing commodities into a new matrix containing n rows representing commodities and m columns representing dates. Unsupervised learning based on this transposed $n \times m$ matrix should reveal mathematically distinctive *temporal* clusters. Since commodities define the rows and trading dates define the columns, each column expresses conditional volatility for all commodities as of a specific date. Trading dates exhibiting similar traits across all 22 commodities would fall within the same cluster.

Since k-means clustering of trading dates in the matrix transpose of conditional volatility is mathematically equivalent to earlier clustering exercises, we now dispense with recitations of the elbow test and the process by which we located nine k-means centroids. These nine centroids summarize information among more than 5000 trading dates.

Closer examination of the nine centroids in Fig. 14 makes it evident why MDS projects those clusters so far from the center. Alone among the nine clusters, clusters 0 and 8 vary by nearly 10 or more than 10 units. Among the others, only the spread between high and low volatility values in cluster 5 exceeds 3 units. All others fall below 2.

By the same token, all dimensions of MDS correspond to the extremity of the values of observations (the clutch of 22 conditional volatility values for each trading date). Hence, clusters 0 and 8 are not merely further removed from the nine clusters' center of gravity in the MDS projection. The minor and major radii of the spheroids representing these clusters are also longer than those of the other clusters. Yet these are also the smallest clusters. At 129 and 53 days respectively, clusters 0 and 8 fall well beneath the 576-day average for these nine clusters.

The MDS visualization does not altogether lack appeal. It shows how extreme clusters 0 and 8 are, relative to all other clusters. If provided no other information, we might choose to investigate these clusters. And clusters 2 and 5, by virtue of their physical location between these extrema and the bulk of the days in the other clusters, might warrant a closer look. Because of the mixed virtues and vices of an MDS-transformed view of these temporal clusters, we sought an additional way to visualize these clusters of trading days. t-distributed stochastic neighbor embedding, or t-SNE successfully showed the distinctiveness of clusters 0 and 8 while preserving a trait that proves vital: these clusters' compactness in time:

In Fig. 14, the t-SNE plot, like MDS, shows clusters 0 and 8 at the edge of the three-dimensional manifold displaying more than 5000 objects whose real dimensionality is 22. Unlike MDS, however, t-SNE shows how compact these clusters are, in the sense that these clusters are by far the smallest and closest to perfectly contiguous in the entire cohort. Clusters 2 and 5 also appear close to the edge of the collection of bubbles. As in the MDS plot, they are closest, respectively, to clusters 0 and 8. That proximity, however, is less visually evident. Projecting these clusters on along the sequential axis of trading dates reveals why the t-SNE projection is arguably more informative:

In Fig. 16, clusters 0 and 8 are now revealed as the most volatile and economically critical periods in the commodity markets of the past two decades (see Fig. 15). Cluster 0 covers the depth of the 2008-09 global financial crisis, a nearly contiguous 129 days covering just over six months. Cluster 8 is an even more concentrated stretch of 53 days during 2020's Covid-19 crisis. Cluster 2's proximity to cluster 0, more evident in the MDS projection than in the t-SNE projection, reveals itself in the sequential view of dates. Cluster 2 straddles either side of the most intense days of the financial crisis, as captured by cluster 0. The relationship between cluster 5 and the depths of the Covid-19 pandemic in cluster 8 is less obvious, since several other periods throughout this century also fall into cluster 5. A principled measure of contiguity within these temporal clusters, reported in Table 3, shows that cluster 8 is perfectly contiguous, while cluster 0 is nearly so.

4.5. *Cosine distance clustermaps*

Two final visualizations in Figs. 17 and 18 summarize the preceding parts of this article rather succinctly. These two clustermaps, implemented through the Seaborn visualization library for Python, combine

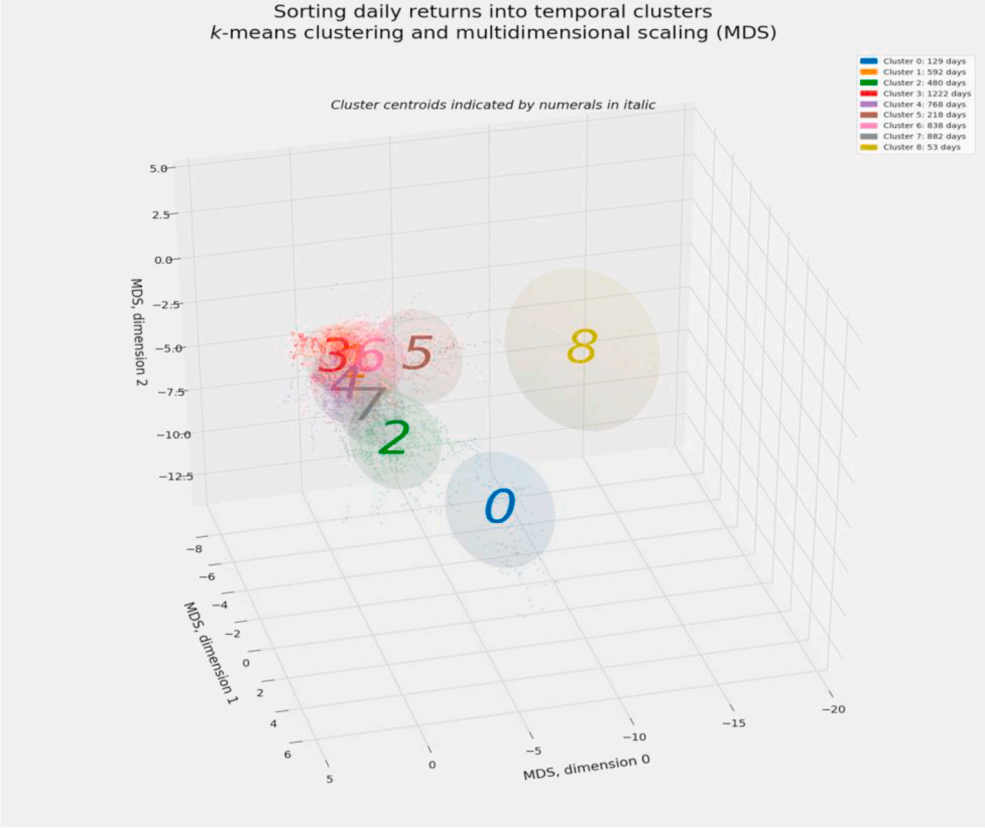

Fig. 14. *K*-means clustering and MDS- returns.

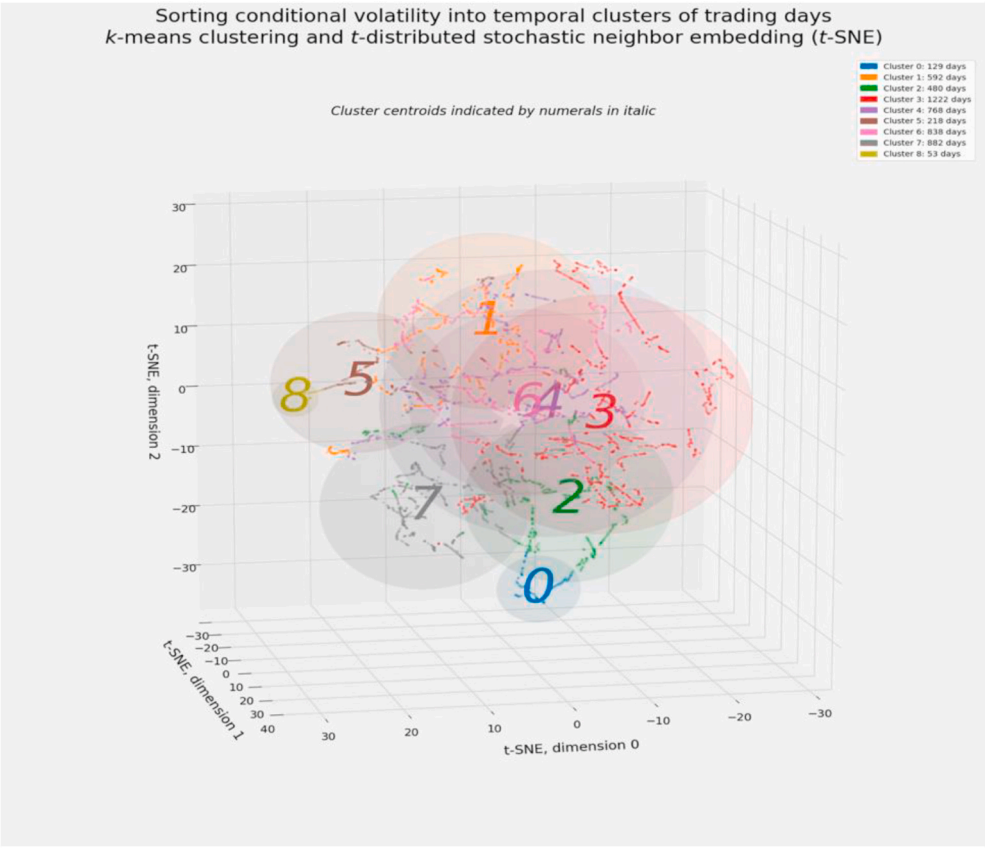

Fig. 15. *K*-means clustering and t-SNE-trading days.

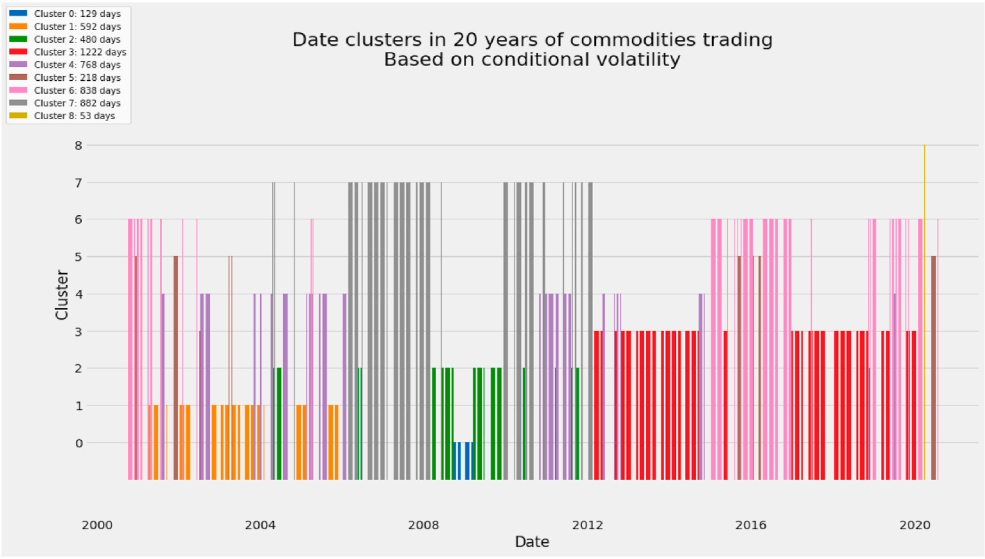

Fig. 16. Data clusters of commodities-conditional volatility.

Table 3
Standard deviation by clusters.

Cluster	No. of trading days	Std. dev. of trading days	Std. dev. of Integers	Std. dev/ Benchmark
0	129	37.2650	37.2380	0.9993
1	592	411.4997	170.8954	0.4153
2	480	598.6796	138.5638	0.2314
3	1222	661.6074	352.7609	0.5332
4	768	1183.4521	221.7023	0.1873
5	218	2067.4613	62.9305	0.0304
6	838	1824.6661	241.9096	0.1326
7	882	562.2167	254.6113	0.4529
8	53	15.2971	15.2971	1.0000

hierarchical clustering of commodities on the vertical axis with heatmap colors showing relationships among dates along the horizontal axis. For the sake of convenience, we relied upon average-linkage implementation of hierarchical clustering among commodities and cosine distance as the distance metric.

Cosine distance is the complement of cosine similarity, which in turn is defined as the inner product of two vectors normalized to the same length of 1 (Singhal, 2001). It is the *uncentered* variant of Pearson's correlation coefficient. (Stated another way, Pearson's correlation coefficient is derived from *centered* cosine similarity.) Cosine distance is used extensively in information retrieval (Zou and Umugwaneza, 2008), natural language processing (Buck and Koehn, 2016), and fuzzy logic (Liu et al., 2019). The cosine distance clustermap based on logarithmic returns combines a perfectly intuitive ontology of commodity classes, albeit with relatively low levels of variation in cosine distance among trading dates:

In Fig. 17, vertically the hierarchical dendrogram at left organizes the commodities in order as agricultural commodities, fuels, base metals, and precious metals. Even the agricultural commodities proceed rather smoothly, as tropical and semitropical crops precede the staple crops of temperate climates (wheat, corn, soybeans). The dates of the 2008–09 financial crisis and the Covid-19 pandemic are visible as vertical streaks of red, even amid a checkered pattern of whites and blues indicating lower cosine distances.

The contrast with the cosine distance heatmap based on conditional volatility could not be starker. The two heatmaps demonstrate the differences between return- and volatility-based evaluation of commodity markets:

The agglomeration of fuels at the bottom of the right axis makes much more intuitive sense within any ontology of commodities trading based on the judgment of human experts. Although the hierarchical clustering of other commodities is seemingly less coherent, the resulting taxonomy is far from useless. Vertical streaks ranging from white to pink and red indicate periods of distressed commodities trading. The distinctive cohort of fuels at the bottom of this clustermap displays the most deviant levels of conditional volatility, as measured through cosine distance. The two most salient critical periods of the past two decades — the financial crisis and the Covid-19 pandemic — are visible as red spikes in a sea of blue.

Moreover, the conditional volatility clustermap reveals a critical difference between those two crises. The financial crisis of 2008–09 transmitted volatility across all commodity markets of interest. Its impact can be seen as a red, pink, and white column covering the entire height of the clustermap. By contrast, the Covid-19 pandemic brought intense volatility to fuels in the lower right, but was less dramatic in its disruption of other commodity sectors.

The horizonal indications in the row for gold provides indirect support as well. Cosine distances between this bellwether "safe harbor" asset and commodity markets overall have remained remarkably low throughout two decades. A lighter shade is visible in the row representing gold in 2008–09, but nothing in this series approaches the extreme levels of volatility observed in fuel markets. If anything, platinum reflects a longer period (in white) of elevated volatility during 2008–09 and a shorter but discernibly more intense level of volatility (in pink) during the pandemic.

With extreme efficiency if not total clarity, these clustermaps extend other tests conducted in this article. The clustermaps illustrate the effect of the average-linkage method of hierarchical agglomerative clustering, distinct from the complete-linkage and Ward's method dendrograms we generated earlier. These clustermaps also revealed the effects of cosine distance, a useful counterweight to our more conventional reliance on Euclidean distance.

Finally, these clustermaps reveal the different benefits of using logarithmic return and conditional volatility as the underlying basis for analyzing financial data. Log returns perform better as ontological guides to commodity markets and their comovement through financial space. Conditional volatility, by contrast, outperforms log returns in identifying crisis periods across market history. The volatility-based clustermap adds even more value to the identification of crisis periods through *k*-means clustering and dimensionality reduction through MDS and *t*-SNE. We can now see with greater clarity the intensity (especially among energy commodities) of the Covid-19 panic and its contrast with

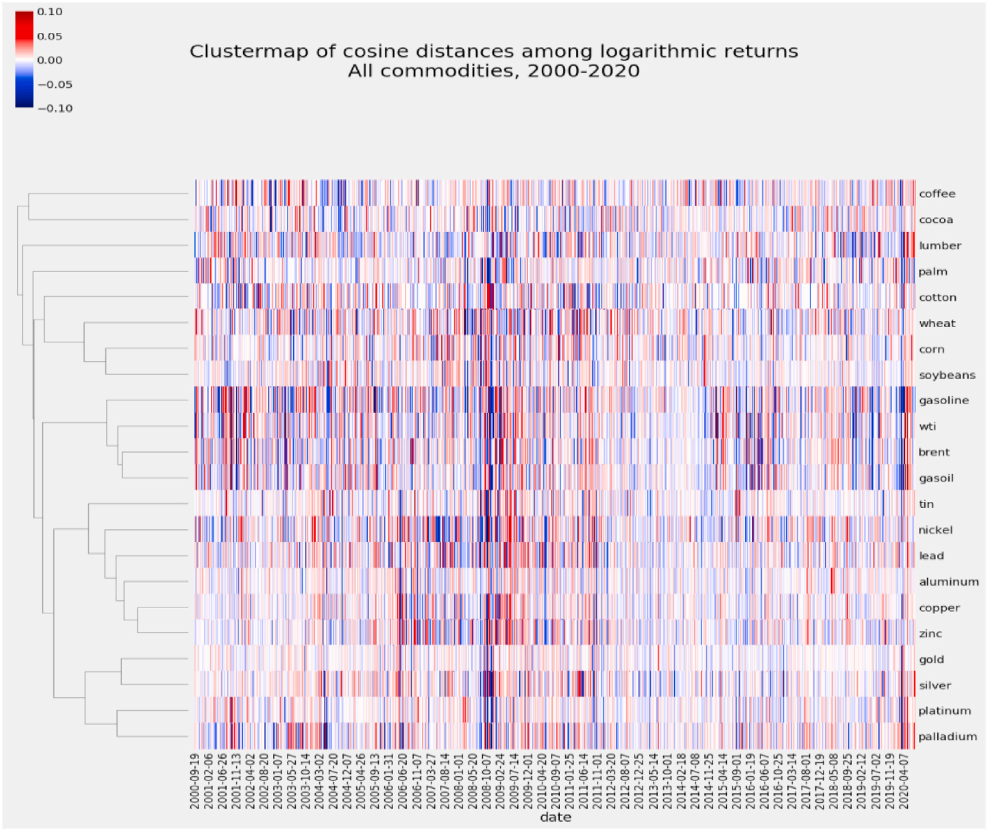

Fig. 17. Cluster maps-log-returns.

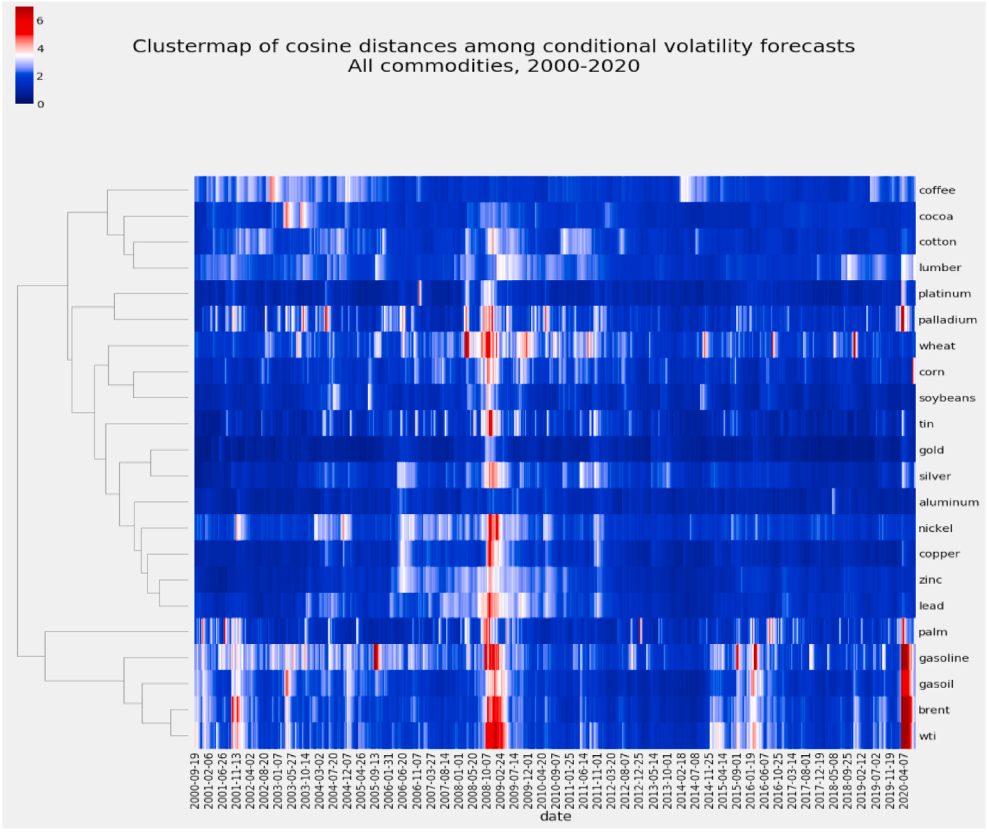

Fig. 18. Cluster maps-conditional volatility.

the widespread, spatially comprehensive nature of the financial crisis of 2008–09.

5. Conclusions and recommendations

We regard these results as the first steps toward more comprehensive analysis of commodity markets. Nothing forces unsupervised learning to work alone or to represent the final step in financial analysis and risk management. Once an ontology of commodity markets in financial space is established through partitioning or hierarchical clustering, further analysis can isolate downside risk or identify arbitrage opportunities.

Risk managers can more readily contain the twin threats of comovement and volatility spillover if they can partition commodity markets or understand them in hierarchical fashion. Individual commodities that do not move together should not be regarded as members of a common asset class. Conversely, greater comovement may signal volatility spillover from more sensitive sectors (such as fuels) to markets across the spectrum of commodities or even to other asset classes such as equities, conventional currencies, and cryptocurrencies.

Far from signaling failure in machine learning, differences between clustering results based on logarithmic returns and conditional volatility present novel insights and learning opportunities for risk management in commodity markets. To be sure, spatial ontologies based on log returns conform more closely to expert human judgment. But different configurations of these markets suggested by volatility-based clustering suggest that expert judgment may have overlooked crucial relationships lurking within the data. Despite the notorious obstacles to the interpretation of many machine learning models, the insights that do arise from machine learning sometimes challenge — and invariably sharpen — intuitions based on expert human judgment as well as inferences based on the closed-form mathematical models of more conventional predictive methods (Chen, 2021).

The value of volatility-based ontologies of commodities and their comovement in financial space increases dramatically in light of the unique success in volatility-based *temporal* clustering of trading days. Temporal clustering of the matrix transpose of logarithmic returns, which admittedly generated more intuitively appealing definitions of asset classes and subclasses in commodity markets, failed to detect the financial crisis of 2008–09 or the Covid-19 pandemic.

As a strictly methodological matter, unsupervised learning also presents an opportunity to bridge the dominant culture of statistics with the relatively new algorithmic culture to which machine learning belongs (Breiman, 2001). Manifold learning, in particular, has already contributed to many conventional economic studies. By simplifying data and making it more analytically tractable, principal components analysis (Abdi and Williams, 2010; Wold et al., 1987) and other methods of manifold learning transform the "*curse of dimensionality*" (Taylor, 1993; Trunk, 1979) into an affirmative blessing (Gorban et al., 2018, 2020).

For their part, *k*-means centroids take exactly the shape of objects intended to cluster around them. Distances to these objects in high-dimensional space can assume analytical meaning outside the specific, originally intended context of clustering through spatial partitioning. Further research may range from smaller, more subtle expedients such as the adoption of alternative distance metrics to the more comprehensive substitution of methods for generating models of space and time in commodities trading. The use of cosine distance in part 8's clustermaps represents a tentative first experiment with alternative distance metrics.

Each of the three primary clustering methods in this paper can guide decisions by commodity investors. First, the primary alignment between returns-based clusters reinforces conventional classifications of commodities markets. Beyond the obvious lines of division between fuels, metals, and crops, returns-based clustering confirms pricing relationships dictated by climatic conditions, industrial applications, etc. Second, the secondary classification supplied by volatility-based clusters provides deeper insight into comovement among commodity prices. Aluminum, despite its contemporary classification among base metals, can align alongside gold and platinum in an echo of the historical expense incurred in its refinement. Finally, the clustering of trading days within the matrix transpose of conditional volatility data identifies distinctive regimes in commodities trading during possible periods of crises. Although there is no assurance that future crises will resemble past crises such as the global financial crisis of 2008–09 or the Covid-19 pandemic, investors can apply the concepts of clustering and distance from other observations as they devise hedging strategies and build resilience into their portfolios.

More ambitious extensions of this work may involve other methods of unsupervised learning, such as kernel density estimation and Gaussian mixture modeling. Clustering and manifold learning can preprocess data in anticipation of time-series forecasting, which in turn may involve some combination of conventional methods and machine learning. Ironically enough, the contrary case may also be true: Conventional methods designed to isolate the time domain, frequency domain, and wavelet structure of time series might fruitfully preprocess data in anticipation of clustering. A distinct literature on the use of clustering methods in time-series forecasting counsels against the direct application of clustering to noisy data, lest this class of methods obliterate the autocorrelation structure underlying the data (Caiado, Maharaj & D'Urso, 2015; D'Urso et al., 2016, p. 2).

Once again, the no-free-lunch theorem of machine learning offers useful guidance (Wolpert, 1996). Since it is unclear *ex ante* which attributes of financial data constitute noise as opposed to signal, complete analysis may require unsupervised learning with and without conventional time-series preprocessing calculated to ensure stationarity and to eliminate seasonal and cyclical effects.

Another fruitful and focused direction for further research lies in reconciling the relatively modest methodological differences between this study and Fernández-Avilés et al. (2020). All spatial analysis in that article relied upon dynamic expected shortfall as the crucial variable of interest. Though we wholly understand these authors' emphasis on extreme downside risk, we have chosen to rely on three rather merely two dimensions generated by manifold learning, whether MDS or *t*-SNE. We are unaware of any principled reason to prefer the pairing of expected shortfall with two arbitrary MDS dimensions as opposed to the generation of three arbitrary MDS or *t*-SNE dimensions. The underlying data, after all, stem from the same GJR-GARCH(1, 1, 1) forecasting process. Whereas we were content to subject the resulting conditional volatility forecasts directly to clustering, Fernández-Avilés, Montero, and Sanchis-Marco took the additional step of computing expected shortfall according to Student's *t*-distribution.

We believe that the real difference between these studies does not hinge on the choice between three wholly arbitrary dimensions of data generated by machine-learning processes and a hybrid models blending merely two such dimensions with a summary statistic generated through conventional economics. Instead, the evident motivation of Fernández-Avilés, Montero, and Sanchis-Marco lies in an effort to capture higher-moment effects through expected shortfall. If unsupervised learning should indeed strive to analyze higher-moment phenomena in financial data, it should apply clustering and manifold learning directly to conditional skewness and kurtosis. The analogy is straightforward: If conditional volatility is the second-moment generalization of logarithmic return data seeking to capture the first-moment expectation from the cross-section of commodity returns, then we should apply *k*-means and hierarchical clustering directly to conditional skewness and kurtosis. (Brooks et al., 2005; Jondeau and Rocklinger, 2003).

We also propose applying Fernández-Avilés, Montero, and Sanchis-Marco's emphasis on the evaluation of expected shortfall, correlation, and other measures of comovement during known or suspected critical periods. Having isolated temporal trading regimes within the data, we could contemplate further research that bridges the methods demonstrated here with our counterparts' crisis-based analysis. Recalibrating

clustering methods according critical subsets of the data, especially in critical temporal clusters that we have identified, may expose the existence of more durable regimes governing commodities during calmer periods, as distinct from periods of extreme perturbation that disrupt "punctuated equilibrium" throughout the long-run evolution of commodity markets (Gould, 1989; Gould & Eldridge, 1993, 1996, 1996; Prindle, 2012).

CRediT authorship contribution statement

James Ming Chen: Writing – original draft, Software, Visualization, Investigation, Writing – review & editing. **Mobeen Ur Rehman:** Conceptualization, Data curation, Writing – review & editing. **Xuan Vinh Vo:** Conceptualization, Supervision, Investigation, Writing – review & editing.

Acknowledgement

This research is partly funded by the University of Economics Ho Chi Minh City, Vietnam.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. WIREs Comput. Stat. 2, 433–459. https://doi.org/10.1002/wics.101.
Agrawal, A., Gupta, U., 2014. Extraction based approach for text summarization using k-means clustering. Int. J. Sci. Res. Publ. 4 (11), 1–4.
Al-Yahyaee, K.H., Mensi, W., Rehman, M.U., Vo, X.V., Kang, S.H., 2020. Do Islamic stocks outperform conventional stock sectors during normal and crisis periods? Extreme co-movements and portfolio management analysis. Pac. Basin Finance J. 62, 101385.
Alexander, C., Lazar, E., Stanescu, S., 2021. Analytic moments for GJR-GARCH (1, 1) processes. Int. J. Forecast. 37, 105–124. https://doi.org/10.1016/j.ijforecast.2020.03.005.
Andrianto, Y., Diputra, Y., 2017. The effect of cryptocurrency on investment portfolio effectiveness. J. Finance Account. 5 (6), 229–238.
Balakrishnan, N., Nevrozov, V.B., 2003. A Primer on Statistical Distributions, first ed. John Wiley & Sons, Hoboken, N.J.
Batten, J.A., Ciner, C., Lucey, B.M., 2010. The macroeconomic determinants of volatility in precious metals markets. Resour. Pol. 35 (2), 65–71.
Baur, D.G., McDermott, T.K., 2010. Is gold a safe haven? International evidence. J. Bank. Finance 34 (8), 1886–1898.
Bholowalia, P., Kumar, A., 2014. EBK-means: a clustering technique based on elbow method and k-means in WSN. Int. J. Comput. Appl. 105, 17–24.
Blashfield, R.K., 1976. Mixture model tests of cluster analysis: accuracy of four agglomerative hierarchical methods. Psychol. Bull. 83, 377–388.
Bollerslev, T., Wooldridge, J.M., 1992. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. Econom. Rev. 11 (2), 143–172. https://doi.org/10.1080/07474939208800229.
Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., Song, A., 2015. Efficient agglomerative hierarchical clustering. Expert Syst. Appl. 42, 2785–2797.
Breiman, L., 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. 16, 199–231.
Broadstock, D.C., Filis, G., 2014. Oil price shocks and stock market returns: new evidence from the United States and China. J. Int. Financ. Mark. Inst. Money 33, 417–433.
Brooks, C., Burke, S.P., Heravi, S., Persand, G., 2005. Autoregressive conditional kurtosis. J. Financ. Econom. 3, 399–421.
Buck, C., Koehn, P., 2016. Quick and reliable document alignment via tf/idf-weighted cosine distance. In: Proceedings of the First Conference on Machine Translation, vol. 2. Shared Task Papers, pp. 672–678.
Cabrera, B.L., Schulz, F., 2016. Volatility linkages between energy and agricultural commodity prices. Energy Econ. 54, 190–203.
Cagli, E.C., Taskin, D., Mandaci, P.E., 2019. The short-and long-run efficiency of energy, precious metals, and base metals markets: evidence from the exponential smooth transition autoregressive models. Energy Econ. 84, 104540.
Cai, F., Le-Khac, N.-A., Kechadi, M.-T., 2016. Clustering Approaches for Financial Data Analysis: A Survey arXiv: 1609.08520.
Caiado, J., Maharag, E.A., D'Urso, P., 2015. Time series clustering. In: Hennig, C., Melia, M., Murtagh, F., Rocci, R. (Eds.), Handbook of Cluster Analysis. Chapman and Hall/CRC, Boca Raton, Fla, pp. 241–264.
Capó, M., Pérez, A., Lozano, J.A., 2017. An efficient approximation to the k-means clustering for massive data. Knowl. Base Syst. 117, 56–69. https://doi.org/10.1016/j.knosys.2016.06.031.
Cashin, P., McDermott, C.J., 2002. The long-run behavior of commodity prices: small trends and big variability. IMF Staff Pap. 49, 175–199.
Cashin, P., McDermott, C.J., Pattillo, C., 2004. Terms of trade shocks in Africa: are they short-lived or long-lived? J. Dev. Econ. 73, 727–744.
Chan, D.M., Rao, R., Huang, F., Canny, J.F., 2018. t-SNE-CUDA: GPU-accelerated t-SNE and its applications to modern data. In: 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), pp. 330–338.
Charles, A., Darné, O., Kim, J.H., 2015. Will precious metals shine? A market efficiency perspective. Int. Rev. Financ. Anal. 41, 284–291.
Chen, J.M., 2021. An introduction to machine learning for panel data. Int. Adv. Econ. Res. 27 forthcoming; preprint at. http://ssrn.com/abstract=3717879.
Chen, M.H., 2010. Understanding world metals prices — returns, volatility and diversification. Resour. Pol. 35 (3), 127–140.
Conlon, T., McGee, R., 2020. Safe haven or risky hazard? Bitcoin during the COVID-19 bear market. Finance Res. Lett. 35, 101607. https://doi.org/10.1016/j.frl.2020.101607.
Corbet, S., Yang, (G.) H., Yang, H., Larkin, C., Oxley, L., 2020. Any port in a storm: cryptocurrency safe-havens during the COVID-19 pandemic. Econ. Lett. 194, 109377. https://doi.org/10.1016/j.econlet.2020.109377.
Cox, M.A.A., Cox, T.F., 2008. Multidimensional scaling. In: Chen, C.-H., Härdle, W., Unwin, A. (Eds.), Handbook of Data Visualization. Springer, Berlin, pp. 315–347. https://doi.org/10.1007/978-3-540-33037-0_14.
D'Urso, P., De Giovanni, L., Massari, R., 2016. GARCH-based robust clustering of time series. Fuzzy Set Syst. 303, 1–28. https://doi.org/10.1016/j.fss.2016.01.010.
Dai, X.-Y., Chen, Q.-C., Wang, X.-L., Xu, J., 2010. Online topic detection and tracking of financial news based on hierarchical clustering. In: *2010* International Conference on Machine Learning and Cybernetics. https://doi.org/10.1109/ICMLC.2010.5580677.
Davidson, I., Ravi, S.S., 2005. Agglomerative hierarchical clustering with constraints: theoretical and empirical results. In: European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, pp. 59–70.
Day, W.H., Edelsbrunner, H., 1984. Efficient algorithms for agglomerative hierarchical clustering methods. J. Classif. 1, 7–24.
Demiralay, S., Ulusoy, V., 2014. Non-linear volatility dynamics and risk management of precious metals. N. Am. J. Econ. Finance 30, 183–202.
Deng, Q., Mei, G., 2009. Combining self-organizing map and K-means clustering for detecting fraudulent financial statements. In: 2009 IEEE International Conference on Granular Computing. https://doi.org/10.1109/GRC.2009.5255148.
Du, X., Yu, C.L., Hayes, D.J., 2011. Speculation and volatility spillover in the crude oil and agricultural commodity markets: a Bayesian analysis. Energy Econ. 33, 497–503. https://doi.org/10.1016/j.eneco.2010.12.015.
Duch, W., Matykiewicz, P., Pestian, J., 2008. Neurolinguistic approach to natural language processing with applications to medical text analysis. Neural Network. 21, 1500–1510.
Falkowski, M., 2011. Financialization of commodities. Contemp. Econ. 5 (4), 4–17.
Feller, W., 1971. An Introduction to Probability Theory and its Applications, second ed., vol. II. John Wiley & Sons, New York.
Fernández-Avilés, G., Montero, J.-M., Sanchis-Marco, L., 2020. Extreme downside risk comovement during distress periods: a multidimensional scaling approach. Eur. J. Finance 26, 1207–1237. https://doi.org/10.1080/1351847X.2020.1724171.
Forster, R., 2006. Document clustering in large German corpora using natural language processing (Doctoral dissertation, University of Zürich). Available at: https://www.zora.uzh.ch/id/eprint/163398/1/20060041.pdf.
Fung, B.C., Wang, K., Ester, M., 2009. Hierarchical document clustering. In: Wang, J. (Ed.), Encyclopedia of Data Warehousing and Mining, second ed. IGI Global, Hershey, Pa, pp. 970–975.
Gil-Garcia, R.J., Badia-Contelles, J.M., Pons-Porrata, A., 2006. A general framework for agglomerative hierarchical clustering algorithms. In: 18th International Conference on Pattern Recognition (ICPR '06), pp. 569–572.
Gorban, A.N., Tyukin, I.Y., 2018. Blessing of dimensionality: mathematical foundations of the statistical physics of data. Phil. Trans. Roy. Soc. A 376 (2118), 20170237. https://doi.org/10.1098/rsta.2017.0237.
Gorban, A.N., Makarov, V.A., Tyukin, I.Y., 2020. High-dimensional brain in a high-dimensional world: blessing of dimensionality. Entropy 22 (1), 82. https://doi.org/10.3390/e22010082.
Gould, S.J., 1989. Punctuated equilibrium in fact and theory. J. Soc. Biol. Struct. 12 (2–3), 117–136.
Gould, S.J., Eldredge, N., 1993. Punctuated equilibrium comes of age. Nature 366 (6452), 223–227.
Hammoudeh, S., Santos, P.A., Al-Hassan, A., 2013. Downside risk management and VaR-based optimal portfolios for precious metals, oil and stocks. N. Am. J. Econ. Finance 25, 318–334.
Haque, M., Kouki, I., 2009. Effect of 9/11 on the conditional time-varying equity risk premium: evidence from developed markets. J. Risk Finance 10, 261–276. https://doi.org/10.1108/15265940910959384.
Hepsen, A., Vatansever, M., 2012. Using hierarchical clustering algorithms for Turkish residential market. Int. J. Econ. Finance 4, 138–150.
Hout, M.C., Papesh, M.H., Goldinger, S.D., 2013. Multidimensional scaling. WIREs Cogn. Sci. 4, 93–103. https://doi.org/10.1002/wcs.1203.
Huang, S., Peng, X., Niu, Z., Wang, K., 2011. News topic detection based on hierarchical clustering and named entity. In: 7th International Conference on Natural Language Processing and Knowledge Engineering, pp. 280–284.
Ishizaka, A., Lokman, B., Tasiou, M., 2020. A stochastic multi-criteria divisive hierarchical clustering algorithm. Omega. https://doi.org/10.1016/j.omega.2020.102370.
Jain, A.K., Murty, M., Flynn, R.J., 1999. Data clustering: a review. ACM Comput. Surv. 31, 265–323.
Jain, H.J., Bewoor, M.S., Patil, S.H., 2012. Context sensitive text summarization using K means clustering algorithm. Int. J. Soft Comput. Eng. 2, 301–304.
Jondeau, E., Rockinger, M., 2003. Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements. J. Econ. Dynam. Contr. 27, 1699–1737.

Kaushik, M., Mathur, B., 2014. Comparative study of k-means and hierarchical clustering techniques. Int. J. Softw. Hardw. Res. Eng. 2 (6), 93–98.
Kodinariya, T.M., Makwana, P.R., 2013. Review on determining number of clusters in k-means clustering. Int. J. 1, 90–95.
Kou, G., Peng, Y., Wang, G., 2014. Evaluation of clustering algorithms for financial risk analysis. Inf. Sci. 275, 1–12. https://doi.org/10.1016/j.ins.2014.02.137.
Kuiper, F.K., Fisher, L.A., 1975. A Monte Carlo comparison of six clustering procedures. Biometrics 31, 777–783. https://doi.org/10.2307/2529565.
Kumar, S., Deo, N., 2012. Correlation and network analysis of global financial indices. Phys. Rev. 86, 026101 https://doi.org/10.1103/PhysRevE.86.026101.
Lengyel, A., Botta-Dukát, Z., 2019. Silhouette width using generalized mean — a flexible method for assessing clustering efficiency. Ecol. Evol. 3, 5774. https://doi.org/10.1002/ece3.5774.
Li, K., Yang, R.J., Robinson, D., Ma, J., Ma, Z., 2019. An agglomerative hierarchical clustering-based strategy using shared nearest neighbours and multiple dissimilarity measures to identify typical daily electricity usage profiles of university library buildings. Energy 174, 735–748. https://doi.org/10.1016/j.energy.2019.03.003.
Lin, D., Wu, X., 2009. Phrase clustering for discriminative learning. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1030–1038.
Liu, C., Naeem, M.A., Rehman, M.U., Farid, S., Shahzad, S.J.H., 2020. Oil as hedge, safe-haven, and diversifier for conventional currencies. Energies 13 (17), 4354. https://doi.org/10.3390/en13174354.
Liu, D., Chen, X., Peng, D., 2019. Some cosine similarity measures and distance measures between q-rung orthopair fuzzy sets. Int. J. Intell. Syst. 34, 1572–1587. https://doi.org/10.1002/int.22108.
MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297.
Mahdizadeh, M., Zamanzade, E., 2019. Goodness-of-fit testing for the Cauchy distribution with application to financial modeling. J. King Saud Univ. Sci. 31, 1167–1174. https://doi.org/10.1016/j.jksus.2019.01.015.
Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge, U.K.
Marti, G., Nielsen, F., Bińkowski, M., Donnat, P., 2020. A Review of Two Decades of Correlations, Hierarchies, Networks and Clustering in Financial Markets arXiv: 1703.00485.
Martino, S., Parson, L.M., 2013. Spillovers between cobalt, copper and nickel prices: implications for deep seabed mining. Min. Econ. 25, 107–127. https://doi.org/10.1007/s13563-012-0027-8.
Mensi, W., Hammoudeh, S., Reboredo, J.C., Nguyen, D.K., 2015. Are Sharia stocks, gold and US Treasury hedges and/or safe havens for the oil-based GCC markets? Emerg. Mark. Rev. 24, 101–121.
Micciche, S., Lillo, F., Mantegna, R.N., 2005. Correlation based hierarchical clustering in financial times series. In: Beck, C., Benedek, G., Rapisarda, A., Tsallis, C. (Eds.), Complexity, Metastability and Nonextensivity: Proceedings of the 31st Workshop of the International School of Solid State Physics, pp. 327–335. https://doi.org/10.1142/9789812701558_0037.
Milligan, G.W., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45, 325–342. https://doi.org/10.1007/BF02293907.
Münnix, M.C., Shimada, T., Schäfer, R., Leyvraz, F., Seligman, T.H., Guhr, T., Stanley, H. E., 2012. Identifying states of a financial market. Sci. Rep. 2, 644. https://doi.org/10.1038/srep00644.
Murtagh, F., 1983. A survey of recent advances in hierarchical clustering algorithms. Comput. J. 26, 354–359. https://doi.org/10.1093/comjnl/26.4.354.
Musmeci, N., Aste, T., Di Matteo, T., 2015. Relation between financial market structure and the real economy: comparison between clustering methods. PloS One 10 (3), e0116201. https://doi.org/10.1371/journal.pone.0116201.
Naeem, M.A., Balli, F., Shahzad, S.J.H., de Bruin, A., 2020. Energy commodity uncertainties and the systematic risk of US industries. Energy Econ. 85, 104589.
Nakhleh, L., 2013. Evolutionary trees. In: Maloy, S., Hughes, K. (Eds.), Brenner's Encyclopedia of Genetics, second ed. Academic Press, Cambridge, Mass, pp. 549–550. https://doi.org/10.1016/B978-0-12-374984-0.00504-0.
Nanda, S.R., Mahanty, B., Tiwari, M.K., 2010. Clustering Indian stock market data for portfolio management. Expert Syst. Appl. 37, 8793–8798. https://doi.org/10.1016/j.eswa.2010.06.026.
Nazlioglu, S., Erdem, C., Soytas, U., 2013. Volatility spillover between oil and agricultural commodity markets. Energy Econ. 36, 658–665. https://doi.org/10.1016/j.eneco.2012.11.009.
Nugroho, D.B., Kurniawati, D., Panjaitan, L.P., Kholil, Z., Susanto, B., Sasongko, L.R., 2019. Empirical performance of GARCH, GARCH-M, GJR-GARCH and log-GARCH models for returns volatility. J. Phys. Conf. 1037, 012003 https://doi.org/10.1088/1742-6596/1307/1/012003.
Ohakwe, J., Osu, B., 2011. The existence of the moments of the Cauchy distribution under a simple transformation of dividing with a constant. Theor. Math. Appl. 1, 27–35.
Pattarin, F., Paterlini, S., Minerva, T., 2004. Clustering financial time series: an application to mutual funds style analysis. Computational Statistics & Data Analysis 47 (2), 353–372.
Perez, H., Tah, J.H.M., 2020. Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-SNE. Mathematics 8 (5), 662. https://doi.org/10.3390/math8050662.
Plourde, A., Watkins, G.C., 1998. Crude oil prices between 1985 and 1994: how volatile in relation to other commodities? Resour. Energy Econ. 20 (3), 245–262.
Prindle, D.F., 2012. Importing concepts from biology into political science: the case of punctuated equilibrium. Pol. Stud. J. 40, 21–44. https://doi.org/10.1111/j.1541-0072.2011.00432.x.
Puerto, J., Rodríguez-Madrena, M., Scozzari, A., 2020. Clustering and portfolio selection problems: a unified framework. Comput. Oper. Res. 117, 104891. https://doi.org/10.1016/j.cor.2020.104891.
Reboredo, J.C., Ugolini, A., 2015. Systemic risk in European sovereign debt markets: a CoVaR-copula approach. J. Int. Money Finance 51, 214–244.
Reboredo, J.C., Rivera-Castro, M.A., Ugolini, A., 2016. Downside and upside risk spillovers between exchange rates and stock prices. J. Bank. Finance 62, 76–96.
Rehman, M.U., 2020. Do bitcoin and precious metals do any good together? An extreme dependence and risk spillover analysis. Resour. Pol. 68, 101737.
Rehman, M.U., Apergis, N., 2019. Determining the predictive power between cryptocurrencies and real time commodity futures: evidence from quantile causality tests. Resour. Pol. 61, 603–616.
Rehman, M.U., Vo, X.V., 2020. Cryptocurrencies and precious metals: a closer look from diversification perspective. Resour. Pol. 66, 101652.
Rehman, M.U., Bouri, E., Eraslan, V., Kumar, S., 2019. Energy and non-energy commodities: an asymmetric approach towards portfolio diversification in the commodity market. Resour. Pol. 63, 101456.
Rehman, M.U., Shahzad, S.J.H., Uddin, G.S., Hedström, A., 2018. Precious metal returns and oil shocks: a time varying connectedness approach. Resour. Pol. 58, 77–89.
Riley, K.F., Hobson, M.P., Bence, S.J., 2006. Mathematical Methods for Physics and Engineering, third ed. Cambridge University Press, Cambridge, U.K.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. 20, 53–65.
Roux, M.A., 2018. Comparative study of divisive and agglomerative hierarchical clustering algorithms. J. Classif. 35, 345–366. https://doi.org/10.1007/s00357-018-9259-9.
Rozga, A., Arnerić, J., 2013. Dependence between volatility persistence, kurtosis and degrees of freedom. Invest. Oper. 30, 32–39.
Sakemoto, R., 2018. Do precious and industrial metals act as hedges and safe havens for currency portfolios? Finance Res. Lett. 24, 256–262.
Selmi, R., Mensi, W., Hammoudeh, S., Bouoiyour, J., 2018. Is Bitcoin a hedge, a safe haven or a diversifier for oil price movements? A comparison with gold. Energy Econ. 74, 787–801. https://doi.org/10.1016/j.eneco.2018.07.007.
Sensoy, A., 2013. Dynamic relationship between precious metals. Resour. Pol. 38 (4), 504–511.
Serra, T., 2011. Volatility spillovers between food and energy markets: a semiparametric approach. Energy Econ. 33, 1155–1164. https://doi.org/10.1016/j.eneco.2011.04.003.
Silvennoinen, A., Thorp, S., 2013. Financialization, crisis and commodity correlation dynamics. J. Int. Financ. Mark. Inst. Money 24, 42–65.
Singhal, A., 2001. Modern information retrieval: a brief overview. Bull. IEEE Comput. Soc. Tech. Committee Data Eng. 24 (4), 35–43.
Song, J.Y., Chang, W., Song, J.W., 2019. Cluster analysis on the structure of the cryptocurrency market via Bitcoin-Ethereum filtering. Physica A 527, 121339. https://doi.org/10.1016/j.physa.2019.121339.
Soni, K.G., Patel, A., 2017. Comparative analysis of k-means and k-medoids algorithm on IRIS data. Int. J. Comput. Intell. Res. 13, 899–906.
Spencer, S., Bredin, D., Conlon, T., 2018. Energy and agricultural commodities revealed through hedging characteristics: evidence from developing and mature markets. J. Commod. Market. 9, 1–20.
Tang, K., Xiong, W., 2010. The Financialisation of Commodities. Vox EU. V November 30.
Tang, K., Xiong, W., 2012. Index investment and the financialization of commodities. Financ. Anal. J. 68 (6), 54–74.
Taylor, C.R., 1993. Dynamic programming and the curses of dimensionality. In: Taylor, C.R. (Ed.), Applications of Dynamic Programming to Agricultural Decision Problems. Westview Press, New York, pp. 1–10.
Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters via the gap statistic. J. Roy. Stat. Soc. B 63, 411–423.
Todorova, N., Worthington, A., Souček, M., 2014. Realized volatility spillovers in the non-ferrous metal futures market. Resour. Pol. 39, 21–31.
Trunk, G.V., 1979. A problem of dimensionality: a simple example. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1, vol. 3, pp. 306–307.
Tsai, C.-F., 2014. Combining cluster analysis with classifier ensembles to predict financial distress. Inf. Fusion 16, 46–58. https://doi.org/10.1016/j.inffus.2011.12.001.
Tumminello, T., Lillo, F., Mantegna, R.N., 2010. Correlation, hierarchies, and networks in financial markets. J. Econ. Behav. Organ. 75, 40–58. https://doi.org/10.1016/j.jebo.2010.01.004.
Uddin, G.S., Shahzad, S.J.H., Boako, G., Hernandez, J.A., Lucey, B.M., 2019. Heterogeneous interconnections between precious metals: evidence from asymmetric and frequency-domain spillover analysis. Resour. Pol. 64, 101509.
van der Maaten, L.J.P., 2009. Learning a parametric embedding by preserving local structure. In: Proceedings of the Twelfth International Conference on Artificial Intelligence & Statistics. (AI-STATS), pp. 384–391.
van der Maaten, L.J.P., 2014. Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. 15, 3221–3245.
van der Maaten, L.J.P., Hinton, G.E., 2008. Visualizing high-dimensional data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.
van der Maaten, L.J.P., Hinton, G.E., 2012. Visualizing non-metric similarities in multiple maps. Mach. Learn. 87, 33–55.
Vijaya, Sharma, S., Batra, N., 2019. Comparative study of single linkage, complete linkage, and Ward method of agglomerative clustering. In: International Conference

on Machine Learning, Big Data, and Cloud and Parallel Computing. https://doi.org/10.1109/COMITCon.2019.8862232.
Wang, Y., et al., 2018. A comparison of word embeddings for the biomedical natural language processing. J. Biomed. Inf. 87, 12–20.
Wazarkar, S.V., Manjrekar, A.A., 2014. Text clustering using HFRECCA and rough k-means clustering algorithm. Discovery 15 (40), 44–47.
Woese, C.R., 2000. Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci. Unit. States Am. 97, 8392–8396. https://doi.org/10.1073/pnas.97.15.8392.
Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometr. Intell. Lab. Syst. 2, 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.
Wolpert, David, 1996. The lack of *a priori* distinctions between learning algorithms. Neural Comput. 8, 1341–1390.
Xu, S., Qiao, X., Zhu, L., Zhang, Y., Xue, C., Li, L., 2016. Reviews on determining the number of clusters. Appl. Math. Inf. Sci. 10, 1493–1512.
Xu, Y., Yang, C., Peng, S., Nojima, Y., 2020. A hybrid two-stage financial stock forecasting algorithm based on clustering and ensemble learning. Appl. Intell. 50, 3852–3867. https://doi.org/10.1007/s10489-020-01766-5.
Zech, J., et al., 2018. Natural language–based machine learning models for the annotation of clinical radiology reports. Radiology 287, 570–580.
Zhang, Y.J., Wei, Y.M., 2010. The crude oil market and the gold market: evidence for cointegration, causality and price discovery. Resour. Pol. 35 (3), 168–177.
Zou, B.J., Umugwaneza, M.P., 2008. Shape-based trademark retrieval using cosine distance method. In: 2008 Eighth International Conference on Intelligent Systems Design and Applications, vol. 2, pp. 498–504.