Financial Innovation

RESEARCH

Open Access

Analysis of the cryptocurrency market using different prototype-based clustering techniques

Luis Lorenzo* and Javier Arroyo

*Correspondence:
luislore@ucm.es
Universidad Complutense de Madrid Facultad de Estudios Estadisticos Madrid, Madrid, Spain

Abstract

Since the emergence of Bitcoin, cryptocurrencies have grown significantly, not only in terms of capitalization but also in number. Consequently, the cryptocurrency market can be a conducive arena for investors, as it offers many opportunities. However, it is difficult to understand. This study aims to describe, summarize, and segment the main trends of the entire cryptocurrency market in 2018, using data analysis tools. Accordingly, we propose a new clustering-based methodology that provides complementary views of the financial behavior of cryptocurrencies, and one that looks for associations between the clustering results, and other factors that are not involved in clustering. Particularly, the methodology involves applying three different partitional clustering algorithms, where each of them use a different representation for cryptocurrencies, namely, yearly mean, and standard deviation of the returns, distribution of returns that have not been applied to financial markets previously, and the time series of returns. Because each representation provides a different outlook of the market, we also examine the integration of the three clustering results, to obtain a fine-grained analysis of the main trends of the market. In conclusion, we analyze the association of the clustering results with other descriptive features of cryptocurrencies, including the age, technological attributes, and financial ratios derived from them. This will help to enhance the profiling of the clusters with additional descriptive insights, and to find associations with other variables. Consequently, this study describes the whole market based on graphical information, and a scalable methodology that can be reproduced by investors who want to understand the main trends in the market quickly, and those that look for cryptocurrencies with different financial performance.In our analysis of the 2018 and 2019 for extended period, we found that the market can be typically segmented in few clusters (five or less), and even considering the intersections, the 6 more populations account for 75% of the market. Regarding the associations between the clusters and descriptive features, we find associations between some clusters with volume, market capitalization, and some financial ratios, which could be explored in future research.

Keywords: Fintech, Unsupervised machine learning, Cryptocurrency, Electronic market, Clustering, Investment portfolios

Introduction

The cryptocurrency market comprises more than 4000 cryptotocoins,[1] with over 800 trades per second, and more than 280 exchanges. It has become a huge new market in the very short term, considering that *Bitcoin* (Nakamoto 2009), the first peer-to-peer and decentralized digital currency was produced in 2008, and the first Bitcoin was mined in 2009. While cryptocurrencies were originally created to enable anonymous wire transfers and online purchases, they have become a powerful investment tool.

However, this new market is diverse. Cryptocurrencies with different technologies, purposes, and user bases, coexist, and form a highly heterogeneous market that is difficult to understand and manage, for those addressing a good investment allocation.

Regarding other assets, the value of cryptocurrencies swing based on news events. However, cryptocurrencies have no physical assets, or governments to return their value to. Moreover, the cryptocurrency market is new, based on a still developing technology that is highly speculative and small compared to others. Consequently, it is highly volatile with large upswings, bubbles, and sudden market downturns.

Being a market so novel, big, diverse, and volatile, it needs to be clearly understood. So far, several categorization efforts have been made. For example, the *Cryptocompare* website[2] analyzed over 200 cryptoassets, according to regulatory aspects, level of decentralization, supply issuance, economic incentives, and others. Such a taxonomy is useful, even if it only covers approximately 5% of the existing cryptocurrencies at that time. Another example, Burniske and Tatar (2017) classifies over 200 cryptocurrencies into three classes of assets, based on traditional financial markets, namely capital, consumable/transformable, and store of value assets. However, this classification is highly subjective, because many times, the cryptocurrencies may be an integration of some of them. Furthermore, these approaches typically cover a small fraction of the cryptocurrencies, which are the most important in terms of volume and popularity and focus on qualitative aspects or aspects that change insignificantly.

A different approach involves analyzing the financial performance of cryptocurrencies, and describing them from a statistical point of view. Chan et al. (2017) analyzed a few cryptocoins (Bitcoin, Dash, Dogecoin, Litecoin, MaidSafeCoin, Monero and Ripple), which exhibited heavy-tailed distributions, that fitted the generalized hyperbolic distributions. Hu et al. (2019) analyzed the stylized facts, and the return properties of 222 cryptocurrencies, and found a large degree of skewness, and volatility in the population of returns. Furthermore, according to Pele et al. (2020), cryptocurrencies can be clearly separated from classical assets, mainly owing to their tail behavior, high variance, and high departure from normality. However, their results also show that the behavior of cryptocurrencies is diverse.

The same conclusion can be drawn from other clustering analyses using cryptocurrencies. Stosic et al. (2018) represent the correlations of 119 cryptocurrency markets as a complex network, and discover distinct community structures in its minimum spanning tree. Song et al. (2019) analyze 76 cryptocurrencies using the

[1] Although cryptoassets is a more general term, as explained in Burniske and Tatar (2017), we use cryptoasset, cryptocoin, and cryptocurrencies terms indistinguishably

[2] https://cryptocompare.com

correlation-based clustering, and filtering out the linear influences of Bitcoin and Ethereum, and detect 6 clusters, but that do not remain stable after the announcement of regulations from various countries. The time dimension also plays an important role (Sigaki et al. 2019) clustering 437 time series of cryptocurrencies, using hierarchical techniques that detect four different groups with a behavior that evolves differently, in terms of efficiency for the information.

All these approaches show that it is possible to establish different groups of cryptocurrencies in terms of their financial performance. Additionally, it is useful to better understand the cryptocurrency market, but also to build a diversified portfolio. In the same way, they use different representations of the cryptocurrencies: correlations (Stosic et al. 2018; Song et al. 2019), factors extracted from the correlation matrix (Pele et al. 2020), and time series (Sigaki et al. 2019). Each representation focuses on different aspects of cryptocurrency that are meaningful to the purpose of the analysis.

However, it would be possible to combine the clustering results using different representations of cryptocurrencies, where each consider different aspects of cryptocurrencies. Thus, the combination of the clustering results makes it possible to characterize each cryptocurrency in several dimensions, one for each cluster strategy. If the clusters for each cluster strategy are meaningful, their combinations would offer a more detailed characterization of the market and useful insights for portfolio management.

This study aims to propose a new methodology that help us to arrange and understand the main trends of the market at a glance, based on the financial behavior of cryptocurrencies.

Each of the clustering methods considered should offer a complementary view of cryptocurrencies, and a meaningful graphical representation that makes it possible to observe the main characteristics of each segment of the market at a glance.

The combination of the clustering methods will make it possible to profile each cryptocurrency, considering the different clusters to which it belongs. Furthermore, the most populated clustering intersections will help us detect the main trends in the market. In conclusion, the clustering analyst can spot the intersection that has a particular financial behavior by choosing the prototype of interest for each clustering method. This can help investors address interesting cryptocurrencies, depending on the investment profile.

In conclusion, the methodology includes the study of the associations between the clustering results, and other cryptocurrency features not considered in the clustering. Thus, the methodology helps to discover potential relationships between the resulting clusters, and other aspects beyond the data characterization used in the clustering methods.

The proposed methodology is fully supported by different statistical tools that ensure the robustness of the results. Further, it is easily scalable, to manage a growing, and dynamic market. Regarding computations, we used R (R Core Team 2013) and several R libraries, as shown in Table 15.

In our study, we include all cryptocurrencies in the market in 2018 (more than 1,700 cryptocurrencies), going beyond the few dozens (or few hundreds) of cryptocurrencies analyzed in other studies.

Regarding the data characterization, we describe each cryptocurrency considering the log-return transformation of the daily price in 2018 based on three different levels of granularity:

- Mean and standard deviation of the daily returns
- Distribution of the daily returns
- Time-series of the daily returns

In the first case, we provide a meaningful summary commonly used to describe financial assets over time, as it is the annualized return and volatility, or with the central tendency and dispersion of returns. In the second case, we consider the whole distribution of returns that account for not only the central tendency and dispersion of an asset, but also for the whole aggregated behavior including asymmetry, kurtosis, and tails. The methods for analyzing distributional data belong to the field of symbolic data analysis (Noirhomme-Fraiture and Brito 2011), where observations account for internal variation that can be represented as intervals or distributions, and have been previously used in finance (Arroyo and Maté 2009; Arroyo et al. 2011; González-Rivera and Arroyo 2012). In conclusion, we consider the observed data, that is, the log return time series that accounts for variations over time and makes it possible to identify when volatile or stable periods occur in each cryptocurrency.

There is a high diversity of clustering techniques. However, in our case, interest lies in the different perspectives shown at each level of granularity. Therefore, for all representations, we use partitional prototype-based clustering algorithms with a similarity measure (distance), that is meaningful for each kind of representation. Thus, we will have a prototype describing the behavior of each cluster using the same representation of the data. Prototypes make it possible to assign financial meaning to the entire cluster.

We further combined the three clustering results and analyzed the most numerous intersections with the help of visual tools. Such approaches have been successfully used in biostatistics (L'Yi et al. 2015; Kern et al. 2017). In our case, we use them to represent the main trends in the cryptocurrency market. If several cryptocurrencies belong to the same clusters in the three clustering results, we can consider them to be very similar. We further inspect the relationships among the three clustering results with the help of visualization tools.

The proposed approach provides a screening mechanism that allows us to explore the entire market, despite its complexity and size. The intersection of the clustering results can also help investors in selecting a suitable cryptocurrency for the portfolio, as it characterizes the market in more detail.

In a further step, we investigate the association between the clustering results and different features of cryptocurrencies, such as technological variables, market capitalization, the *maturity* (age) of the cryptocurrency, and some of the asset portfolio ratios. We aim to inspect whether some clusters are tightly associated with some aspects that do not consider the clustering process. We conducted inference statistical tests, to assess whether the associations were significant. These associations enhance the profiling of the different clusters. We keep continuous references to concrete cryptocurrencies of the market, where most of them are not known, which are part of our analysis. In conclusion, we discuss our results and exemplify how they could be used, and further present our conclusions, including some points, to stimulate further research.

Literature review

Clustering financial data

Clustering analysis is a well-known data analysis tool that has been used in different fields (Henning et al. 2016). Particularly, in Finance, the seminal work of Mantegna (1999) used the cross-correlation of the return time series, and minimum spanning trees (MST) to group the stocks of the New York Stock Exchange from, 1989 to 1995. Mantegna (1999) applies the MST to represent the stock market as a network. Bonanno et al. (2004) further applied the same methodology considering different time horizons, to compare the return and volatility networks. The methodology of Mantegna is applied with different variations in other contexts (Onnela et al. 2003; Mizuno et al. 2006; Brida and Risso 2009). Furthermore, Marti et al. (2017) proposed different alternatives and variants for this methodology.

Another important strand applies fuzzy clustering to financial time series, typically grouping stocks to develop portfolios. For example, D'Urso et al. (2013) and D'Urso et al. (2016) applied a model-based approach with different variations of *fuzzy* clusters, to financial markets for different distance metrics (autorregresive, Caiado). Similarly, D'Urso et al. (2020) proposed a fuzzy clustering method based on cepstral representation, using the daily Sharpe ratio as a variable of clustering.

The main application of clustering in finance is building portfolios. For example, Nanda et al. (2010) applied K-means, fuzzy C-means, and self-organizing maps (SOM), to returns and financial ratios from Indian stocks, to classify them into different clusters and subsequently develop portfolios from these clusters. Chaudhuri and Ghosh (2015) propose an approach that groups the daily Indian market volatility comparing Kernel K-means, SOM and Gaussian clustering models to achieve right volatility prediction using the clusters as predictors.

Liao (2007); Liao and Chou (2013) cluster the daily market data, and apply different association rules between the K-means groups, indices, and some market categories. These associations help analyze and describe co-movement among the different markets.

Regarding the use of time series as objects for clustering, Aghabozorgi and Teh (2014) proposed a three-phase clustering model, to categorize companies based on the shape similarity of their stock markets, using dynamic time warping (DTW) (Berndt and Clifford 1994). D'Urso et al. (2019) apply a trimming procedure to a fuzzy clustering of stocks comprising the FTSE MIB with a DTW as a distance metric with good results, to mitigate the outlier effect on time series.

From traditional finance to cryptoasset markets

Yermack (2013) analyzes Bitcoin market in-depth, and consider it an investment that is more speculative than a currency. Apparently, Bitcoin market poses high risk for the management of transactions, and credit markets. In conclusion, a deflationary scenario is anticipated owing to the limited number of bitcoins that can be issued (21 million). This study anticipated many aspects of cryptocurrency markets prevalent today (excessive volatility and high level of computer knowledge required for using and integrating them into the web of international payments). A more updated vision of this innovative market regards cryptocurrency exchange. Drozdz et al. (2018, 2019) show that BTC/USDT and ETH/USDT ETH/BTC were almost indistinguishable from exchange rate quotes in the forex market. The authors show that the exchange of cryptocurrencies has a behavior similar to that of more mature markets such as stocks, commodities, or Forex. Complementarily, the latest study Drozdz et al. (2020a) points to the anticipated disconnection of the cryptocurrency from the conventional markets, and states that the Bitcoin on the cryptomarket plays a role similar to that of the USD in the Forex market, or Drozdz et al. (2020b), where cryptocurrencies began to be correlated with traditional assets, only from 2020.

Corbet et al. (2019) analyzed the high growth of the cryptocurrency market and its heterogeneity since 2014 in depth. They consider different aspects including regulatory, cyber-criminality, market efficiency, and bubble dynamics, and make recommendations for further investigations on different domains. We consider a couple of them, and address some characteristics based on liquidity with the volume as a proxy, market cap, and other key metrics or ratios, such as the beta or Sharpe ratio. More recently, Fang et al. (2021) updated a survey covering various cryptocurrency trading aspects, including unsupervised machine learning techniques and others (e.g., cryptocurrency trading systems, bubble, and extreme conditions, prediction of volatility and return, crypto-assets portfolio construction and crypto-assets, technical trading, and others).

The characterization of cryptocurrencies from a statistical point of view has been tackled by different studies. Chan et al. (2017) analyze the distributions of a few cryptocurrencies (Bitcoin, Dash, Dogecoin, Litecoin, MaidSafeCoin, Monero and Ripple) and show that they exhibit heavy-tailed distributions that fit the generalized hyperbolic distributions. Our study considers heavy-tail and associated power-law distribution analyses. As part of a benchmark with other markets, Baek and Elbeck (2015) show that Bitcoin market volatility is 26 times more volatile than the S&P 500 Index.

Zhang et al. (2018) analyzes the stylized facts of eight cryptocurrencies that represent almost 70% of the market capitalization and find, that among other things, heavy tails for the returns, return autocorrelations that decay quickly, while the autocorrelations for absolute returns decay slowly, whose returns display strong volatility clustering, and leverage effects, and a power-law correlation between price and volume. The study of stylized facts has been extended by increasing the number of digital coins to 222 (Hu et al. 2019). Similarly, we consider it important to include as many cryptocurrencies as possible in our study, to characterize the market fully.

Clustering of cryptocurrencies

The classical methodology based on MST algorithms (Mantegna 1999) is applied by Song et al. (2019), to filter out the influence of *Bitcoins* and *Ethereum*; it detects six homogeneous clusters. However, the structure found does not remain stable after the announcement of regulations from various countries. Interestingly, the use of clustering together with other methods, such as VAR models and Granger causality tests (Zieba et al. 2019) show that Bitcoin shock prices are not transmitted to the prices of other cryptocurrencies, with *Litecoin* and *Dogecoin* being the most influential actors. According to the results, Bitcoin exhibits a lower relationship with other cryptocurrencies.

Another approach is the use of the random matrix theory, and hierarchical structures in an MST on 119 cryptocurrencies, from 2016 to 2018 (Stosic et al. 2018). They find multiple collective behavior in the cryptocurrency market, which contrasts with the intuitive idea that Bitcoin has a global influence on the entire market.

Furthermore, the time dimension was also considered. Sigaki et al. (2019) first classify 437 cryptocurrencies according to information efficiency, using permutation entropy and statistical complexity, and then cluster their time series using dynamic time warping and hierarchical clustering, to find four groups where the behavior in terms of information efficiency evolves differently.

All these articles show the complexity of the underlying structure in the cryptocurrency market, where some cryptocurrencies influence others, even in unexpected ways.

The comparative study of cryptocurrency markets and traditional financial markets is also a key research area. Corbet et al. (2018) show that cryptocurrencies are highly connected among themselves, and disconnected from mainstream assets (bonds, stocks, S&P500, gold). Consequently, Pele et al. (2020) merged classification based on asset profiles and the dynamic evolution of clusters. First, they characterize a selected group of log-returns assets, including 150 cryptocurrencies, stock commodities, and exchange rates, to estimate a multidimensional vector by applying a dimensionality reduction with factor analysis. They further used classification, where K-means is one of the techniques applied. The main difference between cryptocurrencies and traditional assets is the higher variance and longer tails of the log-return distribution. The work also shows that individual cryptocurrencies tend to develop over time, with similar characteristics (synchronic evolution).

Methodology

Dataset description

We retrieved data from https://www.cryptocompare.com/ for all cryptocurrencies traded in 2018. Many new cryptocurrencies have emerged in recent years, but many of them are short-lived and barely traded. We aim to include as many of them as possible in our study. First, we eliminate *NaN* and *Inf* values, which are mostly caused by zero prices in log transformations. Second, we filter out cryptocurrencies that were in the market less than 95% of the days (92 cryptocurrencies in 2018). We kept for clustering, those that were in the market but were not traded, that is, zero return and volatility or zero volume, as they are part of the market. In 2018, there were 306 cryptocurrencies on exchanges, that were barely traded. However, we decided to include them in the clustering as they are a substantial part of the cryptocurrency market. Even if they have no interest in investors, we are interested in knowing where they are allocated.

Our final dataset in 2018 consisted of 1,723 cryptocurrencies. However, we decided to eliminate cryptocurrencies with low or no activity from the second part of our analysis, the association tests. Low market activity may cause heavy tails in the return distribution and affect the consistency of the results. The remaining dataset for association tests consisted of 1,262 cryptocurrencies with higher statistical quality, ensuring the existence of the first and second statistical moments.

We also downloaded the data for 2019, to extend our experiment for a longer timeframe, analyzing the generalization of the results.

In addition to cryptocurrency data, we use daily data from CCI30[3] to represent the global behavior of the cryptocurrency market. The CCI30 is a market cap weighted index (Rivin and Scevola 2018), which represents the 30 largest cryptocurrencies by market capitalization, which makes it a good representative of the market. However, other crypto indexes (such as, CRIX or BCGI that stands for Bloomberg Galaxy Crypto Index) could be used in our methodology, as all of them highly correlate with the market (Häusler and Xia 2021). We chose the CCI30 index owing to its data availability and transparent methodology. However, the proposed methodology could be used with other indices, provided that it ensures an accurate representation of the trend for the entire market. We also retrieve data from the US Department of the Daily Treasury Bill market (T-Bill).[4] We use both for the computation of some financial benchmarking rates as the Beta and Sharpe ratio, which we explain in the subsequent sections.

Regarding the cryptocurrencies, we constructed the following variables:

- **Daily log-returns**: The use of returns instead of prices in Finance price time-series is very extended and consolidated, owing to its more suitable statistical properties and better comparability. It has also been used in cryptocurrency markets Letra (2016), Stosic et al. (2018). The return for cryptocurrency i on day t is computed as:

  $$r_i(t) = ln(P_i(t)) - ln(P_i(t-1))$$

  where $P_i(t)$ is the daily cryptocurrency price for the i cryptoasset on day t.
- **Heavy tail**: Heavy tail behavior in a return distribution means that extreme price fluctuations are relatively frequent. This might be related to the finite-size effects in the number of active agents, linked to the liquidity and volume of the market (Watorek et al. 2020). The rates of return distributions for less liquid cryptocurrencies are characterized by thicker tails, and poorer scaling.[5] We aim to identify cryptocurrencies prone to extreme behavior and whether they associate with some clusters. We define a cryptocurrency with heavy tail behavior by a binary variable if it has a tail index lower than 2, according to Newman (2005)

  This would question the existence of the finite first, and the second moments of the underlying distributions, which is not a problem in our case, as we use the observed sample statistics in a descriptive manner.
- **Volume**: The daily traded volume in the units of the base cryptocurrency, is used as a liquidity proxy. We transform the volume into an ordinal variable using the quantile functions. Three cryptocurrencies represent 66% of the trading volume of the market in 2018, namely Bitcoin (46%), Ethereum (16.5%), and EOS (4%); in total, 10 cryptocurrencies (`BTC, ETH, EOS, BCH, XRP, LTC, ICX, HSR, ETC, IOT`) represent 80% of the daily volume.
- **Market cap**: it is the one-day market capitalization of February 4, 2019. Three cryptocurrencies represent 60% of the *market cap*: `WBTC*` (26.8%), `BTC`(22.4%), and `NPC`

[3] https://cci30.com/.

[4] https://home.treasury.gov/.

[5] A power-law distribution is also referred to as a *scale-free distribution* because a power-law is the only distribution that is the same, regardless of the scale (Newman 2005).

(11.5%), and five cryptocurrencies (`WBTC*, BTC, NPC, XRP, AMIS`) represent 80% of the total *market cap*.

- **Beta and Sharpe ratios**: We compute and discretize *Beta* and *Sharpe ratio* for each cryptocurrency. These variables enrich the characterization and give us a financial flavor of the clusters that will help us with interpretability.
- **Technological variables**: We represent the encryption, and consensus algorithms of the cryptocurrency as nominal variables:
  - **Encryption**: There are 105 different values. The most relevant are *Scrypt*, *SHA256*, *SHA256D*, *X11*, *X13*, *X15*, *PoS*, *Multiple*, and *CryptoNight*. We notice that this information is not available for 35% of cryptocurrencies (599 obs.) in 2018.
  - **Consensus**: There are 60 possible values, including the well-known Proof of Work (PoW) and Proof of Stake (PoS). The most predominant are obviously *PoW/PoS*, *PoW*, and *PoS*, although this information is missing in 31% of the cryptocurrencies (536 obs.) in 2018
- **Age**: We estimate the time on the market of each cryptocurrency, and transform it into an ordinal variable, by a quantile function. *Age* and *maturity* terms are interchangeable in our study.

Methods

We aim to group the cryptocurrencies based on the behavior of their log-returns in 2018, which will be described later. For this purpose, we use different clustering algorithms that deal with the three representations of the log-returns, described in the previous section: statistic moments, observed probability distribution, and observed daily time series.

We use centroid-based clustering algorithms as the centroids provide an interpretable summary of the elements of each cluster, which will help us identify the most relevant features of the cluster elements. However, this type of clustering algorithm assumes knowledge about the desired number of clusters (k), which is a drawback. We applied different quality criteria, to determine the optimum number of clusters, depending on the technique used. The evaluation of clustering performance is intrinsically difficult, owing to the lack of objective measures—no true table. Moreover, different approaches have been applied, to compare different clustering techniques, for instance, by applying the multiple criteria decision making (MCDM) in Kou et al. (2014) with different methods (i.e., TOPSIS, DEA, and VIKTOR) including 11 performance measures. Particularly, in our case, and for K-means, as we will detail later on, we mostly rely on the straightforward *majority rule* criteria implemented in the R-package (Charrad et al. 2014), which applies 30 performance measures, which is a simpler methodology than MCMD, but adequate in our case, as we do not benchmark different clustering techniques.

Moreover, we use distance-based clustering algorithms, which are simple, intuitive, and applicable to a wide variety of scenarios (Aggarwal et al. 2013). The algorithms considered are based on meaningful dissimilarity measures or distances that help in the interpretability of the clusters. This is especially important for more complex representations, such as distributions or time series. For example, in the case of distributions, the measure should relate to the properties of the density function (central tendency, spread, symmetry), while in the case of time series, it will be more with the shape of the time series. Meaningful measures will help us better understand the resulting clusters, and interpret the nearness of the observations to the centroid. Additionally, our clustering algorithm provides a prototype or centroid of the clustering, which facilitates the characterization of the resulting clusters.

The cluster intersections help us merge the results of the different clusters, and identify the most prominent cryptocurrency profiles in 2018, according to different characteristics through the three techniques. Furthermore, we analyzed the association between the clustering results found for the three representations, and the different attributes of cryptocurrencies.

K-means clustering algorithm for the first and second statistical moments

Regarding the bi-variate (or two-moments) representation, where the two variables are the yearly mean, and standard deviation of the log-returns, we use K-MEANS clustering (MacQueen 1967), which is one of the most extensively used clustering algorithms (Wu et al. 2008) globally and on cryptocurrency markets, particularly (Fang et al. 2021). We standardized the two variables to homogenize the differences between their ranges. K-MEANS clustering minimizes within-cluster variances, that is, squared Euclidean distances in our case, which makes the result easy to understand and interpret. Before clustering, we compute the Hopkins statistic (Banerjee and Dave 2004) to rule out the possibility that a uniform random distribution generated the dataset.

To select the number of clusters (k), we compute several internal cluster validity indices (CVIs) for crisp partitions (Arbelaitz et al. 2013), including Silhouette, Dunn, COP Davies-Bouldin, Calinski-Harabasz, or the score function, and then apply the *majority rule* to choose the best number of clusters.

We apply clustering ensemble techniques (Acharya 2011) to reduce the randomness of partitional cluster results. We run the K-means algorithms 10 times, and ensemble the outcomes by minimizing the Euclidean distance. We confirm that the dissimilarity among the different runs is closer to zero, which makes the ensemble cluster a more stable representation. For each algorithm run, we apply the Hartigan-Wong method for clustering (Hartigan and Wong 1979) with ten iterations, to reach convergence and consider 50 random starts for each iteration. Once we have the 10 algorithm runs, we compute the medoid of an ensemble of partitions, that is, the element of the ensemble minimizing the sum of dissimilarities to all other elements (Hornik 2005, 2019).

Dynamic clustering algorithm for histograms

Regarding the yearly log-return distribution, we apply a clustering algorithm that deals with the histogram-data form. More precisely, we apply the dynamic clustering algorithm for histogram data based on the l_2 Wasserstein distance (Irpino and Verde 2006; Irpino et al. 2014). Thus, we group the cryptocurrencies with similar distributions of log-returns in 2018.

The dynamic clustering algorithm needs a dissimilarity function to assign the observations to the clusters, which is the l_2 Wasserstein distance. Given two histograms h_1 and h_2, the l_2 Wasserstein distance is defined as

$$d_W(h_1, h_2) := \sqrt{\int_0^1 \left[F_1^{-1}(t) - F_2^{-1}(t)\right]^2 dt} \tag{1}$$

where F_1^{-1} and F_2^{-1} are the inverse of the cumulative distribution functions, that is, the quantile functions of h_1 and h_2, respectively. This distance can be decomposed as follows:

$$d_W(h_1, h_2) = \sqrt{(\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2\sigma_1\sigma_2\left(1 - \rho_{1,2}\right)} \tag{2}$$

where μ_i and σ_i are the mean and the standard deviation of the h_i respectively, and $\rho_{1,2}$ is the correlation of h_1 and h_2 (Irpino and Verde 2015). Cpnsequently, the l_2 Wasserstein distance can be decomposed by adding three elements that account for the histogram differences in terms of location, spread, and shape. Interestingly, this distance matches the perceptual similarity that humans observe when comparing distributions (Arroyo and Maté 2009). All these aspects make it a suitable distance for clustering distributions and, in our case, log-return distributions.

The dynamic clustering algorithm for histogram data based on the Wasserstein distance (Hist-DAWass) is a k-means-like algorithm for clustering a set of observations described by histogram variables (Irpino and Verde 2006; Irpino et al. 2014). Each of the k clusters is represented by a centroid or prototype, and observations are assigned to the closest prototype. The prototype is an average histogram of the histograms observed for each variable. In our case, observations are described by a single histogram variable representing the distribution of log-returns, and the resulting prototype is a histogram that averages the histograms of the observations that belong to the cluster (Irpino and Verde 2015). Consequently, the prototypes can be interpreted in a financial context as log-return distributions.

We used the clustering implementation in the R-package Hist-DAWass (Irpino 2016). This implementation provides a quality measure, which is the percentage of the sum of squared (SS) deviation explained by the model running the algorithm several times for each k. We run the clustering algorithm 20 times for each k, of which the solution is the best among the repetitions, that is, the one that maximizes the SS.

TADPole clustering for time-series

Time-series clustering is a challenging domain for clustering owing to the high dimensionality of objects and their ordering. Consequently, many approaches have been proposed over time (Liao 2005; Rani and Sikka 2012; Aghabozorgi et al. 2015).

We aim to cluster the time series with similar volatility patterns in the same period. For this purpose, the Euclidean distance may fail to produce an intuitively correct measure of similarity between two time series as it is very sensitive to small distortions in the time axis. However, other measures, such as dynamic time warping (DTW), manage this problem using *warping* the non-linearly of the time dimension, to estimate their similarity. Currently, DTW is considered one of the most popular and useful shape-based measures (Aghabozorgi et al. 2015).

However, DTW is intrinsically slow owing to its quadratic time complexity, which hampers its applicability in clustering. Therefore, we use the enhanced DTW algorithm TADPole (Time-series Anytime Density Peak) (Begum et al. 2015), which extends the density peak (DP) clustering framework (Rodriguez and Laio 2014) and exploits the upper and lower bounds of DTW, to prune unnecessary distance computations, which accelerates the convergence of the algorithm. Consequently, TADPole produces a correct answer quicker, and then refines it until it converges to the exact answer. Moreover, the clustering algorithm only requires two parameters, which makes it easy to use. First, a cut-off distance that defines the thresholds to select the series. We further set it as 2; and second, a window size that defines the time frame to make the comparison between the series that we set as 3. Optionally, we can also select the number of clusters (k), or let the algorithm choose the optimal one, based on the local density of points (closer series at some time based on some cut-off distance) using a "knee point finding" algorithm, where points with higher values of $\rho_i \cdot \delta_i$, where ρ_i refers to the local density and δ_i is the distance from points with higher local density.

We consider a different number of clusters k, and compute the internal cluster validity index (CVI) for each cluster. As this clustering algorithm uses three distances, we use Calinski-Harabasz as the CVI index to secure the convergence of the algorithm for the asymmetric distance measure.

TADPole allows for the clustering of time-series with arbitrary shapes, which is very useful in our case owing to the heterogeneity of the cryptocurrency market. In contrast, DTW is not a geometric distance with the three fundamental metric properties: non-negativity, symmetry and triangle inequality. TADPole clusters cannot be represented as "balls" in a metric plane, as in K-means. The result is a partition around the medoid (PAM) type centroid, using the DTW similarity measure that can be represented only in a *DTW* space. This centroid is a time-series that helps to identify the volatility patterns of the resulting clusters.

We apply the implementation of the TADPole algorithm of the R-libraries DTWCLUST by Sarda-Espinosa (2019); Sardá-Espinosa (2019). The time-series are log-return values that facilitate the characterization of the clusters from a financial perspective. The DTW measure implemented in the package follows the estimation in Lemire (2008).

Combination of clustering results

Once we have the results of the clustering algorithms, we combine them by intersecting the clusters. Potentially, we have $T_1 \cdot T_2 \cdot \ T_3$ intersections, where T_n is the number of clusters obtained for the clustering algorithm n. The combination of the clustering results makes it possible to characterize each cryptocurrency in several dimensions, one for each cluster strategy. The resulting multidimensional categorical datasets can be shown using visualization techniques supported by graph theory (L'Yi et al. 2015; Kern et al. 2017). To better highlight the changes in the clustering between the different techniques, we visualized such changes by means of a so-called alluvial diagram, which is considered a good example in Rosvall and Bergstrom (2010). We use the alluvial visualization implemented in R (Bojanowski and Edwards 2016) to show the main *flows* of cryptocurrencies.

We can also numerically compare two partitions represented as a $c_1 \times c_2$ matrix, where n_{ij} is the number of objects in group i of partition 1 ($i = 1, ..., c_1$) and group j of partition 2 ($j = 1, ..., c_2$). The labeling of the two partitions was arbitrary. Hubert and Arabie (1985) developed the *Adjusted Rand Index* (ARI) with a correction for chance as

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2} / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2} / \binom{n}{2}} \tag{3}$$

The index computes the proportion of the total of $\binom{n}{2}$ object pairs that agree, that is, they are either (i) in the same cluster according to partition 1 and the same cluster according to partition 2, or (ii) in different clusters according to partition 1 and in different clusters according to partition 2. The higher the ARI index, the higher the agreement.[6] In our case, this means that more cryptocurrencies share clusters for the different partitions. We used the function implemented in the R package MCLUST by Scrucca et al. (2016). We also focus on the cluster intersections with higher cardinality for a better profiling of the main trends of the cryptocurrency market.

Association test

In conclusion, we enhance the descriptive information of each cluster by examining the level of association with different independent variables not considered by the clustering algorithms. We analyze the association among clusters and the categorical variables defined in Table 1 by applying Fisher's exact tests, and analyzing the Pearson's residuals of the contingency tables that we explain later. First, quantitative variables must be transformed into ordinal by quantile functions.

Below, we introduce some variables that are financial ratios, which we borrow from the portfolio theory (Bacon 2008), and apply it to characterize the behavior of cryptocurrencies from an investor perspective, enhancing the association study as well. We take advantage of the R-library PerformanceAnalytics by Peterson et al. (2018), for the computation of the Sharpe ratio.

Beta is a volatility measure of the systematic risk of an asset, the risk inherent to the entire market that is non-diversifiable, in statistic terms, the *beta* is the slope of the regression of our asset compared with a reference on the market:

$$\beta = \frac{Cov(R_c, R_b)}{Var(R_b)}, \tag{4}$$

where R_c is the return of our cryptocurrency, R_b is the return of the benchmark market, and the CCI30 index that tracks the 30 largest cryptocurrencies by market capitalization.

The *Beta* value shows whether an asset moves in the same direction as the reference index, and how volatile or risky it is compared to it. The *beta* for the entire market is 1.0. A positive beta means that the asset moves in the same direction as the market,

[6] The Rand Index yields a value between 0 and 1, but the Adjusted Rand Index can yield negative values if the index is less than the expected index.

Table 1 Categorical variables used on the association tests and values

Variable	# Levels	Values
Algorithm	73	Encryption algorithm (SHA256, Ethash, X13, X11,...)
ProofType	39	Consensus algorithm (PoW, PoW/PoS,DPoS..)
Volume	5	Percentiles of the volume negotiated. Namely, P_{70} for volume values lower than the P_{70} percentile, P_{80} for values higher than the P_{70} and lower than the P_{90}, and similarly P_{90}, P_{99} and P_{100}.
MkCap	5	Percentiles of the market capitalization. Namely, P_{70} for market cap values lower than the P_{70} percentile, P_{80} for values higher than the P_{70} and lower than the P_{90}, and similarly P_{90}, P_{99} and P_{100}.
Beta	6	Beta values divided into the following categories:
		NegBeta for beta values lower than -0.01
		CashLike if beta is to equal or higher than -0.01 and lower than 0.01
		LowVol if beta is equal to or higher than 0.01 and lower than 0.95
		Indexlike if beta is equal to or higher than 0.95 and lower than 1.05
		HighVol if beta is equal to or higher than 1.05 and lower than 100
		Extreme if beta is higher than 100
Sharpe	4	Sharpe ratio divided into the following categories:
		SRF (Small Risk-free) for negative values
		ERP (Excess return positive) for positive values lower than 0.5
		ACC (Acceptable) for values equal to or higher than 0.5 and lower than 1.0
		GOOD for values equal to or higher than 1.0
Age	7	Deciles of the age variable (time on the market). Namely, D_4 for age values lower than the P_{40} percentile, D_5 for values higher than the P_{50} and similarly D_6, D_7, D_8, D_9 and D_{10} for P_{100}.
HeavyTail	2	Binary variable that take value 1 if the cryptocurrency has a heavy-tail behaviour or 0 if it does not.

while a negative beta means that the asset moves in the opposite direction. Furthermore, an absolute value higher than 1 indicates greater sensitivity to systematic risk (i.e., higher risk), while values lower than 1 indicate less sensitivity.

The **Sharpe ratio**(Sharpe variable) is the average return of risk-free by volatility unit or total risk. The ratio determines the risk of investment with respect to the return of an investment with zero risk:

$$SR_c = \frac{E[R_c - R_f]}{\sigma_c}, \tag{5}$$

where R_c is the return of cryptocurrency, σ_c is the standard deviation or the volatility of our cryptocurrency, and R_f is the *risk-free* rate considered the reference; we considered the daily value of the annualized T-Bill over 90 days. Its daily value for 2018 was $E[R_f] = 0.00525\%$, which is almost zero. The greater the value of the Sharpe ratio, the more attractive the risk-adjusted return of cryptocurrency is.

Typically, the **chi-square test** is used to examine the significance of the association between categorical data on a contingency table. However, the significance value is an approximation that is not adequate when the sample size is small. We ruled out the chi-square test as the results are insignificant if the expected frequency is not typically higher than 5 in at least 80% of the cells of the contingency table (Yates 1984). This assumption is not fulfilled in our case, for many of the categorical variables, and for some levels. We used Fisher's exact test (Fisher 1922) to test the association between the variables in Table 1 and the cluster results, which are applicable for all sample sizes. This test assumes no dependency between the categorical variables as a null hypothesis, and assumes a multivariate hypergeometric distribution for the cells in the contingency tables (Mehta and Patel 1983).

For large datasets, the *Monte Carlo* method provides an unbiased estimate of the exact *p-value* (Mehta and Patel 1996). *Monte Carlo* comprises a repeated sampling method that for any observed table, there are many tables, each with the same dimensions and columns and row margins as the observed table. *Monte Carlo* simulations are implemented in R stats-package for the *chisq.test* function. We ran 8,000 simulations for each association, that is, for each pair of variables under analysis, generating simulated contingency tables filled with a sampling of a multivariate hyper-geometric distribution. We further compute the probability that we have a distribution, as we have effectively observed, that is, the *p-value*. A cell-by-cell comparison of the observed and estimated frequencies indicates the nature of the dependence. If the p-values of the Fisher association tests between a couple of variables are lower than 0.01, then we consider the association to be significant. For each significant association between categorical variables of the contingency table, we analyze standardized (adjusted) Pearson's residuals for cell *ij* (Agresti 2018), which are defined as follows:

$$r_{(Adj)ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}(1 - \frac{m_i}{N})(1 - \frac{n_j}{N})}} \tag{6}$$

where O_{ij} and E_{ij} are the observed and expected frequencies, respectively, m_i is the row total, n_j is the column total, and N is the total number of observations.

The sign of the residual (positive or negative) indicates whether the observed frequency in cell *ij* is higher or lower, respectively, than the value fitted under the model, while the magnitude indicates the degree of departure. A standardized residual having an absolute value that exceeds a value of approximately 2.00, when there are a few cells, or approximately 3.00 when there are many cells, indicates that the cell does not satisfy H_0 (Agresti 2018). In our case, we assume a more conservative position, and consider it a cut-off for significant standardized residuals that exceed 3.50.

Replication within a longer time-frame

We propose a time-frame agnostic methodology that aims to describe the behavior of a period of time, regardless of its length or frequency. We use 2018, an extremely active period in the cryptocurrency market, as the subject of analysis. However, it would be interesting to replicate the methodology for other time periods, and/or consider other time frequencies.

In this respect, we re-apply our methodology to an extended time frame that includes both 2018 and 2019, to validate the stability of the results obtained, and the robustness of the methodology. For this purpose, we consider an extended timeframe, including both 2018 and 2019.

We further consider only the cryptocurrencies traded on the market during the entire period (730 days), that is, 440 cryptocurrencies in total. For this particular shortlist, we compute the associations on the extended period for the financial ratios (Beta and

Table 2 Cluster cardinality, mean value and standard deviation of the centroid or prototypes for the clustering methods

	K-means			**Hist-DAWass**			**TADPole**		
	Card.	**Mean**	*Std.Dev.*	**Card.**	**Mean**	**Std.Dev.**	**Card.**	**Mean**	*Std.Dev.*
Clus. 1	19	− 0.008	1.795	496	− 0.134	0.337	22	− 0.001	0.080
Clus. 2	903	− 0.002	0.130	147	− 0.503	0.378	843	0.026	0.046
Clus. 3	801	− 0.009	0.229	1007	− 0.011	0.108	858	− 0.028	0.047
Clus. 4				57	− 0.044	0.867			
Clus. 5				16	− 0.095	3.123			

For Hist-DAWass and TADPole we compute the mean and standard deviation of the prototypes

Sharpe Ratio), Volume. However, the variables Algorithm, ProofType, and Age remain unchanged. For the case of MkCap, we do not have values at regular intervals. Hence, we use those that we took the 4th of February, 2019, as we explained in the variable description section.

We re-run the three clustering techniques for the 2018-19 timeframe. However, to determine the number of clusters, we confirmed that the results were quite similar to those obtained in 2018. Therefore, to ease the comparison, we chose to use the exact same number of clusters used in 2018.

This experiment will help us determine whether some of the underlying structures on the market persist when we consider a longer period and the same for the associations found.

Results

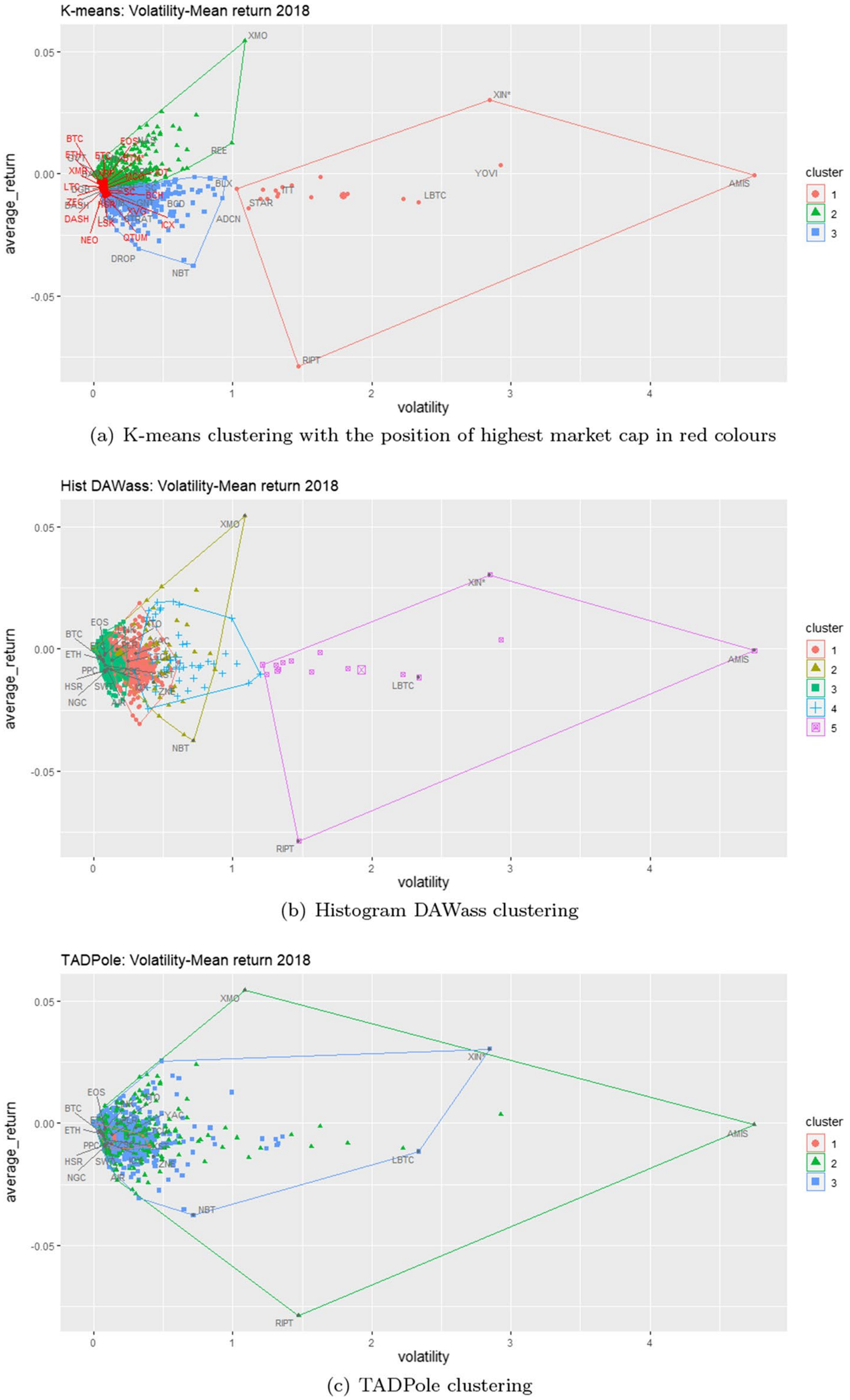
In this section, we present the results of the three clustering algorithms: intersection, clustering and association tests. In Table 2, we summarize the three clustering results, showing the cardinality of each cluster and, for the sake of comparison, the observed mean and standard deviation of the prototypes (for the K-means, we show the centroid values).

Clustering results of the bi-dimensional representations

Regarding the existence of clusters, the Hopkins statistic computed on scaled average returns, and volatility is 0.01552. The value is below 0.5, which indicates the existence of an underlying structure.

According to the CVI index, the optimum number of clusters was 3. The descriptive statistics of the three centroids in ordinary values are shown in Table 2; Figure 1a show the scatter plot of the clusters.

The clustering algorithm clearly discriminates the cryptocurrencies between lower (*Cluster 2* and *3*) and higher volatility (*Cluster 1*), which is the less populated cluster as well. From a financial perspective, *Cluster 1* includes riskier cryptocurrencies. *Cluster 3* mostly allocates negative mean returns, while those in *Cluster 2* have the higher returns, some of them positive and others negative. However, the three centroids are close to the zero-mean return point.



(a) K-means clustering with the position of highest market cap in red colours

(b) Histogram DAWass clustering

(c) TADPole clustering

Fig. 1 Volatility-Average return plane in ordinary values with the vertex names and the more representative cryptocurrencies in terms of market cap for the different clustering techniques for 1723 cryptocurrencies in 2018 time frame

Figures 3, 4 and 5 show a more detailed view of each cluster. In these figures, we represent the density in the bi-dimensional (two-moments) space $(\bar{r}, \sigma)$ in a contour plot that helps us to locate areas in which cryptocurrencies tend to be more concentrated.

- **Cluster 1**: This cluster allocates cryptocurrencies with negative average returns, but with very high volatility, ranging from 1 to 5. It includes only 19 cryptocurrencies that represent approximately 1% of the sample, as shown in Fig. 3a. The higher concentration of cryptocoins in this cluster is surrounding volatility 1.5, and the mean return is approximately -0.1, as shown in Fig. 3b. ELTCOIN (token that run on Ethereum blockchain network released in October 2017) has a central position in the cluster, and around it, we can find B2X, ADCN (no traded on the market since November 2019), BLX, WAND (derivative market platform), GOOD, SBIT, ZCG, ITT, REX, STAR (it is a token and operates on the Ethereum platform, higher volume in Ethereum along 1st and last quarter of 2018) and PFR. This cluster contains a mix of Ethereum tokens and cryptocoins with its own blockchain, most of them with low traded volume, which may cause a few operations to trigger volatility.
- **Cluster 2**: This cluster is the more populated, with approximately 900 cryptocurrencies (52% of the total). It allocates *moderate* behaviors, including higher mean return cryptoassets. It is also less homogeneous than the others, with different dense areas of concentration, as shown in Fig. 4a which point out the existence of other potential cluster. Most of the higher capitalization cryptocurrencies (BTC, EOS, ETC, ETH, and LTC) are in the sub-cluster with very low volatility, and moderate negative returns, as shown in Fig. 4b. However, we also find some cryptocurrencies with moderate positive returns (134 cryptotocoins), and very low volatility, as shown in Fig. 4c. These cryptocurrencies include ALEX (low trading in the first half of 2018, and higher activity in the second half of 2018), BST (BlockStamp had very low activity in 2018), ETL (EtherLite is an ERC20 token based on Ethereum with high peaks of activities in the first quarter of 2018, and no activity in the remaining part of the year), or OPES (OpesCoin had moderate activity in the first half of 2018, and were flat in the second); all of them can be considered low-medium market capitalization (under the 70^{th} percentile). However, the details of Fig. 4c, show the high homogeneity on the selected area where there is no any density contour curve.
- **Cluster 3**: This cluster has 801 cryptocurrencies, most of them with negative average returns, and a volatility lower than 0.5. According to Fig. 5a, the highest concentration of cryptocurrencies is located in a mean return closer to zero, and a volatility of approximately 0.1. Some of the more representative cryptocurrencies of this cluster in terms of market capitalization are XEM, VIA, QRL, DASH, QTUM, XST, and BCH, which are close to each other in the cluster (see Fig. 5b).

We confirm that K-means clearly identifies three different behaviors of cryptocurrencies, in terms of mean returns and volatility.

Clustering results of histogram representations

According to the CVI, the clustering algorithm for histogram data based on the l_2 Wasserstein metric (Irpino et al. 2014) separates the cryptocurrencies into five clusters. Each cluster is represented by its prototype, which is a log-return distribution. Table 3 shows the descriptive statistics of the prototypes of the five clusters.

The five distributions exhibit slightly negative central tendency measures, with Cluster 1 having the lowest values. They are quite symmetrical, with low skewness and heavy tails, as indicated by a high kurtosis. Skewness is closer to zero in all cases, but positive, which means that the right tail of the distribution is fatter; in other words, it has more extreme positive return values (or over the mean) on the right tail.

It is important to note that the coefficients of variation for the centroids are quite different for the clusters that range from -0.75 to -32.90, which indicates that this clustering algorithm is especially sensitive to this particular statistic. This is particularly relevant in the financial context as the coefficient of variation evaluates the degree of volatility assumed in comparison to the amount of return expected from investments. However, because the mean returns are negative, its financial interpretation would be misleading.

The last column of Table 3 provides a measure of variance (Var.Wass) that quantifies the deviation of the distributions of objects into a cluster with respect to its prototype. It is a dispersion measure for histogram data based on the L2 Wasserstein metric (Irpino and Verde 2015). This statistic measures how representative the prototype of a cluster is. According to this statistic, *Cluster 3* would be the more uniform cluster, whereas *Cluster 5* would be more heterogeneous.

The first column in Fig. 6 represents the prototypes of the five clusters, while the rest of the columns show some of the relevant cryptocurrencies of each cluster. Interestingly, except for the prototype of *Cluster 1*, the others exhibit a similar shape, where the main differences lie in the range of the distribution (note that each plot has a different range for the X-Y axis) and in the tail behavior. We describe them below:

- **Cluster 1**: The prototype in Fig. 6a has a mean return of -0.13, and the highest kurtosis (13.43). The standard deviation of this prototype is slightly lower than that of the *Cluster 1* prototype; however, the shape of the distribution is different, as the tails are heavier. The Wasserstein variance in Table 3 associated with the mean distribution (0.025) suggests that the cluster is homogeneous and has a cardinality closer to 500 cryptocurrencies, which represent approximately 30% of the samples. Some of the most representative cryptocurrencies in this cluster have a high market cap (*P99*), for example, `BITUSD` (high capitalization along 2018 but in a downward trend), `CHAT` in Fig. 6b, `KEY` in Fig. 6c (high trading volume in the second half of 2018), `MAN` (increasing trading volume along 2018, maximum at the end of the year) and `OCN` (a token for peer-to-peer sharing economies such as Airbnb).
- **Cluster 2**: The prototype shown in Fig. 6d (green color distribution) has the lowest mean (-0.50) and median (-0.51) return among the clusters, and the highest coefficient of variation (-0.75). The cluster variance (0.079) indicates a homogeneous cluster. This cluster has 147 cryptocurrencies and concentrates all non-traded cryptocurrencies (92) and the lowest market cap (*P70*) for most cryptocurrencies in this cluster. Representative into the cluster by the market cap is 365 in Fig. 6e, `ACN`, `CBX` or `ALT` in Fig. 6f.
- **Cluster 3**: The prototype shown in Fig. 6g has a mean return close to zero (−0.01) and the most moderated volatility (0.11), and the shortest observed range between

minimum and maximum returns. According to the Wasserstein variance, this cluster is the most homogeneous, which is especially interesting given that it has the highest cardinality with more than 1000 cryptocurrencies (around 60% of the sample). Unsurprisingly, this cluster allocates cryptocurrencies with the highest market capitalization, including BTC in Fig. 6h, BCH, EOS, ETC, ETH in Fig. 6i and others (HSR, ICX or LTC). Given the size of the cluster, these cryptocurrencies represent the predominant behavior in the market, which is unsurprisingly the most moderate behavior, and includes the most popular cryptocurrencies.

- **Cluster 4**: The prototype shown in Fig. 6j is characterized by negative mean returns (-0.04), notable volatility (standard deviation of 0.87) and fat tails with very high kurtosis (11.95). The coefficient of variation was also low (-19.97). The cluster was not very homogeneous compared with the mean distribution (0.128). The cardinality of this cluster is low (around 60 cryptocurrencies). Further, some of the representative cryptocurrencies are NAS (most of the trading volume in the 2^{nd} and 3^{rd} quarters of 2018), NKC (high trading volume since February 2018 and very important trading volume in August 18), POLY in Fig. 6k (launched in January 2018), FSN (higher volume activity in the 2nd and 3rd quarter of 2018 with a peak in August), JNT in Fig. 6l (higher volume activity in 3rd quarter) or MNTP (no continuity on the trading volume with sporadic peaks).
- **Cluster 5**: The prototype shown in Fig. 6m has a mean return closer to zero (-0.09), but the highest standard deviation (3.12), which causes the lowest coefficient of variation (-32.90). The shape of the cluster is almost symmetric (0.05) with moderate kurtosis compared with the other clusters (5.66). We find the highest negative and positive returns in this cluster. This cluster was the most heterogeneous compared with the mean distribution (1.116). Unsurprisingly, it has the lowest cardinality, with only 16 cryptocurrencies, which is approximately 1% of the sample. Some representative cryptocurrencies are B2X in Fig. 6n (low trading in 2018), ITT in Fig. 6o (very active trading volume in January 2018 and along and a peak in July but no activity since that time, no trading volume in 2019), LBTC (launched in 2017, discontinuous activity along 2018 with no activity at all from September to end of November), PFR, STAR (noted in *Cluster 2* of K-MEANS), YOVI, AMIX, ELTCOIN (noted in *Cluster 2* of K-MEANS) or FLLW (some activity the first 2-3 months of 2018, low trading volume in the remaining part of the year).

HIST-DAWASS clustering shows that it is possible to effectively discriminate the log-return distributions, considering central tendency, dispersion, and shape.

Clustering results of the time-series representation

TADPOLE clustering Begum et al. (2015) has better performance with a $k = 3$ value according to the Calinski-Harabasz index. Figure 7 represents the time axis, and the medoids of each cluster. Hence, they are observed objects (time series of cryptocurrencies). Figure 8 shows the annual and quarterly density functions of the three medoids. Also, Fig. 8a represents how different the density plot of the *Cluster 1* compared with the others corresponding with the higher volatile cryptocurrencies is.

- **Cluster 1**: The medoid of this cluster in Fig. 8b shows a time-variation of approximately zero, with return peaks positive and negative up to (-0.2, +0.2). The central part of the distribution is heavily concentrated around zero, but with extreme volatility. The quarterly average returns change smoothly, starting with a low but positive value in the first quarter, negative in the second and third, and positive in the fourth quarter. This cluster had the lowest cardinality (22 cryptoassets). The medoid of this cluster is the time-series LINK (Chainlink's native token, known as LINK, which is used to pay the network's node operators, or oracles, to provide secure data feeds). Other cryptocurrencies in this cluster are LTCU, PPC, SWT, AIR, NGC, PLR or ZSC.
- **Cluster 2**: The medoid of this cluster in Fig. 8c shows consistent average returns above zero. The density functions have three modes, which are greater than or equal to zero. However, the last two quarters of 2018 exhibited fat-negative tails with ranges over -0.1. The cardinality of this cluster represents approximately 49% of the cryptocurrencies, including some of the highest market cap cryptocurrencies BTC, HSR (noted in *Cluster 3* of Hist-DAWass), ICX (noted in *Cluster 3* of Hist-DAWass), LTC (noted in *Cluster 3* of Hist-DAWass), and XRP. The medoid is the cryptocurrency XTO (referred to as Tao coin, as well as a token for music streaming services).
- **Cluster 3**: The medoid of this cluster in Fig. 8d has average returns below zero in all the quarters; the densities further exhibit two modes smaller than or equal to zero, and occasionally large positive returns. The cardinality of the cluster was approximately 50% of the total. This cluster includes most of the remaining highest market cap cryptocurrencies (e.g., EOS, ETC, ETH). The medoid is the cryptocurrency ZNE (Zone coin with more trade activity in the first quarter of 2018, with the more important peak of trade volume in July, flat trading volume in the remaining part of the year).

Evidently, the TADPole clustering for the return time series effectively identifies three different clusters by considering the time series trend and dispersion over time. In Table 4, we show the variability of the clusters, measuring the variability as the mean distance (DTW+LB) to the centroid, and its standard deviation with LB as *Lower Bound*. The variability is quite similar in all clusters, with cluster 1 being the most homogeneous, and *Cluster 3* being the least. However, according to the standard deviation and coefficient of variation, the dispersion within the clusters is quite high.

Intersection of clusters

Regarding comparison, Fig. 1 shows the three clustering results on the same annual return-volatility plane. In each plot, all cryptocurrencies are located in the same location, but the color scheme in each plot represents the respective clustering results. In the plot, we marked the cryptocurrencies with the highest market capitalization and polygon vertices. Notably, most of them are located in a precise area below the point (0, 0). The polygons and colors reveal that the results of the three techniques overlap owing to the different dimensionality of the objects. The only exception is *Cluster 1* of K-means in Fig. 1a and *Cluster 5* of Hist-DAWass in Fig. 1b which are mostly the same. These plots confirm that each clustering algorithm considers different aspects of cryptocurrencies, and that their combinations may provide further insights into the cryptocurrency market. TADPole clustering in Fig. 1c is the more different in the groups, compared with previous techniques, overlapping all the cluster areas when represented on the same return-volatility plane than the other techniques.

We further analyze the main groups of cryptocurrencies that remain together through the three clustering algorithms, which we call the *intersection of clusters*. Only 24 out of 45 (3 × 5 × 3 intersections) are populated. Table 5 shows all the intersections, and those with a cardinality greater than 100 (first six intersections), representing 75% of the total market.

Intersection 1 and *2* have almost 300 cryptocurrencies each. Both are characterized by cryptocurrencies that belong to *Cluster 2* and *3* in the K-means and Hist-DAWass algorithms, which are the most populated clusters for each technique. Both are characterized by low volatility, and (negative) close to zero average returns. However, in *Intersection 1*, we find *Cluster 3* of the TADPole algorithm, whereas in *Intersection 2*, we find *Cluster 2*, which mainly differs from that in the first case, in that, it has negative quarterly average returns, while in the second case they are positive. In *Intersection 1*, we find cryptocurrencies such as EOS, GVT, MANA, ETH, and ETC. In Section 2, we find some of the most popular market cap cryptocurrencies (BTC, LTC, XRP), and some others with lower market caps and higher returns (AE, USDT, ZRX).

Intersection 3 and *4* have approximately 200 cryptocurrencies, each with a high influence of K-means and Hist-DAWass clusters. These intersections are characterized by cryptocurrencies that belong to *Cluster 3* of K-means and *Cluster 3* of the Hist-DAWass technique 6(g). The main difference from the previous intersections is that *Cluster 3* of K-means corresponds, on average, with negative daily mean returns but moderate volatility 3(a); therefore, the average returns are also lower for this intersection.

Intersection 3 includes one of the highest market cap cryptocurrency (BCH), and others with high market capitalization (GNT, LSK, QTUM). In Intersection 4, the lower returns introduced by *Cluster 1* of K-means is compensated by the positive effect on the return by the *Cluster 2* of TADPole, with the centroids sited over zero mean returns for all quarters 8(c). There are no high returns cryptocurrencies at this intersection (DASH, SC, STRAT).

In conclusion, in *Intersection 5* and *6*, we have *Cluster 3* from K-means, *Cluster 1* from Hist-DAWass, *Cluster 3*, and *2* from TADPole, respectively. *Cluster 2* from Hist-DAWass was more volatile than *Cluster 3* and has the heavier tails. In *Intersection 5*, we find cryptocurrencies with an average-high risk and average returns (CMT, ETT, HST). Intersection 6 allocates some cryptocurrencies with high market caps but low returns (BCD, SBTC, GEO).

In the alluvial plot shown in Fig. 9, we show how the different clusters of the different algorithms are related. This makes it possible to appreciate both the main trends already mentioned, and those that are more subtle. For example, notably, the smallest clusters in K-means and Hist-DAWass (*Cluster 1* and *5*, respectively) share the most cryptocurrencies. In the subgroup of very volatile cryptocurrencies, we find that AMIS, B2X, ELTCOIN, FLLW, GOOD, ICE, ITT, LBTC, PFR, REX, RIPT, STAR, WAND, XIN, YOVI, and ZCG. Further, the group diverges and relates to the two main clusters of the TADPole without a clear pattern, which means that the temporal evolution is more conventional with a mean return on a quarterly basis, positive or negative, but not related to other forms of multidimensionality. Curiously, the smallest TADPole *Cluster 1* is not strongly related to any other cluster. This means that its peculiar time series evolution is not particularly related to the prototypes of the aggregated representations of the other clustering techniques, namely the return distributions and mean-standard deviation bi-variate or two-moments representations.

In conclusion, we notice that DEUR is the only cryptocurrency that was not pair-combined with any other cryptocurrencies along the three techniques with no activity at all in the market, during our analysis period.

With regard to the ARI values, we obtained extremely low agreement values. The highest value is 0.0123, which is very close to zero, which means no agreement. We obtain this value for the agreement between the K-means and Hist-DAWass results. For the rest of the intersections, we find even lower values. This means that there is no agreement between the different clustering results, which matches our aim of using clustering results that provide complementary views on the market.

Association tests

As explained in the Methods section, we rely on exact Fisher tests based on *Monte Carlo* simulations for the significance tests of the associations between the variables. The p-values of Fisher's test are depicted graphically in Fig. 10, with the results of the Fisher tests among the categorical variables in Table 1 and the clusters (including the intersections of the clustering results). P-values lower than 0.01 are represented in purple color addressing the more significant associations.

The association tests are aimed at enhancing the characterization of the clusters, by adding value to the prototyping descriptions that we explained in Result section. In the red box, we group the areas with the associations between clusters, and market categorical variables.

Association between market cap, volume and clusters

According to Table 6, the *Cluster 3* of K-means (the one with the more pronounced negative mean return prototype with -0.009) is associated with cryptocurrencies of high volume, but not those with the highest (Volume variable with *P80*, *P90*, and *P99* values) with standardized residuals of 4.36, 4.35, and 5.71, respectively. However, *Cluster 2* with the least pronounced negative mean returns (-0.002) is associated with the lower percentiles (*P70*) with a very high residual of 8.93.

While Volume was not considered in the clustering algorithm, the K-means results show an interesting association with volume; more precisely, lower volume or liquidity cryptocurrencies are strongly associated with the Cluster 2 profile. Curiously, some of the cryptocurrencies with the highest volume (BTC, EOS, ETC, ETH, LTC, and XRP) are also located in *Cluster 2*, even if the association is not statistically representative.

Regarding the Hist-DAWass and Volume variable, according to Table 6, *Clusters 1*, and *2*, whose prototypes had the lowest mean returns (-0.134, -0.503), are strongly associated with the lower Volume cryptocurrencies (standardized residuals 12.25 and 3.72). However, *Cluster 3*, whose prototype had the least pronounced negative average returns (-0.011) and the lowest volatility (0.108), is associated with the highest percentiles *P90*, *P99*, and *P100* with residuals of 5.09, 7.85, and 3.71, respectively. It is also possible to see weaker but relevant associations in one of the lowest cardinality clusters (*Cluster 4*) with a standardized residual of 3.36 for *P80*.

Consequently, we can conclude that Hist-DAWass provides a more accurate screening by Volume than K-means, as it separates the cryptocurrencies in the three groups more clearly.

Regarding the MKCap variable, in Table 7, the K-means association is not very strong. For example, we see the lowest and highest market cap percentiles (*P70, P100*) sharing the same *Cluster 2* with standardized residuals of 3.91 and 3.57, respectively.

However, in the association between the MKCap variable and Hist-DAWass, we observe a significant association between *Cluster 1*, and the lowest market cap percentiles *P70* with a value of 8.07. In contrast, *Cluster 3* is linked to high market cap cryptocurrencies (*P90* and *P99* percentiles).

Association between financial ratios and clusters

Regarding the associations with Beta, Table 8 shows a link (standardized residual of 17.79) between the *Cluster 1* of K-means and the *Extreme* Beta (ICE, ITT, PFR, and STAR), which is consistent with the cluster being the one with the most volatile cryptocurrencies.

Cluster 2 of K-means allocates cryptocurrencies with positive and moderate negative mean returns, and is strongly related to low volatility (*LowVol*) Betas (BTC, DCN, WAVES or WBTC*).

In conclusion, *Cluster 3* is associated (standardized residual of 5.79) with cryptocurrencies with high volatility (*HighVol*) (ADA, BCH or SALT).

The *beta* value acts as a proxy of the risk, and the association with the K-means results reveal that it discriminates three groups of different behaviors that could interest the investor depending on his/her risk-aversion profile, sufficiently.

However, we can confirm the higher screening capacity of the Hist-DAWass clustering with the help of Table 8. This technique separates with a high significance *NegBeta* in *Cluster 1*, *Cluster 2*, and *Cluster 3* (the high negative value of -15.24 of the standardized residual means that *NegBeta* cryptocurrencies are not significantly allocated in *Cluster 3*); *IndexLike* in *Cluster 3*; and *Extreme* beta values in *Clusters 4* and *5* with the highest standardized residuals (10.33 and 19.24). The association of *Cluster 3* and *Indexlike* Beta values can be explained by many of the components of the CCI30 index (BTC, BCH, DASH, ETC, ETH, LTC).

Regarding the Sharpe ratio variable, Table 9 shows that TADPole is capable of reflecting a strong association between *ERP* (excess return positive) and *cluster 2* (BTM, SC, DNT, LEND, and WINGS) with a residual of 10.6, and between *SRF* (small risk-free) class and *Cluster 3* (EOS, ETC, ETH, NEO, or ZEC) with a residual of 11.65.

The *Sharpe ratio* represents the excess return with respect to a risk-free asset or, in other words, the risk-reward for the investment on the asset (cryptoasset in our case). Interestingly, there was no association with the *GOOD* categorical label –a Sharpe ratio higher than 1.0– in the 2018 dataset. In the best case, there are weak associations with *ACC* -higher than 0.5, and lower than 1.0, which is a suboptimal category (see Table 1).

These cryptocurrencies are located in *Cluster 2* and are represented by a distribution with a positive mean return, as shown in Table 2.

In summary, the clustering results of the K-means and Hist-DAWass clusters are associated with the Market cap, Volume, and Beta variables, and the TADPole results are only associated with the *Sharpe ratio*.

Associations results for the intersection of clusters

As shown in Fig. 10, the cluster intersection (Combi variable) is significantly associated with most of the variables. The intersection of the cluster provides a complementary characterization of the cryptocurrencies as the intersections successfully combine the idiosyncrasy of each clustering algorithm. Tables 6, 7, 8, and 9 show the association between the different categorical variables and the higher cardinality intersections (first six rows in Table 5).

Regarding the Volume variable, *Intersection 3* with standardized residuals of 6.21 and 6.21, and *Intersection 4* with standardized residuals of 4.53 to 7.35 are associated with high-volume cryptocurrencies in percentiles *P90, P99* (ADT, BLOCK, CND). *Intersection 4* is also linked with *P80* percentile (FLIX, LDC, RVT). The highest percentile *P100* is associated with *Intersection 1* (EOS, ETC, ETH) with a standardized residual of 4.02. In conclusion, the lowest percentile *P70* is allocated in *Intersection 1* (again, it coincides with *P100*), *Intersection 2*, *Intersection 5*, and *Intersection 6*.

Regarding the MKCap variable, the low market cap cryptocurrencies *P70* are mostly allocated in *Intersection 5* (ANTI, BBT, XMG) and *Intersection 6* (BTA, CNT, NTRN) with standardized residuals of 4.88 and 4.87, respectively. The percentiles *P80, P90* are linked to *Intersection 3* (ADT, BTX, ION) and *P80* with *Intersection 4* (BAY, LEND, SKY).

There are no *Extreme* Beta cryptocurrencies in the highest cardinality intersections, as shown in Table 8. However, we can see an association between *HighVol* and *Intersection 3* (BCH, QTUM, XVG) and *Intersection 4* (ADA, HSR, STRAT) with residuals of 6.20 and 6.62, respectively. In contrast, *LowVol* is associated with *Intersection 2* (BTC, WAVES, WBTC*) with a residual of 5.62. In conclusion, the *NegBeta* values are strongly associated with *Intersection 5* (FRX, PPP, XPY) and *Intersection 6* (GLC, TIT, XHI) with standardized residuals of 7.87 and 10.52, respectively.

Regarding the Sharpe ratio, the acceptable cryptocurrencies for investment (*Acc*) are mostly allocated in *Intersection 2* (WAVES, XRP, ZEN) with standardized residuals of 4.13. The excess return positive (*ERP*) are linked to *Intersection 2* (AC, ZRC, ZRX) and *Intersection 4* (ADA, HSR, SC) with a standardized residual of 5.20 and 4.67, respectively.

Notably, the intersections are associated with all the considered categorical variables, except for the technological variables that we review below. Consequently, we can conclude that the intersections of the clustering results improve the characterization by means of the associations. The intersection inherits some of the associations of the different clustering results, despite the significance being lower.

Associations between financial and the technological variables

It is worth mentioning that while technological variables are not associated with any clustering results, they have a significant relationship with other independent financial variables, such as market cap (Tables 10, 11) and trading volume (Table 12), as shown in Fig. 10 (grey square in the upper left area). For example, in Table 10, we can see a relevant association between *Scrypt* (7.58 standardized residual), *SHA256* (3.58), and *X11* (6.65) encryption algorithms, as well as the lower percentile (P70) of the market cap. Encrypted algorithms *CryptoNight-V7, Ethash, Ouroboros* are also associated with the highest market cap percentile (P100).

Regarding the consensus algorithm, the ProofType variable in Table 11, the *PoS* (3.74), *PoW/PoS* (9.09), and *PoW* (4.54) are associated with the lowest quantile (*P70*) of the Market cap variable.

Associations with the age of the cryptocurrencies

In Table 13, we can see the associations of clusters with the age or maturity of the cryptocurrencies. In K-means, the only association is that of *Cluster 2*, which is characterized by low volatility (0.130) and slightly negative average returns (-0.002) in Table 2, with the youngest cryptocurrencies (*D4*) (standardized residual of 4.74).

However, Hist-DAWass showed more interesting associations. For example, *Cluster 1* is associated with cryptocurrencies in the deciles *D5, D6, and D7*, with standardized residuals of 11.88, 7.91, and 4.70, respectively. According to Table 3, this cluster is characterized by a distribution with moderate skewness (0.82) and high kurtosis (13.43). Similarly, *Cluster 2* is linked to cryptocurrencies in decile *D6* with a residual of 4.97.

In contrast, the oldest cryptocurrencies are prominently associated with *Cluster 3* with a standardized residual of 15.08. Interestingly, *Cluster 3* is also a cluster with higher market cap cryptocurrencies (`BCH, BTC, DASH, EOS, ETC, ETH, IOT,LINK, LTC, NEO, WAVES, XLM, XMR, XRP, ZEC, ZRX`); hence, they are the most popular cryptocurrencies for investors, but also those with more stable behavior, according to the cluster prototype.

Regarding the intersections of the clustering results, Table 13 shows that the oldest (*D10*) cryptocurrencies are allocated in *Intersection 3* and *4* with high standardized residuals (7.20, 8.37, respectively). Middle-age cryptocurrencies (*D5*, *D6*) are linked to *Intersection 5* and *Intersection 6*, whereas the younger ones (*D4*) are significantly allocated in *Intersection 6* (`EBC, ICOB, PULSE`) and *Intersection 4*.

We further confirm that Hist-DAWass offers stronger associations than K-means, and that the main intersections provide even more associations. Consequently, clustering intersections are very good for characterizing cryptocurrencies owing to their higher granularity, and their tendency to display more significant associations that are better distributed.

Association with heavy-tail behavior

We counted 461 out of 1723 cryptocurrencies with heavy-tail behavior in 2018. According to the tests, heavy-tail behavior is mainly associated with *Cluster 2* in K-means and *Cluster 2* in Hist-DAWass (standardized Pearson's residuals of 7.02 and 17.08, respectively), but the association is also high for *Cluster 1* of K-means and Hist-DAWass, as shown in Table 14. We have already mentioned that *Cluster 1* in K-means allocates the highest volatile cryptocurrencies, and in this case, they correspond to heavy-tail cryptocurrencies as well. As stated earlier, *Cluster 2* of Hist-DAWass allocates more negative-return cryptocurrencies. Therefore, we can conclude that heavy-tails are stronger for the left tail. In conclusion, there is no clear link between the TADPole technique and the heavy-tail distributions (very low values for the standardized Pearson's residuals); therefore, we conclude that there is no relationship between shape-base clustering and distribution characterization.

Analysis of the extended time frame

According to Fig. 2, we can see that the shapes of the clusters in the extended period is quite similar to those found for 2018 (Fig. 1). We notice that the xy-axis has different ranges for volatility and mean return, owing to the variation in the data sets, which are primarily different in the number of cryptocurrencies. However, a comparison of both time-frame periods shows that the shapes are mostly the same.

If we analyze the ARI index of the K-means results in 2018 and 2018-2019, we find a high agreement or similarity (0.349). The agreement for Hist-DAWass is similar (0.304), whereas for TADPole is null (-0.0021). The TADPole result can be explained as it uses the 'raw' data and not a summary, which makes it more difficult to find similar trajectories through longer time periods. Additionally, TADPole clustering appears to be more sensitive to changes in the objects to be clustered, than the classic K-means and Hist-DAWass.

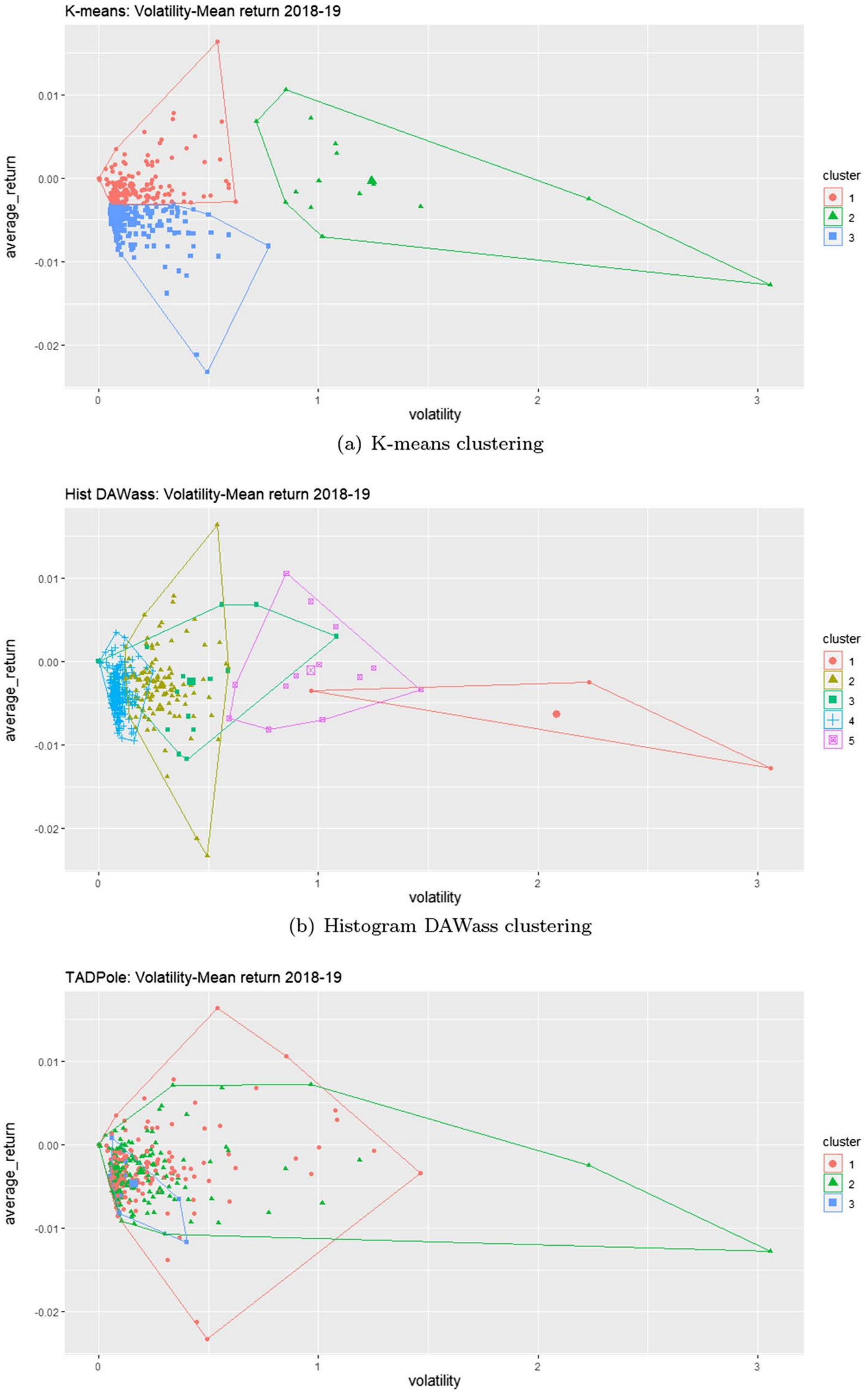
Regarding the application of the association tests for the extended time frame, the results in Fig. 11 are quite similar to those in 2018 shown in Fig. 10 with some exceptions: there is no association between TADPole and ClassSharpeR variable. Rather, we find a significant association between K-means and Hist-DAWass with ClassSharpeR. We further confirm the association between technological variables and Volume and MKCap.

We can conclude that there is a persistence on the structures detected by K-means and Hist-DAWass confirmed on the extended period but not for TADPole clustering. However shape-based clustering as TADPole is helpful in enhancing the description of the market for the chosen time frame.

Discussion

In this section, we summarize the main results obtained from the clustering and association tests.

- We confirm the existence of a structure in the market, that allows us to segment the cryptocurrencies into clusters. Interestingly, the optimum number of clusters remains low, independent of the representation considered, which indicates a high degree of homogeneity despite the high number of cryptocurrencies. This result is consistent with others that find evidence of different behaviors among cryptocurrencies (Song et al. 2019; Sigaki et al. 2019; Stosic et al. 2018) .
- The Hist-DAWass clustering offers a more subtle discrimination of the cryptocurrencies that are offered by the K-means on the mean and standard deviation,

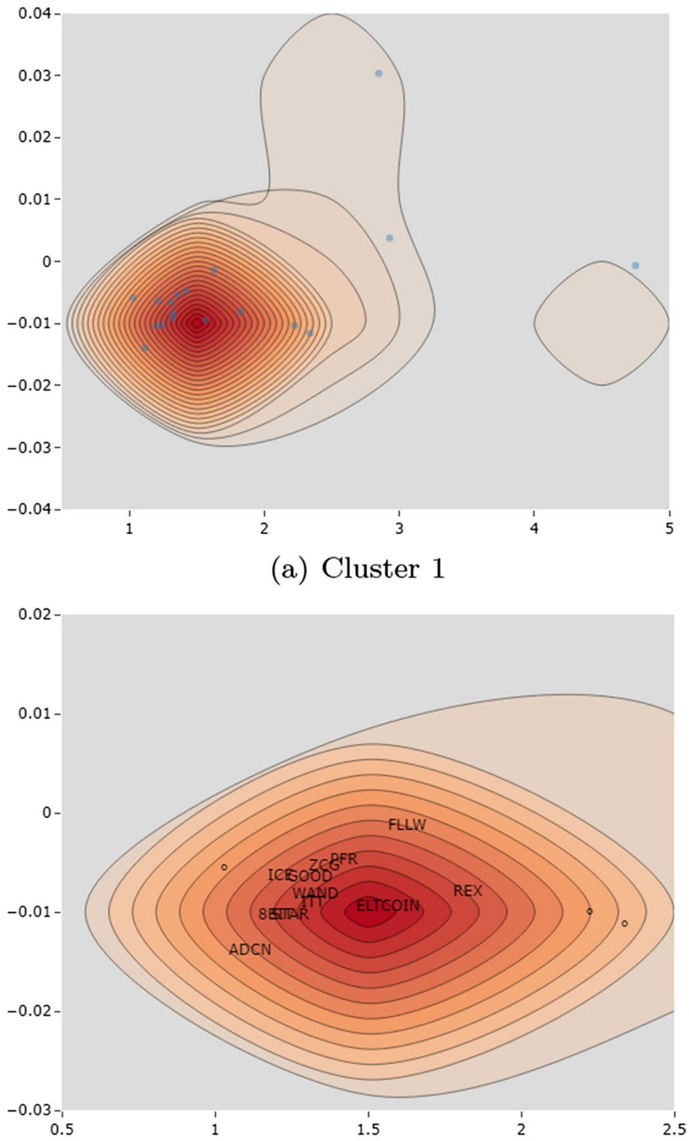


(a) K-means clustering

(b) Histogram DAWass clustering

(c) TADPole clustering

Fig. 2 Volatility-Average return plane in ordinary values for the different clustering techniques for 440 cryptocurrencies in 2018-19 time frame

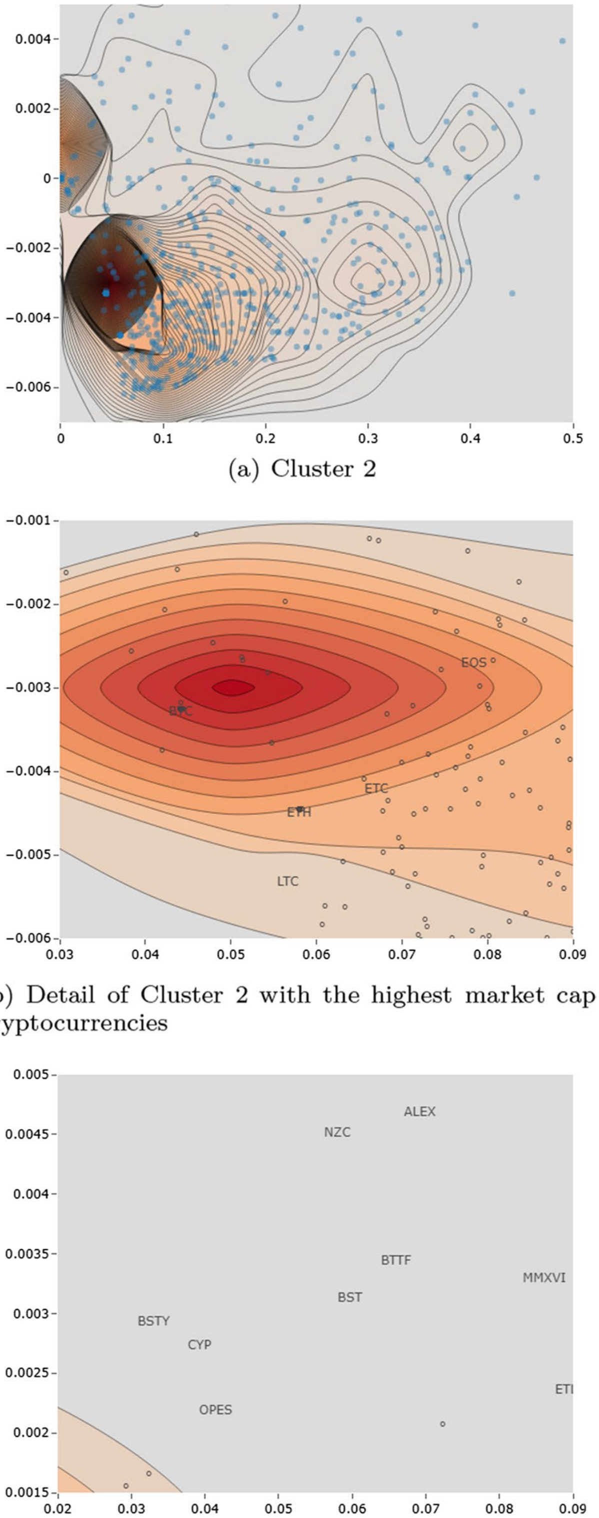


(a) Cluster 1

(b) Detail of Cluster 1

Fig. 3 Cluster 1 represented in the bi-dimensional space $(\bar{r}, \sigma)$ by a 2D density contour plot

or by the TADPole on the time series. From our point of view, the log distribution offers a nuanced summary that considers not only the mean and the standard deviation of the distribution, but also other aspects such as symmetry, kurtosis, or tail behavior. Obviously, the distribution aggregation does not consider the evolution of through time, but this problem can be partially overcome by considering a sequence of distributions that aggregate data, say, quarterly, instead of yearly. Therefore, we consider that Hist-DAWass is a suitable and promising profiling tool for investors, and believe that it could be used in financial markets in general.

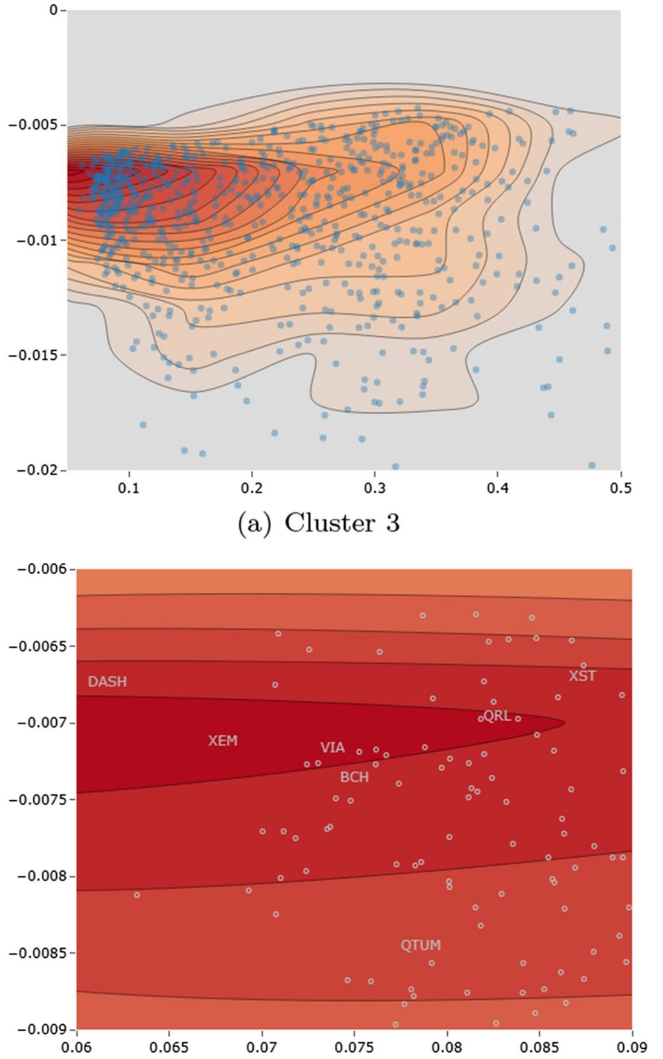


(a) Cluster 2

(b) Detail of Cluster 2 with the highest market cap cryptocurrencies

(c) Detail of cluster 2 for the high returns and low volatility cryptocurrencies

Fig. 4 Cluster 2 represented in the bi-dimensional space $(\bar{r}, \sigma)$ by a 2D density contour plot



(a) Cluster 3

(b) Detail of Cluster 3 with the highest market cap cryptocurrencies)

Fig. 5 Cluster 3 represented in the bi-dimensional space $(\bar{r}, \sigma)$ by a 2D density contour plot

Table 3 Descriptive statistics for the prototypes of the Hist-DAWass clustering

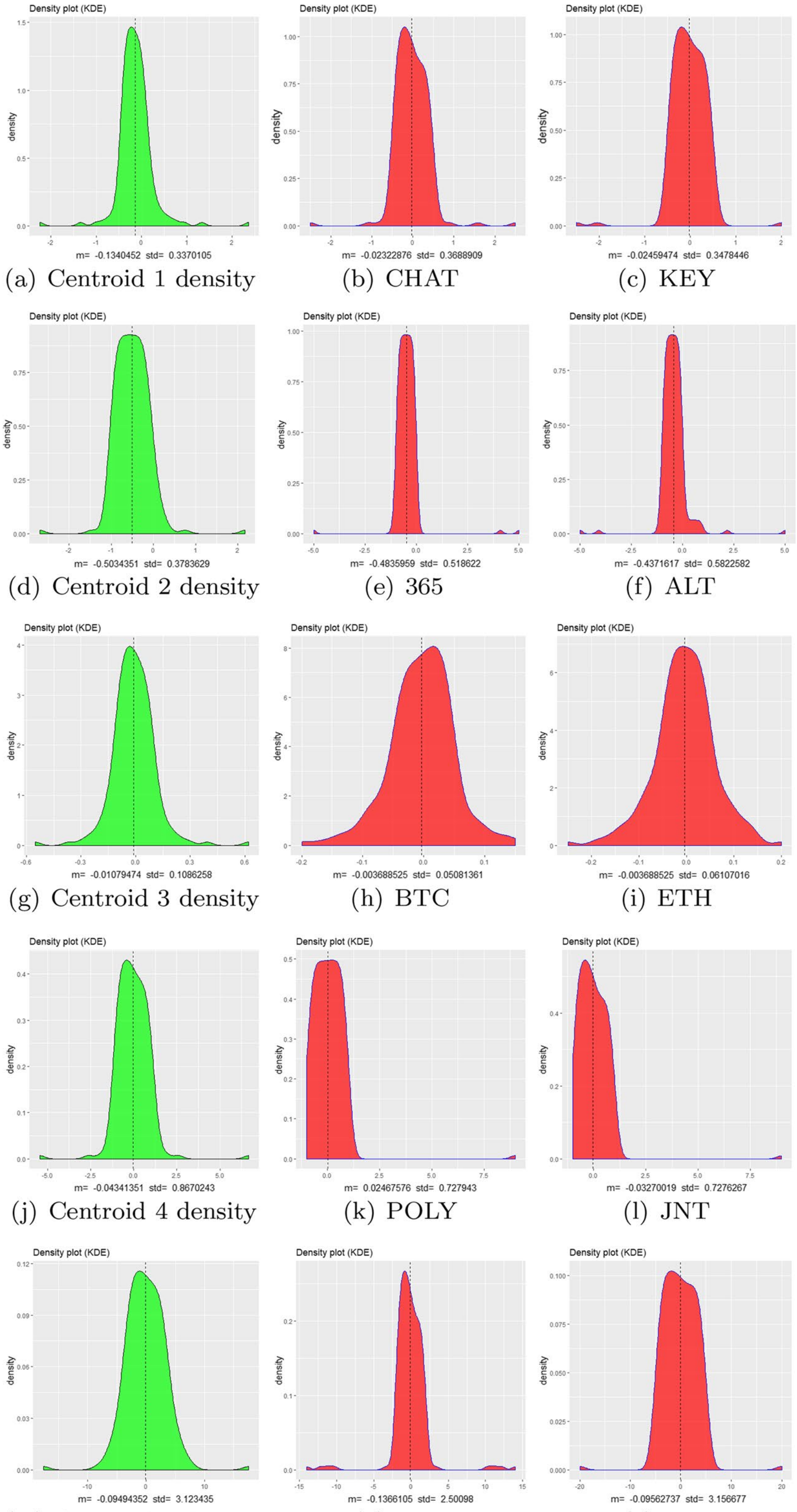
	Mean	Std. Dev.	Coef.Var.	Skew.	Kurt.	Med.	Min.	Max.	Var.Wass.
Clus. 1	− 0.13	0.34	− 2.51	**0.82**	**13.43**	− 0.16	− 2.24	2.36	0.025
Clus. 2	− **0.50**	0.38	− **0.75**	0.56	9.33	− 0.51	− 2.69	2.18	0.079
Clus. 3	− **0.01**	0.11	− 10.06	0.28	7.10	− 0.01	− 0.55	0.62	0.005
Clus. 4	− 0.04	0.87	− 19.97	0.54	11.95	− 0.08	− 5.44	6.67	0.128
Clus. 5	− 0.09	**3.12**	− **32.90**	0.05	5.66	− 0.17	− 17.56	17.56	1.116

Bold figures show extreme performance values

Table 4 Variability of TADPole clusters with the mean distance (Mean Dist.) to the centroid, standard deviation (Std. Dev.) and coefficient of variation (Coef. Var.)

Cluster	Mean Dist.	Std. Dev.	Coef. Var.
1	4.31	3.04	0.71
2	4.60	3.29	0.72
3	4.85	3.53	0.73

- Our results show that the K-means partition is strongly associated with Beta values. This is not surprising, as beta is computed using the mean return and volatility, which are the variables considered for clustering.
- Both the K-means and Hist-DAWass partitions are associated with the market capitalization and the volume. The relation between price and volume for bitcoin has been documented in the literature (Balcilar et al. 2017; Sahoo et al. 2019; Szetela et al. 2021) but our results hints that such a relationship could be extended to other cryptocurrencies in the market. Particularly, the association seems to be stronger in the case of Hist-DAWass, which means that the shape of the distribution plays a role in the association with the volume and market cap.
- The K-means and Hist-DAWass clusters also show an interesting association with the age of the cryptocurrencies. The results indicate that younger and older cryptocurrencies have particular and different return and volatility behaviors as detected by the clustering techniques. This *maturity* effect was previously observed in the Bitcoin financial behavior (Drozdz et al. 2018). However, our results point out that it happens to other cryptocurrencies. Pele et al. (2020) shows that the behavior of cryptocurrencies evolves and follows a synchronic evolution. Our results show that younger cryptocurrencies tend to have higher kurtosis and skewness, while the oldest cryptocurrencies are more stable. Interestingly, the cluster with the most extreme behavior did not show a significant association with a particular age category.
- TADPole clustering of the time-series representation produced a small number of clusters and was associated with the Sharpe ratio. However, in the extended time frame, the clustering results do not show a low similarity with the results in 2018, and further show no associations. Therefore, the TADPole or shape-based clustering seem to offer more unstable results, probably because they use disaggregated data, and the results are more sensitive to small changes.
- The intersection of clusters seemingly inherit the association that we observed separately for each one of the methods. This is confirmed for the Age, MkCap variables, and different financial ratios. Consequently, clustering intersections characterize the main trends of the cryptocurrency market in a comprehensive manner, providing a manageable number of clusters with a multi-faceted characterization, and display significant associations with other relevant variables not considered in the clustering process.
- We confirm the persistence of many of the associations in a longer period which seems to confirm that these associations are not conjectural, but prolonged.
- The proposal of techniques other than those proposed in this research may yield better results measured in terms of a higher level of significance in the associations, which is what determines the quality of the detected clusters that could ultimately be a part of



(a) Centroid 1 density (b) CHAT (c) KEY

(d) Centroid 2 density (e) 365 (f) ALT

(g) Centroid 3 density (h) BTC (i) ETH

(j) Centroid 4 density (k) POLY (l) JNT

(m) Centroid 5 density (n) B2X (o) ITT

Fig. 6 Density plot for prototypes (first column), and some representative cryptocurrencies of each cluster in terms of market capitalization (2nd and 3rd columns)

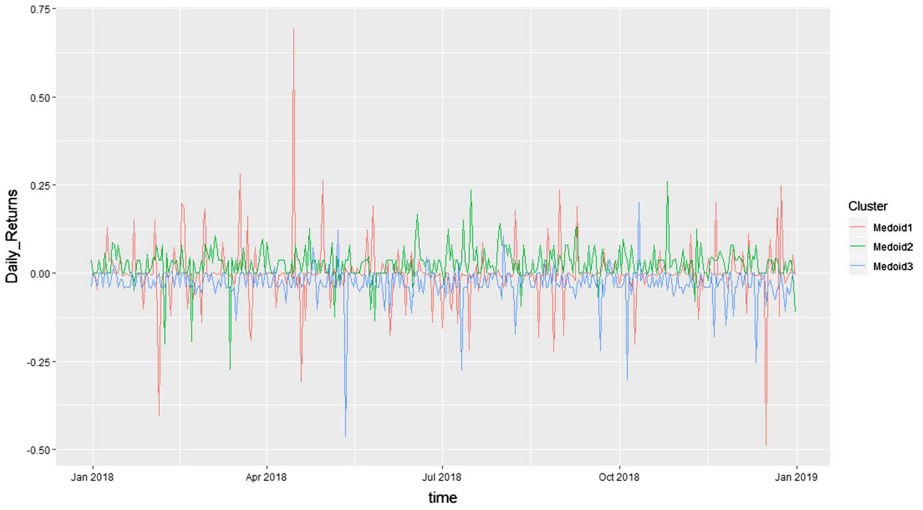


Fig. 7 Medoids of the clustering results of the TADPole algorithm (time series of daily returns)

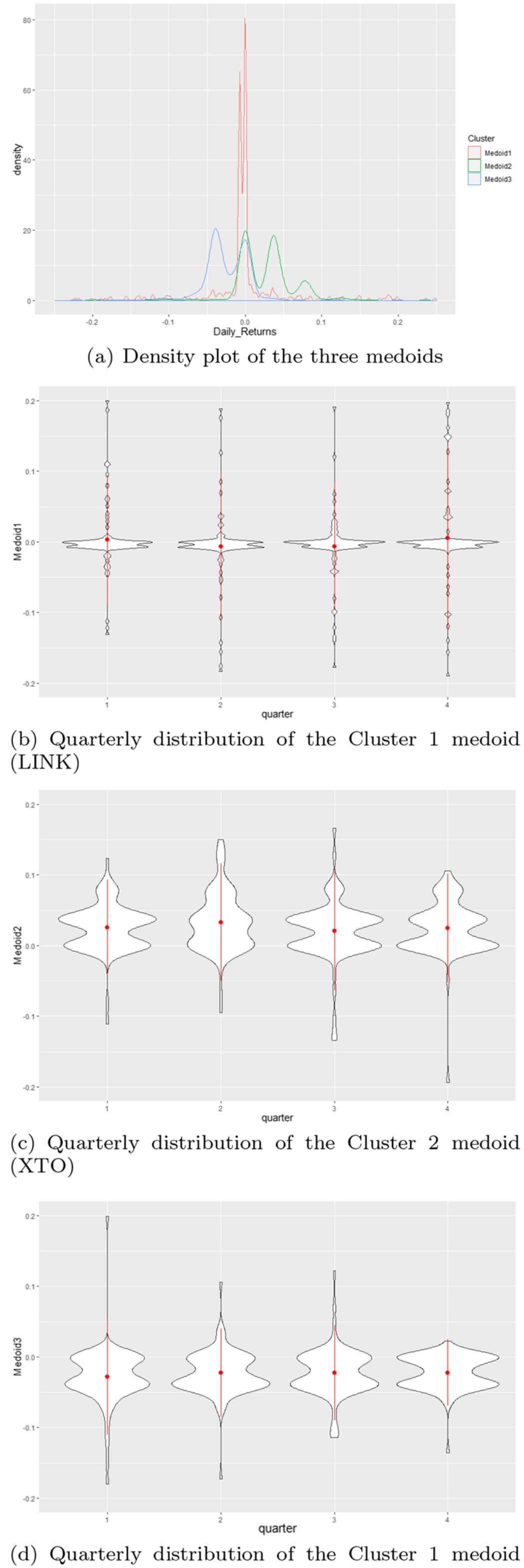
further investigations. In any case, the proposed methodology is fully open to test other techniques but always ensures the descriptive capacity of the prototypes, which should be considered if other potential techniques are proposed.

Use of the methodology results

The methodology proposed can be used as a way to map the main trends of the market by integrating different partitional clustering techniques. A similar example in the literature is the work by Soleymani and Vasighi, which groups stocks using enhanced K-means and computes value-at-risk measures to find the most and least riskiest groups of stocks.

In our methodology, the interested user, for example, an investor, could use the clusters to browse the cryptocurrency market and select cryptocurrencies or groups that are interesting or differ from their behavior. The search depends on its aims, for example, looking for diversification in a portfolio with traditional stocks, building a portfolio of cryptocurrencies that behave differently, etc. For such purposes, they can use one clustering result, or a combination of some of them. A graphical representation and summary of the prototypes should inform the search. Furthermore, the association of some of the clusters with some financial ratios can also guide the selection, for example, looking for clusters associated with interesting beta values or Sharpe ratios. The association with other descriptions, such as age, market cap, or volume, can also refine the search. After the search, a group of stocks is available, and the analyst should look at their behavior specifically, for example, the distribution of returns, the value of the financial ratios, the temporal evolution, etc.

We consider, for example, a trader with a high risk aversion, but that unsurprisingly would like to obtain high returns. Based on such investment criteria and aligned with the characteristics of the different centroids, the most suitable selection should be the cryptocurrencies in *Intersection 2* in Table 5. This intersection comprises *Cluster 2* of



(a) Density plot of the three medoids

(b) Quarterly distribution of the Cluster 1 medoid (LINK)

(c) Quarterly distribution of the Cluster 2 medoid (XTO)

(d) Quarterly distribution of the Cluster 1 medoid (ZNE)

Fig. 8 Density plot in **a** and quaterly distribution of the TADPole medoids represented by violin graphs in **b**–**d** with mean and standard deviation

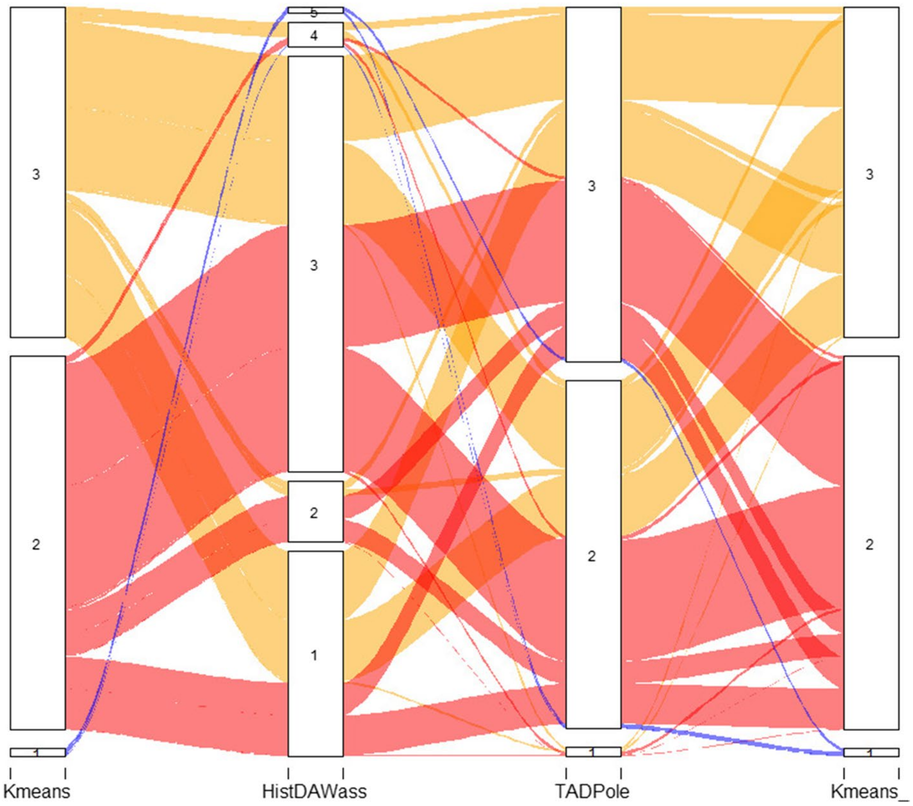


Fig. 9 Alluvial plot showing the 'flows' of cryptocurrencies through the three clustering algorithms

K-means, which includes higher mean returns cryptocurrencies, *Cluster 3* of HistDAWass, which has shorter tails and a higher median, and *Cluster 2* of TADpole with a consistent average return above zero.

Further, the trader can look for the higher market cap cryptocurrencies (*P99, P100*) at the intersection, and those with a Beta, *IndexLike* value, and a Sharpe ratio equal to or higher than *ERP*. The outcome of the search was cryptocurrencies DCR, LTEC, XMR, and ZEN.

Another trend in the literature shows the use of clustering as a preceding stage, before applying well-known portfolio optimization models. For example, after a clustering process, Gubu et al. (2020) computes the Sharpe ratio of each stock and selects the stocks with the highest value from each cluster. Similarly, Nguyen Cong et al. (2014) cluster the stocks based on their associated return rate and risk after a multi-objective optimization, to find a global solution Pareto optimal by selecting the stocks of each cluster (previously less interesting clusters from a financial point of view are discarded). Based on this premise, we can use our methodology to reduce the cryptocurrency market universe to those cryptocurrencies allocated on what we consider the more interesting clusters or intersections. We can further apply the portfolio optimization model that we consider appropriate Burggraf (2019) on the selected clusters or intersections.

For example, we consider a portfolio manager interested in the features of the aforementioned *Intersection 2*. Particularly, besides the centroid properties, they also consider

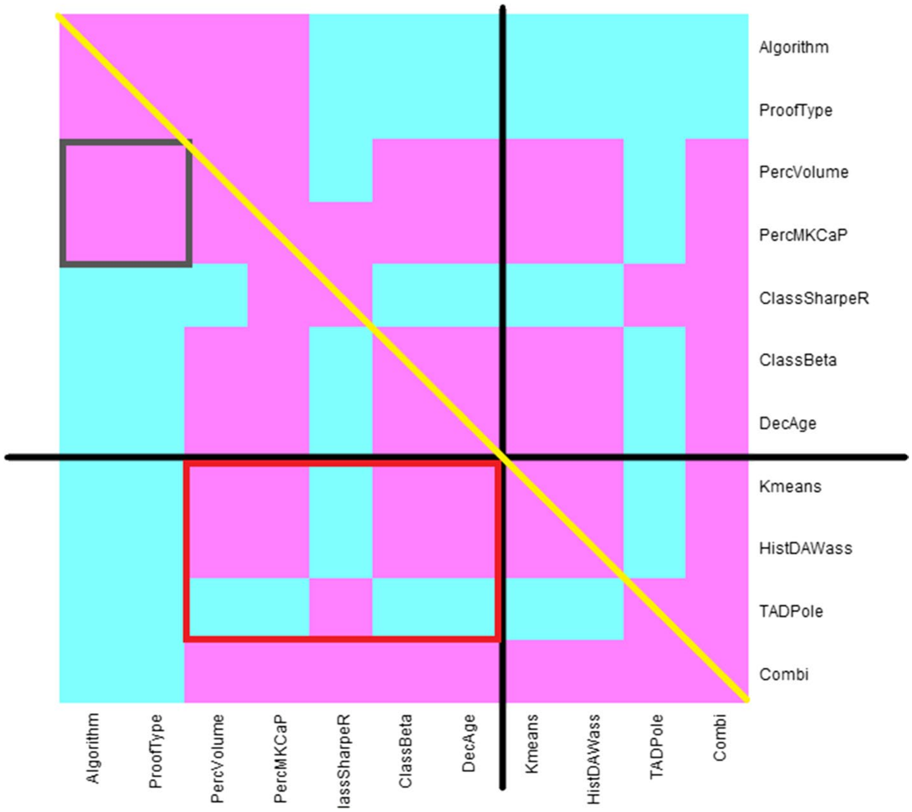


Fig. 10 Matrix-type representation of the association tests in 2018 using the Fisher's exact test. Binary colored, where pink color indicates significant association at p-values lower than 0.01. The red box for cluster-categorical variables and the gray box focused on the particular associations with the technological variables. The yellow line represents the trivial association between a variable and its own

the significant associations of *Intersection 2* with the other descriptive values. For example, there is a significant association between the *Intersection 2* and the *LowVol* value Beta variable, which represents low volatility. Similarly, *Intersection 2* has a significant association with the *ERP* value, which represents a moderate excess positive return of the Sharpe ratio. Furthermore, *Intersection 2* is associated with cryptocurrencies that are not very popular, according to the significant associations with the value *P70* of variable Volume, which represents the lowest liquidity according to our categorization, and with the value *D4* of the variable Age, which represents the younger cryptocurrencies in our categorization. With this information, the portfolio manager could consider only the cryptocurrencies of *Intersection 2* as a reduced universe of cryptocurrencies to apply a portfolio optimization model.

Conclusions

In this study, we analyzed the cryptocurrency market in 2018, that is, all cryptocurrencies traded in 2018, using a novel method that involved the integration of three different clustering algorithms. Each method uses a meaningful representation considering different aggregation or granularity levels of the daily returns, from the yearly average return and volatility, the yearly distribution of returns, and finally, the observed time series of

Table 5 Intersection of clusters across the different clustering algorithms, each column represent the cluster number. Intersections (Combi variable) are ordered by backwards cardinality

Intersection (Combi)	Kmeans	Hist-DAWass	TADPole	N
1	2	3	3	295
2	2	3	2	294
3	3	3	3	208
4	3	3	2	196
5	3	1	3	166
6	3	1	2	148
7	2	1	2	97
8	2	1	3	78
9	2	2	3	57
10	2	2	2	54
11	3	2	3	20
12	3	4	2	18
13	3	4	3	18
14	3	2	2	15
15	2	4	2	10
16	2	4	3	8
17	1	5	2	8
18	1	5	3	8
19	2	3	1	7
20	3	3	1	7
21	3	1	1	5
22	1	4	2	3
23	2	1	1	2
24	2	2	1	1

daily returns. Given the meaningful data representation, the cluster prototypes are useful to obtain an informative summary, and a visual representation of the main trends of the entire market.

Furthermore, we enhanced our profiling of the cryptocurrency market with association tests to validate the potential relationship between the clustering results and other descriptive features of cryptocurrencies (technological attributes, financial ratios, market cap, volume and age). These tests make it possible to ascertain whether some features are related to a particular financial performance detected by the clustering algorithms. Additionally, we found a significant association between technological attributes and the behavior of the market. These associations discovered open venues for future research to confirm them and determine their scope more precisely.

Our analysis confirmed that there is an underlying structure of the data, which also persisted when considering a longer time period. Each of the clustering algorithms helped to reveal different aspects of the cryptocurrency market. Furthermore, we show that the combination of the different clustering results proved valid for detecting the main trends in the cryptocurrency market. The cluster partitions along with the prototypes, and the cluster description provide a manageable summary in the financial terms of the entire market. It is also possible to obtain more sophisticated profiles by examining the intersection of the clusters. Furthermore, the analyst or

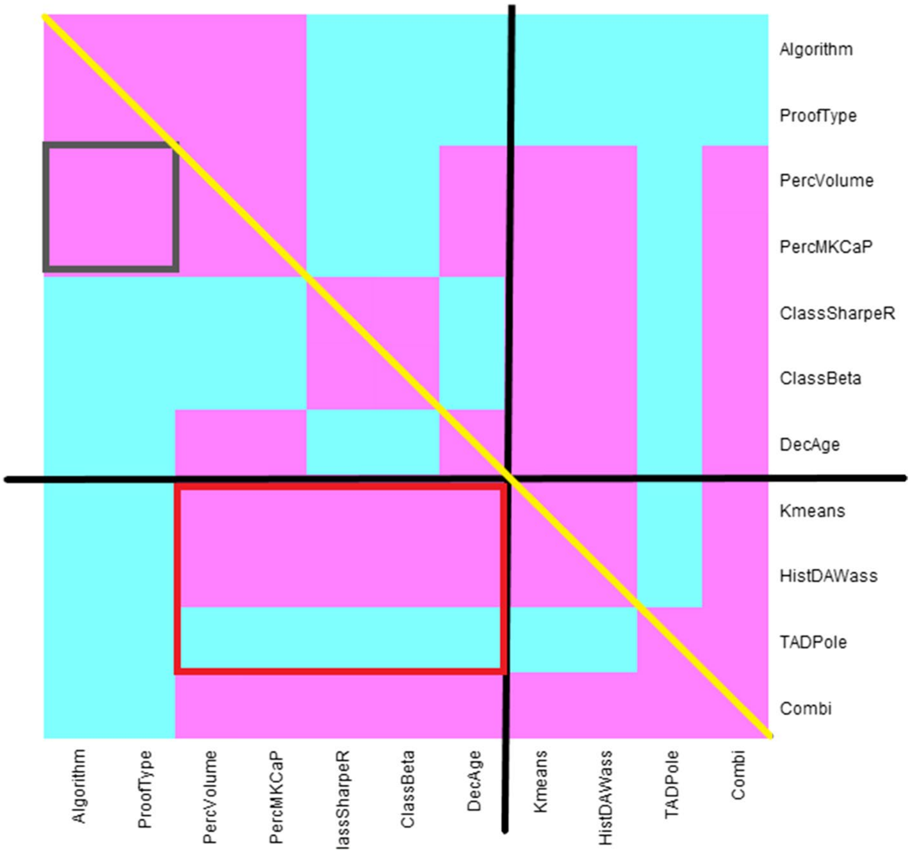


Fig. 11 Matrix-type representation of the association tests in 2018–19 using the Fisher's exact test. Binary colored, where pink color indicates significant association at *p*-values lower than 0.01. The red box for cluster-categorical variables and the gray box focused on the particular associations with the technological variables. The yellow line marks the trivial maximum association for the same variables

Table 6 *Volume* - Standardized Person's residuals

Technique	Cluster/ Intersection	Volume				
		P70	P80	P90	P99	P100
K-means	1	− 0.13	2.43	− 1.05	− 0.97	− 0.43
	2	**8.93**	− 4.73	− 4.20	− 5.57	2.61
	3	− 8.90	**4.36**	**4.35**	**5.71**	− 2.54
Hist-DAWass	1	**12.25**	− 4.33	− 4.82	− 7.26	− 3.67
	2	**3.72**	− 1.71	− 1.78	− 1.66	− 0.74
	3	− 12.01	3.09	**5.09**	**7.85**	**3.71**
	4	− 2.02	**3.36**	0.44	− 0.95	0.22
	5	− 0.48	2.78	− 0.97	− 0.90	− 0.40
Combi	1	**3.66**	− 2.20	− 2.62	− 2.52	**4.02**
	2	**5.98**	− 2.52	− 2.51	− 3.49	− 0.41
	3	− 10.25	2.57	**6.21**	**6.21**	− 0.26
	4	− 12.42	**6.39**	**4.53**	**7.35**	− 0.21
	5	**7.16**	− 3.21	− 2.24	− 3.95	− 2.14
	6	**7.30**	− 1.25	− 3.96	− 4.39	− 1.97

Bold figures show the more significant positive residuals

Table 7 *Market cap* - Standardized Person's residuals

Technique	Cluster/ Intersection	Market cap (MKCap)				
		P70	P80	P90	P99	P100
K-means	1	1.15	0.28	− 0.96	− 0.91	− 0.37
	2	**3.91**	− 5.90	− 1.77	0.20	**3.57**
	3	− 4.08	**5.85**	1.92	− 0.06	− 3.51
Hist-DAWass	1	**8.07**	− 1.70	− 4.98	− 4.64	− 2.21
	2	2.36	− 0.87	− 1.63	− 0.81	− 0.63
	3	− 8.37	1.88	**4.89**	**4.84**	2.59
	4	− 0.44	− 0.24	1.37	− 0.16	− 0.78
	5	0.94	0.44	− 0.89	− 0.84	− 0.34
Combi	1	2.27	− 3.02	− 2.45	0.86	2.97
	2	2.17	− 3.34	0.40	− 0.90	1.31
	3	− 6.69	**3.98**	**4.80**	1.89	− 1.54
	4	− 6.33	**3.63**	2.72	3.31	− 0.33
	5	**4.88**	− 0.99	− 3.38	− 2.22	− 1.72
	6	**4.87**	0.21	− 2.94	− 3.96	− 1.58

Bold figures show the more significant positive residuals

Table 8 *Beta* - Standardized Person's residuals

Technique	Cluster/ Intersection	Beta					
		NegBeta	CashLike	LowVol	Indexlike	HighVol	Extreme
K-means	1	3.05	− 0.17	− 2.92	− 0.97	− 1.45	**17.79**
	2	− 2.87	0.57	**6.70**	− 0.18	− 5.58	− 1.51
	3	2.41	− 0.54	− 6.26	0.32	**5.79**	− 1.13
Hist-DAWass	1	**12.09**	1.57	− 0.32	− 5.84	− 3.37	− 1.90
	2	**3.97**	− 0.28	− 2.24	− 0.94	0.74	− 0.38
	3	− 15.24	− 1.29	3.10	**6.66**	2.83	− 4.29
	4	**6.93**	− 0.36	− 5.47	− 2.05	1.23	**10.33**
	5	2.02	− 0.15	− 2.70	− 0.89	− 1.34	**19.24**
Combi	1	− 3.71	− 0.90	4.50	− 0.19	− 2.92	–
	2	− 4.07	0.49	**5.62**	0.57	− 4.82	–
	3	− 3.47	− 0.79	− 5.74	3.28	**6.20**	–
	4	− 3.42	0.76	− 5.41	1.99	**6.62**	–
	5	**7.87**	− 0.65	1.73	− 3.61	− 3.53	–
	6	**10.52**	1.29	− 1.69	− 3.16	− 1.61	–

Bold figures show the more significant positive residuals

investor can look for specific cryptocurrencies and determine the clusters to which they belong, and how far they are from the prototypes, according to the dissimilarity measure used for each clustering method.

We believe that the proposed methodology provides a consistent and descriptive tool supported by both well-known, and modern clustering techniques that may be useful for investors who need to understand the cryptocurrency market, as it reduces the dimensionality of the data set and identifies the main trends in a descriptive manner.

Particularly, the associations between financial ratios and clusters could play an important role in enhancing the performance of the optimization algorithms for

Table 9 *Sharpe ratio* - Standardized Person's residuals

Technique	Cluster/Intersection	Sharpe ratio		
		SRF	ERP	Acc
TADPole	1	− 1.43	1.52	− 0.44
	2	− 11.32	**10.60**	**3.69**
	3	**11.65**	− 10.95	− 3.58
Combi	1	**5.83**	− 5.49	− 1.74
	2	− 6.02	**5.20**	**4.13**
	3	**3.92**	− 3.63	− 1.53
	4	− 4.67	**4.67**	0.11
	5	**3.93**	− 3.68	− 1.24
	6	− 3.00	3.04	− 0.15

Bold figures show the more significant positive residuals

Table 10 Relevant associations between *Encrypted algorithm* - *Market cap* using the standardized Person's residuals

Encrypted algorithm (Algorithm)	Market cap (MKCap)				
	P70	P80	P90	P99	P100
Counterparty	− 1.92	4.00	− 0.51	− 0.49	− 0.19
CryptoNight−V7	− 1.92	− 0.50	1.69	− 0.49	**5.16**
Ethash	− 0.99	− 0.11	− 1.15	0.98	**4.37**
Leased POS	− 1.36	− 0.35	− 0.36	− 0.34	7.42
Ouroboros	− 1.36	− 0.35	− 0.36	− 0.34	**7.42**
Scrypt	**7.58**	− 2.54	− 3.05	− 5.06	− 2.16
SHA256	**3.58**	− 1.64	− 2.14	− 1.52	− 0.30
X11	**6.65**	− 2.34	− 3.68	− 3.72	− 0.87

Bold figures show the more significant positive residuals

Table 11 Relevant associations between *Consensus algorithm* - *Market cap* using the standardized Person's residuals

Consensus algorithm (ProofType)	Market cap (MKCap)				
	P70	P80	P90	P99	P100
LPoS	−1.36	−0.35	−0.36	−0.34	7.42
PoI	−1.36	−0.35	−0.36	−0.34	7.42
PoS	**3.74**	−1.77	−0.76	−2.82	−0.88
PoW	**4.54**	−2.39	−1.69	−3.12	0.63
PoW/PoS	**9.09**	−3.38	−4.69	−4.72	−2.43

Bold figures show the more significant positive residuals

asset selection, and the diversification of portfolios (Liu 2019; Platanakis et al. 2018; Brauneis and Mestel 2019) or improving the forecasting performance of predictive models (Mallikarjuna and Rao 2019) to tackle the difficulty of investing in a new and unknown market. It is also promising to analyze the connections between the

Table 12 Relevant associations between *Consensus algorithm* - *Volume* using the standardized Person's residuals

Consensus algorithm (ProofType)	Volume				
	P70	P80	P90	P99	P100
LFT	−1.20	−0.38	−0.40	−0.37	6.23
PoS	**3.66**	−1.65	−1.41	−1.74	−1.29
PoW/PoS	**7.34**	−1.51	−3.80	−4.65	−1.84

Bold figures show the more significant positive residuals

Table 13 *Age* - Standardized Person's residuals

Technique	Cluster/ Intersection	Age						
		D4	D5	D6	D7	D8	D9	D10
K-means	1	−1.14	0.32	0.76	2.48	0.63	1.18	−2.14
	2	**4.74**	−0.34	−1.75	0.91	1.33	−1.14	−2.79
	3	−4.56	0.29	1.63	−1.29	−1.43	0.96	3.11
Hist-DAWass	1	3.27	**11.88**	**7.91**	**4.70**	−0.51	−3.35	−13.70
	2	2.41	2.85	**4.97**	−1.15	−1.32	−1.78	−3.64
	3	−3.99	−11.80	−8.80	−4.65	0.38	2.80	**15.08**
	4	1.08	−1.35	−0.83	0.83	1.06	2.05	−1.95
	5	−1.06	0.49	0.94	1.08	0.80	1.43	−1.98
Combi	1	1.70	−2.76	−1.40	0.19	0.24	0.36	0.51
	2	**4.71**	−1.10	−1.73	−0.15	1.93	−1.22	−2.13
	3	−5.47	−4.02	−3.05	−1.46	−1.82	3.07	**7.20**
	4	−5.61	−4.24	−2.99	−2.84	−0.54	1.55	**8.37**
	5	1.35	**6.34**	**7.38**	2.87	1.16	−2.07	−8.53
	6	**3.87**	**8.38**	**3.62**	2.16	−1.15	−2.38	−7.92

Bold figures show more significant positive residuals

Table 14 Heavy-tail cryptocurrencies, Standardized Person's residuals for the association between the heavier tail distributions and clusters

Technique	Cluster	Heavy-tail
K-means	1	**3.60**
	2	**7.02**
	3	− 7.78
Hist-DAWass	1	0.52
	2	**17.08**
	3	− 11.97
	4	3.27
	5	3.24
TADPole	1	− 0.91
	2	− 0.28
	3	0.48

Bold figures show more significant positive residuals

Table 15 Summary of the main R libraries applied on the different parts of the investigation

DOMAIN	R package	Functions
Data wrangling	*reshape2*	Table convertion between Wide and Long
	data.table	Flexible and faster handling of big tables
	xts	Financial time-series operation
	rjson	For Json file convertion to R objects
	dplyr	Manipulation of operations with R commands, mainly %>% operator
	magick	Figure format conversion from png to eps
Quality	*poweRlaw*	An implementation of maximum likelihood estimator of heavy tail distributions
	mclust	Adjusted Rand Index computation
	clusterend	Hopkins index computation
Clustering	*clue*	Ensemble of k-means outcomes
	Hist-DAWass	Histogram clustering
	dtwclust	TADPole clustering
Graphs	*ggplot2*	General graph package
	FactoMineR	Scatter-plots of clustering
	factoextra	Scatter-plots of clustering
	plotly	Joint Probability Density (K-mean clusters)
	magrittr	Joint Probability Density
	RSelenium	Joint Probability Density
	ggridges	Density plots of TADPole medoids
Finance	*PerformanceAnalytics*	Sharpe ratio
Association tests	*vcd*	Pearson's residuals representation

technological implementation of the blockchain networks, and the formation of a given cryptocurrency's prices that we have only noted.

In conclusion, in our proposal, the combination of the results of the different clustering techniques by intersecting the clusters is quite an intuitive and straightforward method. However, studies could examine much more sophisticated techniques. For instance, the heterogeneous large-scale group decision maker approach applied by Chao et al. (2021) is mostly based on *fuzzy* clustering, where we could consider each clustering technique as a decision-maker. Thus, the heterogeneous LSGDM allows us to modulate the weight of each technique upon integrating them (Table 15).

Acknowledgements
We are grateful to the anonymous referees of the journal for their useful suggestions for improving the quality of the article. We are also very grateful for the selflessness, and effort of the huge community of R developers https://www.r-project.org/, which we relied on to develop our research.

Authors' contributions
The initial idea was conceived by JA. The experiments were designed using the LL. The search on the databases, statistical analysis, and software design was performed using LL. The work was drafted by LL and revised thoroughly, by JA. All the authors read and approved the final manuscript.

Funding
Funding was provided by EIT Digital (Grant no 825215) and European Cooperation in Science and Technology (COST Action 19130).

Availability of data and materials
The datasets generated and/or analyzed during the current study are available in the OSF repository https://osf.io/mr4pg/?view_only=4053eb6e7b46421b9bae194151997ae8.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 26 September 2020 Accepted: 14 November 2021
Published online: 12 January 2022

References

Acharya JGA (2011) Cluster ensembles. WIRES Data Mining and Knowledge discovery 1(4):305–315

Aggarwal CC, Reddy KC (2013) Data clustering: algorithms and applications, 1st edn. Chapman & Hall/CRC

Aghabozorgi S, Teh YW (2014) Stock market co-movement assessment using a three-phase clustering method. Expert Syst Appl 41(4, Part 1):1301–1314. https://doi.org/10.1016/j.eswa.2013.08.028

Aghabozorgi S, Shirkhorshidi AS, Wah TY (2015) Time-series clustering-a decade review. Inf Syst 53:16–38

Agresti A (2018) An introduction to categorical data analysis. Wiley

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recognit 46:243–256

Arroyo J, Maté C (2009) Forecasting histogram time series with k-nearest neighbours methods. Int J Forecast 25(1):192–207. https://doi.org/10.1016/j.ijforecast.2008.07.003

Arroyo J, González-Rivera G, Maté C, San Roque AM (2011) Smoothing methods for histogram-valued time series: an application to value-at-risk. Stat Anal Data Min The ASA Data Sci J 4(2):216–228. https://doi.org/10.1002/sam.10114

Bacon CR (2008) Practical portfolio performance measurement and attribution. The Wiley Finance Series. John Wiley & Sons

Baek C, Elbeck M (2015) Bitcoins as an investment or speculative vehicle? a first look. Appl Econ Lett 22(1):30–34

Balcilar M, Bouri E, Gupta R, Roubaud D (2017) Can volume predict bitcoin returns and volatility? a quantiles-based approach. Econ Model 64:74–81. https://doi.org/10.1016/j.econmod.2017.03.019

Banerjee A, Dave RN (2004) Validating clusters using the hopkins statistic. In: 2004 IEEE international conference on fuzzy systems (IEEE Cat. No.04CH37542), vol 1, pp 149–153

Begum N, Ulanova L, Wang J, Keogh E (2015) Accelerating dynamic time warping clustering with a novel admissible pruning strategy. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '15, pp 49–58. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2783258.2783286

Berndt DJ, Clifford J (1994) Using dynamic time warping to find patterns in time series. In: KDD workshop, vol 10, pp 359–370. Seattle, WA

Bojanowski M, Edwards R (2016) alluvial: R Package for Creating Alluvial Diagrams. R package version: 0.1-2. https://github.com/mbojan/alluvial

Bonanno G, Caldarelli G, Lillo F, Micciche S, Vandewalle N, Mantegna RN (2004) Networks of equities in financial markets. Eur Phys J B Condens Matter 38(2):363–371. https://doi.org/10.1140/epjb/e2004-00129-6

Brauneis A, Mestel R (2019) Cryptocurrency-portfolios in a mean-variance framework. Finance Res Lett 28:259–264

Brida J, Risso W (2009) Dynamics and structure of the 30 largest North American companies. Soc Comput Econ 35(1):85–99

Burggraf T (2019) Risk-based portfolio optimization in the cryptocurrency world. Inf Syst Econ eJournal. https://doi.org/10.2139/ssrn.3454764

Burniske C, Tatar J (2017) Cryptoassets: the innovative investor's guide to bitcoin and beyond. McGraw-Hill Education. https://books.google.es/books?id=-5AtDwAAQBAJ

Chan S, Chu J, Nadarajah S, Osterrieder J (2017) A statistical analysis of cryptocurrencies. J Risk Financ Manag 10(2):12

Chao X, Kou G, Peng Y, Viedma EH (2021) Large-scale group decision-making with non-cooperative behaviors and heterogeneous preferences: An application in financial inclusion. Eur J Oper Res 288(1):271–293. https://doi.org/10.1016/j.ejor.2020.05.047

Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) NbClust: an R package for determining the relevant number of clusters in a data set. J Stat Softw 61(6):1–36

Chaudhuri TD, Ghosh I (2015) Using clustering method to understand Indian stock market volatility. Commun Appl Electron 2(6):35–44

Corbet S, Meegan A, Larkin CJ, Lucey B, Yarovaya L (2018) Exploring the dynamic relationships between cryptocurrencies and other financial assets. Econ Lett 165:28–34

Corbet S, Lucey B, Urquhart A, Yarovaya L (2019) Cryptocurrencies as a financial asset: a systematic analysis. Int Rev Financ Anal 62:182–199

Drozdz S, Gebarowski R, Minati L, Oswiecimka P, Watorek M (2018) Bitcoin market route to maturity? evidence from return fluctuations, temporal correlations and multiscaling effects. Chaos Interdiscip J Nonlinear Sci 28(7):071101. https://doi.org/10.1063/1.5036517

Drozdz S, Minati L, Oswiecimka P, Stanuszek M, Watorek M (2019) Signatures of the crypto-currency market decoupling from the forex. Future Internet 11(7):154. https://doi.org/10.3390/fi11070154

Drozdz S, Minati L, Oswiecimka P, Stanuszek M, Watorek M (2020) Competition of noise and collectivity in global cryptocurrency trading: Route to a self-contained market. Chaos Interdiscip J Nonlinear Sci. https://doi.org/10.1063/1.5139634

Drozdz S, Minati L, Oswiecimka P, Stanuszek M, Watorek M (2020) Complexity in economic and social systems: cryptocurrency market at around covid-19. Entropy 22(9):1043

D'Urso P, De Giovanni L, Massari R (2016) Garch-based robust clustering of time series. Fuzzy Sets Syst 305(C):1–28. https://doi.org/10.1016/j.fss.2016.01.010

D'Urso P, De Giovanni L, Massari R (2019) Trimmed fuzzy clustering of financial time series based on dynamic time warping. Ann Oper Res 229. https://doi.org/10.1007/s10479-019-03284-1

D'Urso P, Cappelli C, Di Lallo D, Massari R (2013) Clustering of financial time series. Phys A Stat Mech Appl 392(9):2114–2129

D'Urso P, Giovanni LD, Massari R, D'Ecclesia RL, Maharaj EA (2020) Cepstral-based clustering of financial time series. Expert Syst Appl 161:113705. https://doi.org/10.1016/j.eswa.2020.113705

Fang F, Ventre C, Basios M, Kong H, Kanthan L, Li L, Martinez-Regoband D, Wu F (2021) Cryptocurrency trading: a comprehensive survey. arxiv:2003.11352

Fisher RA (1922) On the interpretation on teh x2 from contingency tables and the calculation of the p. R Stat Soc 85(1):87–94

González-Rivera G, Arroyo J (2012) Time series modeling of histogram-valued data: the daily histogram time series of s&p500 intradaily returns. Int J Forecast 28(1):20–33. https://doi.org/10.1016/j.ijforecast.2011.02.007

Gubu L, Rosadi DA (2020) Robust mean variance portfolio selection using cluster analysis: a comparison between kamila and weighted K-mean clustering. Asian Econ Financ Rev 10(10):1169–1186

Hartigan JA, Wong MA (1979) Algorithm as 136: a k-means clustering algorithm. J R Stat Soc Ser C (Applied Statistics) 28(1):100–108

Henning C, Meila M, Murtagh F, Rocci R (2016) Handbook of cluster analysis. CRC Press

Hornik K (2005) A CLUE for CLUster Ensembles. Journal of Statistical Software **14**(12). https://doi.org/10.18637/jss.v014.i12

Hornik K (2019) Clue: Cluster Ensembles. R package version 0.3-57. https://CRAN.R-project.org/package=clue

Hu AS, Parlour CA, Rajan U (2019) Cryptocurrencies: stylized facts on a new investible instrument. Financ Manag 48(4):1049–1068

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218

Häusler K, Xia H (2021) Indices on cryptocurrencies: an evaluation. IRTG 1792 Discussion Papers 2021-014, Humboldt University of Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series". https://ideas.repec.org/p/zbw/irtgdp/2021014.html

Irpino A, Verde R (2006) Dynamic clustering of histograms using wasserstein metric. COMPSTAT 2006. Proceedings in Computational Statistics. Physica-Verlag, Heidelberg, pp 869–876

Irpino A (2016) HistDAWass Package: An R Tool for Histograms-values Data. R package version 1.0.4. https://cran.r-project.org/package=HistDAWass

Irpino A, Verde R (2015) Basic statistics for distributional symbolic variables: a new metric-based approach. Adv Data Anal Classif 9:143–175

Irpino A, Verde R, De Carvalho Francisco de AT (2014) Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Syst Appl 41(7):3351–3366. https://doi.org/10.1016/j.eswa.2013.12.001

Kern M, Lex A, Gehlenborg N, Johnson CR (2017) Interactive visual exploration and refinement of cluster assignments. BMC Bioinform 18(1):1–13

Kou G, Peng Y, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using mcdm methods. Inf Sci 275:1–12. https://doi.org/10.1016/j.ins.2014.02.137

Lemire D (2008) Faster retrieval with a two-pass dynamic-time-warping lower bound. CoRR. arxiv:0811.3301

Letra IJS (2016) What drives cryptocurrency value? a volatility and predictability analysis. PhD thesis, Instituto Superior de Economia e Gestão

Liao S-H (2007) Mining stock category association and cluster on Taiwan stock market. Expert Syst Appl 35:19–29

Liao S-H, Chou S-Y (2013) Data mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio. Expert Syst Appl 40(5):1542–1554. https://doi.org/10.1016/j.eswa.2012.08.075

Liao TW (2005) Clustering of time series data-a survey. J Pattern Recognit Soc 1(38):1857–1874

Liu W (2019) Portfolio diversification across cryptocurrencies. Financ Res Lett 29:200–205

L'Yi S, Ko B, Shin D, Cho Y-J, Lee J, Kim B, Seo J (2015) XCluSim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. BMC Bioinform 16(S11):5

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley sympposium on mathematical statistics and probability, vol 1, pp 281–297

Mallikarjuna M, Rao RP (2019) Evaluation of forecasting methods from selected stock market returns. Financ Innov 5(1):1–16. https://doi.org/10.1186/s40854-019-0157-x

Mantegna R (1999) Hierarchical structure in financial markets. Eur Phys J B 11(1):193–197

Marti G, Nielsen F, Bi'nkowski M, Donnat P (March 2017) A review of two decades of correlations, hierarchies, networks and clustering in financial markets. Papers 1703.00485, arXiv.org. https://arxiv.org/abs/1703.00485

Mehta CR, Patel NR (1983) A network algorithm for performing fisher exact test in rxc contingency table. J Am Stat Assoc 78(382):427–434

Mehta CR, Patel NR (1996) Exact tests^{TM}. SPSS exact tests 7:12

Mizuno T, Takayasu H, Takayasu M (2006) Correlation networks among currencies. Phys A Stat Mech Appl 364:336–342. https://doi.org/10.1016/j.physa.2005.08.079

Nakamoto S (2009) Bitcoin: A peer-to-peer electronic cash system. http://www.bitcoin.org/bitcoin.pdf

Nanda SR, Mahanty B, Tiwari MK (2010) Clustering Indian stock market data for portfolio management. Expert Syst Appl 37:8793–8798

Newman M (2005) Power laws, pareto distributions and zipf's law. Contemp Phys 46(5):323–351. https://doi.org/10.1080/00107510500052444

Nguyen Cong L, Wisitpongphan N, Meesad P, Unger H (2014) Clustering stock data for multi-objective portfolio optimization. Int J Comput Intell Appl. https://doi.org/10.1142/S1469026814500114

Noirhomme-Fraiture M, Brito P (2011) Far beyond the classical data models: symbolic data analysis. Stat Anal Data Min ASA Data Sci J 4(2):157–170. https://doi.org/10.1002/sam.10112

Onnela J-P, Chakraborti A, Kaski K, Kertész J, Kanto A (2003) Dynamics of market correlations: taxonomy and portfolio analysis. Phys Rev E. https://doi.org/10.1103/physreve.68.056110
Pele D, Wesselhöfft N, Härdle W, Kolossiatis M, Yannis Y (2020) A statistical classification of cryptocurrencies. https://ssrn.com/abstract=3548462
Peterson BG, Carl P, Boudt K, Bennet R, Ulrich J, Zivot E, Lestel M, Balkissoon K, Wuertz D (2018) PerformanceAnlytics: econometric tools for performance and risk analysis. R package version 1.5.2. https://cran.r-project.org/package=PerformanceAnalytics
Platanakis E, Sutcliffe C, Urquhart A (2018) Optimal vs naïve diversification in cryptocurrencies. Econ Lett 171:93–96
R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. http://www.R-project.org/
Rani S, Sikka G (2012) Recent techniques of clustering of time series data: a survey. Int J Comput Appl 52(15):1–9
Rivin I, Scevola C (2018) The cci30 index. arXiv: General Finance
Rodriguez A, Laio A (2014) Clustering by fast search and find of density peaks. Science 344(6191):1492–1496. https://doi.org/10.1126/science.1242072
Rosvall M, Bergstrom CT (2010) Mapping change in large networks. PLoS ONE 5(1):e8694
Sahoo PK, Sethi D, Acharya D (2019) Is bitcoin a near stock? linear and non-linear causal evidence from a price-volume relationship. Int J Manag Financ. https://doi.org/10.1108/IJMF-06-2017-0107
Sarda-Espinosa A (2019) Dtwclust: time series clustering along with optimizations for the dynamic time warping distance. R package version 5.5.6. https://CRAN.R-project.org/package=dtwclust
Sardá-Espinosa A (2019) Time-series clustering in R using the dtwclust package. R J 11(1):22–43
Scrucca L, Fop M, Murphy TB, Raftery AE (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J 8(1):289–317
Sigaki HYD, Perc M, Ribeiro HV (2019) Clustering patterns in efficiency and the coming-of-age of the cryptocurrency market. Sci Rep. https://doi.org/10.1038/s41598-018-37773-3
Soleymani F, Vasighi M. Efficient portfolio construction by means of cvar and k-means++ clustering analysis: evidence from the nyse. Int J Financ Econ. https://doi.org/10.1002/ijfe.2344
Song J, Chang W, Song J (2019) Cluster analysis on the structure of the cryptocurrency market via bitcoin–ethereum filtering. Phys A Stat Mech Appl. https://doi.org/10.1016/j.physa.2019.121339
Stosic D, Stosic D, Ludermir TB, Stosic T (2018) Collective behavior of cryptocurrency price changes. Phys A Stat Mech Appl 507:499–509. https://doi.org/10.1016/j.physa.2018.05.050
Szetela B, Mentel G, Bilan Y, Mentel U (2021) The relationship between trend and volume on the bitcoin market. Eurasian Econ Rev 11:25–42. https://doi.org/10.1007/s40822-021-00166-5
Watorek M, Drozdz S, Kwapien J, Minati L, Oswiecimka P, Stanuszek M (2020) Multiscale characteristics of the emerging global cryptocurrency market. Phys Rep. https://doi.org/10.1016/j.physrep.2020.10.005
Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY et al (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37
Yates F (1984) Tests of significance for 2 x 2 contingency tables. R Stat Soc 147(3):426–463
Yermack D (2013) Is bitcoin a real currency? An economic appraisal. Working Paper 19747, National Bureau of Economic Research. https://doi.org/10.3386/w19747. http://www.nber.org/papers/w19747
Zhang W, Wang P, Li X, Shen D (2018) Some stylized facts of the cryptocurrency market. Appl Econ 50(55):5950–5965
Zieba D, Kokoszczyski R, Sledziewska K (2019) Shock transmission in the cryptocurrency market. is bitcoin the most influential? Int Rev Financ Anal 64:102–125

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.