RESEARCH

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

Luis Lorenzo[1*†] and Javier Arroyo[2,3]

*Correspondence: luislore@ucm.es
[1]Complutense University, Faculty of Statistical Studies, Madrid, Spain
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Since the appearance of Bitcoin, cryptocurrencies have experienced enormous growth not only in terms of capitalization but also in number. As a result, the cryptocurrency market can be an attractive arena for investors as it offers many possibilities, but a difficult one to understand as well. In this work, we aim to summarize and segment the whole cryptocurrency market in 2018 with the help of data analysis tools. We will use three different partitional clustering algorithms each of them using a different representation for cryptocurrencies, namely: yearly mean and standard deviation of the returns, distribution of returns, and time series of returns. Since each representation will provide a different and complementary perspective of the market, we will also explore the combination of the three clustering results to obtain a fine-grained analysis of the main trends of the market. Finally, we will analyze the association of the clustering results with other descriptive features of the cryptocurrencies, including the age, technological attributes, and financial ratios derived from them. This will help to enhance the profiling of the clusters with additional insights. As a result, this work offers a description of the market and a methodology that can be reproduced by investors that want to understand the main trends on the market and that look for cryptocurrencies with different financial performance.

Keywords: Fintech; Data Sciences; Cryptocurrency; Electronic market; Clustering

Content

The cryptocurrency market consists of more than 4,000 cryptocoins[1] with over 800 trades per second and more than 280 exchanges. It has become a huge new market in a very short term, considering that *Bitcoin* (Nakamoto, 2009), first peer-to-peer and decentralised digital currency was created in 2008 and the first bitcoin was mined in 2009. While cryptocurrencies were originally intended to enable anonymous wire transfers and online purchases, they have become a powerful investment tool.

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

As other assets, the value of cryptocurrencies swing based on news events, but cryptocurrencies have no physical assets or governments to back their value. More-

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

over, the cryptocurrency market is new, based on a still developing technology, highly speculative and small in comparison to others. As a result, it is highly volatile with big upswings, bubbles, and sudden market downturns.

Being a market so novel, big, diverse and volatile, it needs to be understood. Several categorization efforts have been made so far. For example, the *Cryptocompare* website [2] analyzed over 200 cryptoassets according to regulatory aspects, level of decentralization, supply issuance, economic incentive and others. Such taxonomy is useful even if it only covers approximately the 5% of existing cryptocurrencies at that time. Another example, Burniske and Tatar (2017) classify over 200 cryptocurrencies into three classes of assets based on traditional financial markets, namely: capital asset, consumable/transformable assets and store of value asset. However, this classification is highly subjective as many times the cryptocurrencies may be a combination of some of them. Furthermore, these approaches typically cover a small fraction of the cryptocurrencies, which are the most important ones in terms of volume and popularity, and focus on qualitative aspects or aspects that do not change much.

A different approach consists of analyzing the financial performance of the cryptocurrencies and describing it from a statistical point of view. Chan et al. (2017) analysed 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) analyse the stylized facts and return properties of 222 cryptocurrencies and find 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 due to their tail behaviour. However, their cluster results also reveal that the behaviour of the cryptocurrencies is diverse.

The same conclusion can be drawn in other clustering analysis using cryptocurrencies. Stosic et al. (2018) represent the correlations of 119 cryptocurrencymarket as a complex network and discover distinct community structures in its minimum spanning tree. Song et al. (2019) analyse 76 cryptocurrencies using the 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 plays an important role as well (Sigaki et al., 2019) clustering 437 time series of cryptocurrencies using hierarchical techniques where detect 4 different groups with a behavior evolving differently in terms of efficiency for the information.

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

[2] https://cryptocompare.com

However, it would be possible to combine the clustering results using different representations of the cryptocurrencies where each one take into account different aspects of the cryptocurrencies. In this way, the combination of the clustering results would make 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.

In this work, we will explore the combination of clustering of cryptocurrencies using R (R Core Team, 2013) as a main data-tool, we will support our investigation on the huge amount of R libraries available. We will go beyond a few hundreds cryptocurrencies as most studies do, and we will explore all the cryptocurrencies on the market in 2018 (more than 1,700 cryptocurrencies) by a methodology that is easily scalable for a growing and dynamic market. We will describe each cryptocurrency considering the log-return transformation of the daily price in 2018 with 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 the dispersion of the returns. In the second case, we consider the whole distribution of returns that accounts not only for the central tendency and dispersion of an asset, but for the whole aggregated behavior including asymmetry, kurtosis and the tails. Methods to analyze 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). Finally, we consider the observed data, that is the log return time series that accounts for variations over time and makes possible to identify when volatile or stable periods take place in each cryptocurrency.

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

Then, we will combine the three clustering results and analyze the most numerous intersections with the help of visual tools. Such approaches are successfully used in biostatistics (Kern et al., 2017; L'Yi et al., 2015). In our case, we will use them to represent the main trends in the cryptocurrency market. If several cryptocurrencies belong to the same clusters in the three clustering results then we can consider them as very similar. We will inspect the relationship among the three clustering results with the help of visualization tools.

The approach proposed provides a screening mechanism that allows us a meaningful exploration of the whole market in spite of its complexity and size. The

intersection of the clustering results can also be helpful for investors in order to select a suitable cryptocurrency for the portfolio as it characterizes the market in more detail.

In a further step, we will investigate the association between the clustering results and different features of the cryptocurrencies such as technological variables, the market capitalization, the *maturity* (age) of the cryptocurrency, and some of asset portfolio ratios. We aim to inspect whether some clusters are tightly associated or not with some aspects not taken into account the clustering process. we apply some inference statistical tests to assess whether associations are significant. These associations enhance the profiling of the different clusters. We keep continuous references to concrete cryptocurrencies of the market, most of them no very known, that are part of our analysis. Finally, we present our conclusions with some points open to debate.

Related work

Clustering financial data

Clustering analysis is a well-known data analysis tool that have served well in different fields (Henning et al., 2016). In particular 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. Later, Bonanno et al. (2004) apply the same methodology considering different time horizons comparing the return and volatility networks. The methodology of Mantegna is applied with different variations in other contexts (Brida and Risso, 2009; Mizuno et al., 2006; Onnela et al., 2003). Furthermore, Marti et al. (2017) propose different alternatives and variants of this methodology.

Another important strand of applies fuzzy clustering to financial time series typically grouping stocks for developing portfolios. For example, D'Urso et al. (2013) and D'Urso et al. (2016) apply 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) propose a fuzzy clustering method based on cepstral representation using the daily Sharpe ratio as variable of clustering.

The main application of clustering in finance is building portfolios. For example, Nanda et al. (2010) apply K-means, Fuzzy C-means and Self Organizing Maps (SOM) to returns and financial ratios from Indian stocks to classify them in different clusters and subsequently develop portfolios from these cluster. 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. That associations help to analyze and describe the co-movement among the different markets.

Regarding the use of time series as objects to cluster, Aghabozorgi and Teh (2014) propose 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 composing 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) analyses Bitcoin market in-depth and consider it an investment more speculative than a currency.It is considered that it poses high risk for the management of transactions and credit markets. Finally, a deflationary scenario is anticipated because of the limited number of bitcoins that can be issued (21 millions). The paper anticipated many aspects of the cryptocurrency markets that we are experiencing today (excessive volatility, high level of computer knowledge required for using and integration into the web of international payments). A more updated vision on this innovative market is regarding the cryptocurrency exchange, Drozdz et al. (2018, 2019) show that BTC/USDT, ETH/USDT ETH/BTC were almost indistinguishably from exchange rates quotes on Forex market. The authors show that the exchange of cryptocurrencies have a behavior similar to more mature markets such as stocks, commodities or Forex. Complementary, latest study Drozdz et al. (2020a) points to the anticipated disconnection of the cryptocurrency from the conventional markets and that the Bitcoin on the cryptomarket comes to plays a similar role as the USD in the Forex or Drozdz et al. (2020b) where shows that cryptocurrencies began to be correlated with traditional assets only from 2020.

The high growth of the cryptocurrency market and its heterogeneity since 2014 was analyzed in depth by Corbet et al. (2019), who consider different aspects including regulatory, cyber-criminality, market efficiency or bubble dynamics and make recommendations for further investigations on different domains. We take a couple of them and we address in our work some characteristics based on liquidity with the volume as a proxy, market cap and other key metrics or ratios, for instance, Beta or Sharpe ratio.

The characterization of the cryptocurrencies from a statistical point of view has been tackled by different works. Chan et al. (2017) analyze the distributions for 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. Heavy-tail and associated power-law distribution analysis will be taken account on our study. As part of a benchmark with other markets, Baek and Elbeck (2015) show that the volatility of Bitcoin market shows that bitcoins are 26 times more volatile than S&P 500 Index.

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

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*, and 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) helps to find that Bitcoin shock prices are not transmitted to the prices of other cryptocurrencies, being *Litecoin* and *Dogecoin* the more influential actors. According to the results, Bitcoin exhibits a lower relationship with other cryptocurrencies. Other approach is the use of random matrix theory and hierarchical structures in a MST on 119 cryptocurrencies from 2016 to 2018 (Stosic et al., 2018). They find the presence of multiple collective behaviors in the market of cryptocurrencies, which contrast to the intuitive idea that Bitcoin exerts a global influence on the entire market.

Furthermore, the time dimension can also be taken into account. 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 evidence 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). In turn, Pele et al. (2020) merge classification based on asset profiles and dynamic evolution of clusters. First, they characterize of selected group of log-returns assets including 150 cryptocurrencies, stocks commodities and exchange rates to estimate a multidimensional vector applying a dimensionality reduction with factor analysis. Then, they use classification where K-means is one of the techniques applied. The main difference of cryptocurrencies with respect to traditional assets is a higher variance and longer tails of the log-return distribution. The work also shows that individual cryptocurrencies tend to develop over time similar characteristics (synchronic evolution).

Methodology

Dataset description

We retrieved data from `https://www.cryptocompare.com/` for all the cryptocurrencies traded during 2018. Many new cryptocurrencies appeared in the last years, but many of them were short-lived and barely traded. We aim to include in our study as many of them as possible. Firstly, we remove *NaN* and *Inf* values, which were mostly caused by zero-prices in log transformations. Secondly, we filter out those cryptocurrencies that were in the market less than the 95% of the days (92 cryptocurrencies in 2018). We kept for clustering those that were in the market but were not traded, i.e. zero return and volatility or zero volume, because they are part of the market. In 2018 there were 306 cryptocurrencies on the exchange that were

not traded at all. Still we decided to include them onthe clustering, because they are a substantial part of the cryptocurrency market. Even if they have no interest for investor we are interested to know where are allocated.

Our final dataset in 2018 consist of 1,723 cryptocurrencies. However, we decide to remove those cryptocurrencies with low or no activity from the second part of our analysis, the association tests though included on clustering part (first part). Low market activity may cause heavy tails on the return distribution and affect to the consistence of the results. The remaining dataset for association tests consist of 1,262 cryptocurrencies with the higher statistical quality ensuring the existence of first and second statistical moments.

We also download the data for 2019 to extend our experiment for a longer time-frame analysing the generalization of the results.

Besides the cryptocurrency data, we use the CCI30 daily data [3]. The CCI30 is a market cap weighted index that represent the 30 largest cryptocurrencies. We also retrieve data from the US Department of the Treasury[4] that we use for the computation of some financial benchmarking rates as Beta and Sharpe ratio that we will explain in the next sections.

For 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 due to its more suitable statistical properties and better comparability. It has been used in cryptocurrency markets as well Letra (2016); Stosic et al. (2018). The return for the cryptocurrency i at 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 i cryptoasset at day t.
- **Heavy tail**: Heavy tail behaviour 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 are interested in identifying the cryptocurrencies prone to extreme behavior and, whether they associate with some cluster. We define a cryptocurrency with heavy tails behaviour by a binary variable if it has a tail index lower than 2 according to Newman (2005)
  This would question the existence of finite first and second moment of the underlying distributions, which is not a problem in our case, since we use the observed sample statistics in a descriptive manner.
- **Volume**: It is the daily traded volume in units of the base cryptocurrency that is used as a liquidity proxy. We transform the volume into an ordinal

[3] https://cci30.com/

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

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

variable by 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%), and 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*, namely `WBTC*` (26.8%), `BTC`(22.4%) and `NPC` (11.5%), and in total 5 cryptocurrencies (`WBTC*, BTC, NPC, XRP, AMIS`) represent 80% of the total *market cap*.
- **Technological variables**: We represent the encryption and consensus algorithms of the cryptocurrency as nominal variables:
  - **Encryption**: There are 105 different values. The more relevant are *Scrypt*, *SHA256*, *SHA256D*, *X11*, *X13*, *X15*, *PoS*, *Multiple* and *CryptoNight*. We notice that this information is not available for 35% of the 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* though 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 on our study. This variable is qualify by a quantile function as well.

Methods

We aim to group the cryptocurrencies based on the behavior of their log-returns in 2018 and describe later. For such purpose we will 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 because the centroids provide us an interpretable summary of the elements of each cluster, which will help us to identify the most relevant features of the cluster elements. A drawback in this type of clustering algorithms is that they assume knowledge about the desired number of clusters (k). We apply different quality criteria to determine the optimum number of clusters depending on the technique.

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

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

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

For the bi-variate (or two-moments) representation, where the two variables are the yearly mean and standard deviation of the log-returns we use the K-MEANS clustering (MacQueen, 1967), which is one of the most widely used clustering algorithms (Wu et al., 2008). We standardize 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 the clustering, we compute the Hopkins statistic (Banerjee and Dave, 2004) to rule out the possibility that a uniform random distribution generated the data set.

For the selection of 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) aiming at reducing the randomness on the partitional cluster results. We run the K-means algorithms 10 times and we ensemble the outcomes 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 considering 50 random starts for each iteration. Once we have the 10 algorithm runs, we compute the medoid of an ensemble of partitions, i.e, the element of the ensemble minimizing the sum of dissimilarities to all other elements (Hornik, 2005, 2019).

Dynamic clustering algorithm for histograms

For the yearly log-return distribution, we apply a clustering algorithm that deals with 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). In this way, we will 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's the quantile functions of h_1 and h_2, respectively. This distance can be decomposed

as follows:

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

where μ_i and σ_i are respectively the mean and the standard deviation of the h_i and $\rho_{1,2}$ is the correlation of h_1 and h_2 (Irpino and Verde, 2015). As a result, the l_2 Wasserstein distance can be decomposed in the addition of three elements that account for the histogram differences in terms of location, spread and shape, respectively. Interestingly, this distance matches the perceptual similarity that the human observes 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 the average histogram of the observed histograms 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). As a result, the prototypes can be interpreted in a financial context as log-return distributions.

We use the clustering implementation in the R-package Hist-DAWass (Irpino, 2016). This implementation provides quality measure that is the percentage of 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 and the solution is the best one 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 due to the high dimensionality of the objects and how they are ordered. As a result, many approaches have bee proposed over time (Aghabozorgi et al., 2015; Liao, 2005; Rani and Sikka, 2012).

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

However, DTW is intrinsically slow because of its quadratic time complexity, which hampers its applicability in clustering. Thus, we use the enhanced DTW algorithm TADPole (Time-series Anytime Density Peak) (Begum et al., 2015) that 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 speed-up the convergence of the algorithm. As a result, TADPole produces a right answer quicker, and then refines it until it converges to the exact answer. Besides, the clustering algorithm only needs two parameters which makes it easy to use. Firstly a cut-off distance that define the thresholds to select the series and we set it to 2; and secondly, a windows-size that define the time frame to make the comparison between the series that we set to 3. Optionally we can also select the number of clusters (k) or we can let the algorithm to 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 referred to the local density and δ_i is the distance from points with higher local density.

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

TADPole allows to cluster time-series with arbitrary shapes which is very useful in our case because of the heterogeneity of the cryptocurrency market. In contrast, TADPole clusters because of the different distance metrics used cannot be represented as "balls" in a metric plane as in K-means for example. The result is a Partition Around Medoid (PAM) type centroid using the DTW distance that can be represented only in a DTW space. This centroid is a time-series that serves us to identify the volatility patterns of the resulting clusters.

We apply the implementation of TADPole algorithm of the R-libraries DTWCLUST by Sarda-Espinosa (2019); Sardá-Espinosa (2019). The time-series are of log-return values which facilitate the characterization of the clusters from financial perspective. The DTW measure implemented in the package follows the estimation by 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 that we obtain for the clustering algorithm n. The combination of the clustering results make possible to characterize each cryptocurrency in several dimensions, one for each cluster strategy. The resulting multi-dimensional categorical datasets can be shown using visualization techniques supported on the graph theory (Kern et al., 2017; L'Yi et al., 2015). To better highlight the changes in the clustering between the different techniques, we have visualized such changes by means of a so called alluvial diagram taken 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.

Also, we can compare numerically two partitions represented as a $c_1 \cdot c_2$ matrix $N = n_{ij}$ 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 are

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 agrees, that is, that 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, it means that more cryptocurrencies share the clusters for the different partitions. We will use the function implemented in the R package MCLUST by Scrucca et al. (2016). We will focus also on the cluster intersections with the higher cardinality for a better profiling of the main trends of the cryptocurrency market.

Association test

Finally, we enhance the descriptive information of each cluster by studying 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 the exact Fisher's tests and analyzing the Person's residuals of the contingency tables that we explain later. Firstly, we transform quantitative variables into ordinal by quantile functions.

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

Beta is a volatility measure of 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, the CCI30 index that tracks the 30 largest cryptocurrencies by market capitalization. .

The *Beta* value shows if an asset moves in the same direction as the reference index, and how volatile or risky is compared with it. The *Beta* for the whole market is 1.0. A positive beta means the asset moves in the same direction as the market, while negative beta means that the asset moves in opposite direction. Furthermore,

[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.

an absolute value higher than 1 means greater sensitivity to systematic risk, i.e. higher risk; while values lower than 1 mean less sensitivity.

The **Sharpe ratio**(Sharpe variable) is the exceed average return of risk-free by volatility unit or total risk. The ratio determines the risk of the 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 our cryptocurrency, σ_c is standard deviation or the volatility of our cryptocurrency and R_f is the *risk-free* rate taken as reference; we considered the daily of the annualized T-Bill over 90 days, and its daily value for 2018 was $E[R_f] = 0,00525\%$, almost zero. The greater the value of the Sharpe ratio, the more attractive the risk-adjusted return of the cryptocurrency.

Typically, **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 it is not adequate when the sample size is small. We ruled out the Chi-Square test since results are not significant if the expected frequency is not typically higher than 5 in at least 80% of the cells of the contingency table (Yates, 1984) and this assumption is not fulfilled in our case for many of the categorical variables for some levels. we will use the Fisher's exact test (Fisher, 1922) to test the association between the variables of the Table 1 and the cluster results, which is applicable for all sample sizes. This test assumes no dependency between the categorical variables as null hypothesis, and assumes a multivariate hyper-geometric distribution for the cells into the contingency tables (Mehta and Patel, 1983).

For large datasets *Monte Carlo* method provides an unbiased estimate of the exact *p-value* (Mehta and Patel, 1996). *Monte Carlo* consist of 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 run 8,000 simulations for each association, i.e. for each pair of variables under analysis, generating simulated contingency tables filled with a sampling of a multivariate hyper-geometric distribution. Then we compute the probability that we have a distribution as we effectively have observed, that is the *p-value*. A cell-by-cell comparison of observed and estimated frequencies evidences the nature of the dependence. If the p-values of the Fisher association tests between a couple of variables is lower than 0.01 then we consider that the association is significant. For each significant association between categorical variables of the contingency table we analyze standardized (adjusted) Person's residuals for the cell ij (Agresti, 2018), which is 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 frequency, 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 about 2 when there a few cells, or about 3 when there are many cells indicates that the cell do not satisfy H_0 (Agresti, 2018). In our case, we assume a more conservative position and we consider as a cut-off for significant standardized residuals those that exceed 3.5.

Replication within a longer time-frame

We propose a methodology that is time-frame agnostic because aims to describe the behavior of a period of time, regardless its length or its frequency. We take 2018, a extremely active period in the cryptocurrency market, as use case. However, it would be interesting to replicate the methodology on other time periods and/or considering other time frequency.

At 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 as well. For that purpose, we consider an extended time-frame including both 2018 and 2019 years.

We will consider only the cryptocurrencies that were traded on the market during the whole period (730 days), that is 440 cryptocurrencies in total. For this particular shortlist, we will compute the associations on the extended period for the financial ratios (Beta and 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, so we use those that we took the 4th of February in 2019 as we explained in variable description section.

We re-run the three clustering techniques for the 2018-19 time-frame. However, for determining the number of clusters, we checked that the results were quite similar to those reached for 2018. Thus, for easing the comparison we chose to use the exact same number of clusters used in 2018.

This experiment will help us to 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, the intersection clustering and finally the association tests. In Table 2 we summarize the three clustering results, showing for each cluster its cardinality 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

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

The optimum number of clusters according to the CVI indexes is 3. The descriptive statistics of the 3 centroids in ordinary values are shown in Table 2 and Figure 1(a) shows the scatter plot with the clusters.

The clustering algorithm clearly discriminate 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 the 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.

Figures 2, 3 and 4 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 plots that help us to locate the areas where 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 around 1% of the sample as we see in Fig. 2(a). The higher concentration of cryptocoins into this cluster is surrounding volatility 1.5 and mean return around -0.1 as we see in Fig. 2(b). `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`. We see into this cluster a mix of Ethereum tokens and cryptocoins with its own blockchain, most of them with low traded volume which may cause that a few operations trigger the volatility.
- **Cluster 2**: This cluster is the more populated with around 900 cryptocurrencies (52% of the total). It allocates the *moderate* behaviours including the higher mean return cryptoassets. It is also less homogeneous than the others, with different dense areas of concentration as we see in Fig. 3(a) which point out the existence of other cluster. Most of the higher capitalization cryptocurrencies (`BTC`, `EOS`, `ETC`, `ETH` or `LTC`) are in the sub-cluster with very low volatility and moderate negative return shown in Fig. 3(b). However, we can also find some cryptocurrencies with moderate positive returns (134 cryptocoins) and very low volatility, as shown in Fig. 3(c). 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 along 2018), `ETL` (EtherLite is a 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 a moderate activity in the first half of 2018, and was flat in the second); all of them can be considered low-medium market capitalization (under 70^{th} percentile). On the other hand, the detail of Fig. 3(c)

shows the high homogeneity on the selected area where does not appear any density contour curve.

- **Cluster 3**: This cluster has 801 cryptocurrencies, most of them with negative average returns, and volatility lower than 0.5. According to Fig. 4(a), the highest concentration of cryptocurrencies is located in mean return closer to zero and volatility around 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` that are close one to each others into the cluster (see Fig. 4(b)).

We confirm that K-MEANS identifies clearly three different behaviours on the 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 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 from the prototypes of the five clusters.

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

It is important to remark that the coefficients of variation for the centroids are quite different for the clusters ranging from -0.75 to -32.90, which point out that this clustering algorithm is specially sensitive to this particular statistic. This is particularly relevant in the financial context, since the coefficient of variation measures how much volatility is assumed in comparison to the amount of return expected from investments. However, since 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 much representative is the prototype of a cluster. According to this statistic, *Cluster 3* would be the more uniform cluster, while *Cluster 5* would be the more heterogeneous.

The first column of Fig. 5 represents the prototypes of the five clusters, while the rest of the columns show some of the relevant cryptocurrencies of each cluster. It is interesting that 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 behaviour. We describe them below:

- **Cluster 1**: The prototype in Fig. 5(a) has a mean return of -0.13 and the highest kurtosis (13.43). The standard deviation of this prototype is slightly

lower than that from *Cluster 1* prototype, however, the shape of the distribution is different, because here 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 that represent around at 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.5(b), `KEY` in Fig.5(c) (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. 5(d) (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) points out a quite homogeneous cluster. This cluster has a cardinality of 147 cryptocurrencies and concentrate all no-traded cryptocurrencies (92), lowest market cap (*P70*) for most of the cryptocurrencies into this cluster. Representative into the cluster by market cap are 365 in Fig.5(e), `ACN`, `CBX` or `ALT` in Fig.5(f).
- **Cluster 3**: The prototype shown in Fig. 5(g) 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 specially interesting given that it has the highest cardinality with more than 1000 cryptocurrencies (around 60% of the sample). Unsurprisingly, this cluster allocates the cryptocurrencies with the highest market capitalization, including `BTC` in Fig.5(h), `BCH`, `EOS`, `ETC`, `ETH` in Fig.5(i) and others (`HSR`, `ICX` or `LTC`). Given the size of the cluster, these cryptocurrencies represent the predominant behaviour in the market, which unsurprisingly is the most moderate behaviour and includes the most popular cryptocurrencies.
- **Cluster 4**: The prototype shown in Fig. 5(j) 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 is low too (-19.97). The cluster is not very homogeneous compared with the mean distribution (0.128). The cardinality of this cluster is low (around 60 cryptocurrencies) and some of the representative cryptocurrencies are NAS (most of the trading volume in 2^{nd} and 3^{rd} quarters of 2018), `NKC` (high trading volume since February 2018 and very important trading volume in Aug 18), `POLY` in Fig.5(k) (launched in January 2018), `FSN` (higher volume activity in 2nd and 3rd quarter of 2018 with a peak in August), `JNT` in Fig.5(l) (higher volume activity in 3rd quarter) or `MNTP` (no continuity on the trading volume with sporadic peaks).
- **Cluster 5**: The prototype shown in Fig. 5(m) has a mean returns 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 a moderate kurtosis compared with the others clusters (5.66). We find the highest negative and positive returns in this cluster. This cluster is the most heterogeneous compared with the mean distribution (1.116). Unsurprisingly, it has the lowest cardinality with only 16 cryptocurrencies, which

is around the 1% of the sample. Some of the representative cryptocurrencies are B2X in Fig.5(n) (low trading in 2018), ITT in Fig.5(o) (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 is possible to effectively discriminate the log-return distributions taking into account 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 6 represents on a time-axis, the medoids of each cluster so they are observed objects (time series of cryptocurrencies). Figure 7 shows the annual and quarterly density functions of the three medoids. Also, Fig. 7(a) represent how different is the density plot of the *Cluster 1* compared with the others corresponding with the higher volatile cryptocurrencies.

- **Cluster 1**: The medoid of this cluster in Fig. 7(b) shows a time-variation around 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 changes smoothly starting with a low but positive value the first quarter, negative the second and third, and positive the fourth. This cluster has the lowest cardinality (22 cryptoassets). The medoid of this cluster is the time-series LINK (Chainlink's native token, known as LINK, is used to pay the network's node operators, or oracles, for providing 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. 7(c) shows a consistent average returns above zero. The density functions have three modes and they are greater than or equal to zero. However, the last two quarters of 2018 exhibit fat negative tails with ranges over -0.1. The cardinality of this cluster represents around a 49% percentage of the cryptocurrencies and it includes 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 (called Tao coin as well is a token for music streaming services).
- **Cluster 3**: The medoid of this cluster in Fig. 7(d) has average returns below zero in all the quarters and the densities exhibit two modes smaller than or equal to zero and occasionally large positive returns. The cardinality of the cluster is around a 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 the remaining part of the year).

We can see that the TADPole clustering for the return time series effectively identifies three different clusters taking into account 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 the clusters, being cluster 1 the most homogeneous and *Cluster 3* the less. However, according to the standard deviation and the coefficient of variation the dispersion within the clusters is quite high.

Intersection of clusters

For the sake of comparison, Fig. 1 shows the three clustering results on the same annual return-volatility plane. In each plot, all cryptocurrencies are site in the same location, but the colour scheme in each plot represents the respective clustering results. In the plot, we have marked the cryptocurrencies with highest market capitalization and polygon vertices. We can see that most of them are very located in a precise area, below the point $(0, 0)$.

The polygons and the colours reveal that the results of the three techniques are overlapped because respond to a different dimensionality on the objects. The only exception is the *Cluster 1* of K-means in Fig.1(a) and *Cluster 5* of Hist-DAWass in Fig.1(b) which are mostly the same. These plots confirm that each clustering algorithm takes into account different aspects of the cryptocurrencies and that their combinations may provide us further insights on the cryptocurrency market. TADPole clustering in Fig.1(c) is the more different on the groups compared with previous techniques, overlapping all the cluster areas when represented on the same return-volatility plane that the other techniques.

We analyse now the main groups of cryptocurrencies that remains together through the three clustering algorithms, it is what we call *intersection of clusters*. Only 24 out of 45 ($3 \times 5 \times 3$ intersections) possibles are populated. Table 5 shows all intersections and those with the cardinality greater than 100 (first 6 intersections), represents the 75% of total market.

Intersection 1 and *2* have almost 300 cryptocurrencies each one. Both of them are characterized by cryptocurrencies that belong to *Cluster 2* and *3* in the K-means and Hist-DAWass algorithms, which unsurprisingly are the most populated clusters for each algorithm. Both of them are characterized by low volatility and (negative) close to zero average returns. However, in *Intersection 1* we can find *Cluster 3* of the TADPole algorithm, while in *Intersection 2*, we can find *Cluster 2*, which mainly differ in that in the first case 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` or `ETC`. While in Intersection 2 we find some of the most popular highest market cap cryptocurrencies (`BTC`, `LTC`, `XRP`), and some others with lower market cap and higher returns (`AE`, `USDT`, `ZRX`).

Intersection 3 and *4* have around 200 cryptocurrencies each one 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 to *Cluster 3* of the Hist-DAWass technique 5(g). The main difference with the previous intersections

is that *Cluster 3* of K-MEANS corresponds on average with negative daily mean-returns but moderate volatility 2(a) so the average returns are lower as well 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 the *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-return for all quarters 7(c). There is not any high returns cryptocurrencies on this intersection (DASH, SC, STRAT).

Finally, in *Intersection 5* and *6*, we have *Cluster 3* from the K-MEANS, *Cluster 1* from HIST-DAWASS and *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 average-high risk and average returns (CMT, ETT, HST). Intersection 6 allocate some cryptocurrencies with high market cap but low returns (BCD, SBTC, GEO).

In the alluvial plot shown in Fig. 8 we show how are related the different clusters of the different algorithms. It makes possible to appreciate both, the main trends already commented and those more subtle. For example, we can see that the smallest clusters in K-MEANS and HIST-DAWASS (*Cluster 1* and *5*, respectively) share most of the cryptocurrencies. In that sub-group of very volatile cryptocurrencies we find AMIS, B2X, ELTCOIN, FLLW, GOOD, ICE, ITT, LBTC, PFR, REX, RIPT, STAR, WAND, XIN, YOVI and ZCG. Then the group diverges and relates with the two main clusters of the TADPOLE without a clear pattern, which means that the temporal evolution is more conventional with mean return in quarterly basis positive or negative but not related with other multidimensionality on the objects. Curiously, the smallest TADPOLE *Cluster 1* is not strongly related with any other cluster. This means that its peculiar time series evolution is not particularly related with the prototypes of the aggregated representations of the others clustering techniques, namely the return distributions and mean-standard deviation bi-variate or two-moments representations.

Finally, we notice that DEUR is the only cryptocurrency that was not pair-combined with any other cryptocurrencies along the three techniques with not activity at all on the market aloong our analysis period.

Regarding the ARI values, we obtain extremely low agreement values. The highest one is 0.0123, which is very close to 0 that means no-agreement. We obtain this value for the agreement between the K-MEANS and HIST-DAWASS results. For the rest of 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 we explained in Methods section, we rely on exact Fisher tests based on *Monte Carlo* simulations for the significance tests of the associations between the variables. P-values of Fisher test are depicted graphically in Fig. 9 ,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 goal of the association tests is to enhance the characterization of the clusters adding value upon the prototyping descriptions that we explained in previous Section. We group in a red box the area with the associations between clusters and market categorical variables.

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 Person's residual of 7.02 and 17.08 respectively) but the association is also high on *Cluster 1* of K-means and Hist-DAWass showed on Table 14. We have already mentioned that *Cluster 1* in K-means allocate the highest volatile cryptocurrencies and in this case, they correspond with heavy-tails cryptocurrencies as well. As we said, *Cluster 2* of Hist-DAWass allocates the more negative-return cryptocurrencies so we can conclude that heavy-tails are stronger for left tail. Finally, there is not any clear link between TADPole technique and the heavy-tails distributions (very low values for the standardized Person's residuals) so for this last one we conlcude that there is not a relation between a shape-base clustering and the distribution characterization.

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 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 8.93.

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

For 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). While *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. Also, it is possible to see weaker but relevant associations in one of the lowest cardinality clusters (*Cluster 4*) with an standardized residual of 3.36 for *P80*.

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

Regarding the MKCap variable, in Table 7 we can see that the K-means association are not very strong. For example, we see the lowest and highest market caps

percentiles (*P70, P100*) sharing the same *Cluster 2* with standardized residuals of 3.91 and 3.57 respectively.

However, in the association between MKCap variable and HIST-DAWASS, we observe an important association between *Cluster 1* and the lowest market cap percentiles *P70* with the value 8.07. While *Cluster 3* is linked with 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, it is is strongly related to low volatility (*LowVol*) Betas (`BTC`, `DCN`, `WAVES` or `WBTC*`) .

Finally, *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 reveals that it properly discriminates three groups of different behavior that could interest the investor depending on his/her risk-aversion profile.

However, again we can confirm with the help of Table 8 the higher screening capacity of the HIST-DAWASS clustering. 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 because in that cluster we can find many of the components of the CCI30 index (`BTC, BCH, DASH, ETC, ETH, LTC`).

For 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 respect a risk-free asset or, in other words, the risk-reward for the investment on the asset (cryptoasset in our case). Interestingly, there is no association with *GOOD* label -higher value than 1.0- into 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 represented by a distribution with positive mean return as we showed in Table 2.

Summarizing, the clustering results of K-MEANS and HIST-DAWASS clusters are associated with the Market cap, Volume and Beta variables, and the TADPOLE results is the only associated with the *Sharpe ratio* .

Associations results for the intersection of clusters
As we can see in Fig. 9, 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 because the intersections successfully combine the idiosyncrasy of each clustering algorithm. Tables 6, 7, 8 and 9 show the association for the different categorical variables and the higher cardinality intersections (first 6 rows in Table 5).

Regarding the Volume variable, *Intersection 3* with standardize residuals of 6.21 and 6.21 and *Intersection 4* with standardized residual 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 4.02 as standardized residual. Finally, the lowest percentile *P70* is mostly allocated in *Intersection 1* (again it coincidences *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 4.88 and 4.87, respectively. The percentiles *P80, P90* are linked to the *Intersection 3* (`ADT, BTX, ION`), and *P80* with *Intersection 4* (`BAY, LEND, SKY`).

There are not *Extreme* Beta cryptocurrencies in the highest cardinality Intersections as we see in Table 8. However, we can see an association of *HighVol* with *Intersection 3* (`BCH, QTUM, XVG`) and *Intersection 4* (`ADA, HSR, STRAT`) with residuals of 6.20 and 6.62, respectively. On the other hand, *LowVol* is associated with *Intersection 2* (`BTC, WAVES, WBTC*`) with a residual of 5.62. Finally, *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 into the *Intersection 2* (`WAVES, XRP, ZEN`) with standardized residuals 4.13. Exceed Return Positive (*ERP*) are linked to *Intersection 2* (`AC, ZRC, ZRX`) and *Intersection 4* (`ADA, HSR, SC`) with standardized residual 5.20 and 4.67, respectively.

We can see that the intersections are associated with all considered categorical variables except with the technological ones that we review below. As a result, we can conclude that the intersections of the clustering results improve the characterization by means of the associations. The intersection inherit some of the associations of the different clustering results though the significance is lower.

Associations between clusters and the technological variables
It is worth mentioning that while technological variables are not associated with any clustering results, they have significant relationship with other independent financial variables as the market cap and volume of trading as we show in Fig. 9

(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 and the lower percentile (P70) of market cap as well. Encrypted algorithms *CryptoNight-V7, Ethash, Ouroboros* are also associated to the highest market cap percentile (P100).

Regarding the consensus algorithm, ProofType variable in Table 11, the algorithms *PoS* (3.74), *PoW/PoS* (9.09) and *PoW* (4.54) have relevant associations 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 those of *Cluster 2*, which was 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 shows more interesting associations. For example, *Cluster 1* is associated with cryptocurrencies in the deciles *D5, D6 and D7*, with standardized residuals 11.88, 7.91, 4.70, respectively. According to Table 3 this cluster is characterize by a extreme distribution with skewness (0.82) and kurtosis (13.43). Similarly, *Cluster 2* is linked to the cryptocurrencies in decile *D6* with a residual of 4.97.

On the other hand, the oldest cryptocurrencies are prominently associated to *Cluster 3* with a standardized residual of 15.08. Interestingly, *Cluster 3* is also the higher market cap clusters (`BCH, BTC, DASH, EOS, ETC, ETH, IOT,LINK, LTC, NEO, WAVES, XLM, XMR, XRP, ZEC, ZRX`), thus the most popular cryptocurrencies for investors.

Regarding the intersections of 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*, while the younger ones (*D4*) are significantly allocated in *Intersection 6* (`EBC, ICOB, PULSE`) and *Intersection 4*.

Again, we confirm that Hist-DAWass offer a stronger associations than K-means, and that the main intersections provide even more associations. As a result, clustering intersections are very good for characterizing cryptocurrencies due to their higher granularity and because they tend to display more significant associations better distributed.

Analysis of the extended time frame

According to Fig. 10, we can see that the shapes of the clusters in the extended period is quite the same that we have represented in 2018 (Fig. 1). We notice that the xy-axis have different ranges for volatility and mean return because of the difference on the data sets, mainly different on the number of cryptocurrencies, but comparing both time-frame periods 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), while for the TADPole is null (-0.0021). The TADPole result

can be explained because it uses the 'raw' data and not a summary, which makes more difficult to find similar trajectories through longer time periods. In addition, the TADPole clustering seems 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 quit similar to those in 2018 shown in Fig. 9 with some exceptions: no association between TADPole and ClassSharpeR variable, but instead we find a significant association between K-means and Hist-DAWass with ClassSharpeR. We again 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. Still the shape-based clustering as TADPole shown helpful to enhance the description of the market for the chosen time frame.

Discussion

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

- We confirm the existence of a structure on the market that allow us to segment the cryptocurrencies in clusters. However, the optimum number of clusters remains low, independently of the representation considered, which points out to a high degree of homogeneity despite of the high number of cryptocurrencies.
- The bi-dimensional or two-moments representation by the well-known K-means works quite well to segment separately by the returns and volatility in three major groups. However, Hist-DAWass offered a more subtle discrimination of the cryptocurrencies, for example, taking into account the combination of both moments. Thus, it seems a promising and suitable profiling tool for investors.
- The K-means partition is strongly associated with Beta values. This is not surprising since beta is computed using the mean return and volatility that are the variables considered for the clustering.
- Both K-means and Hist-DAWass partitions are associated with market capitalization and Volume. This is unexpected, because it reveals a connection between the shape of the return distribution and how prominent is the role of the cryptocurrency in the market.
- The K-means and Hist-DAWass clusters also show an interesting association with the age or maturity of the cryptocurrencies. The results seem to point out that younger and older cryptocurrencies have particular and different return and volatility behaviors detected by the clustering techniques.
- 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 remain showing a low similarity with the results in 2018 and no associations. Thus, the TADPole clustering seem to offer more unstable results, probably because it uses disaggregated data and results are more sensitive to small changes.

- The intersection of clusters seemingly inherit the association that we have observed separately for each one of the techniques. This is confirmed for Age, MkCap variables and the different financial ratios. As a result, clustering intersections characterize in a very comprehensive manner the main trends of the cryptocurrency market providing a manageable number of clusters with a multi-faceted characterization, and that 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.

Conclusion

In this work we analyzed the whole cryptocurrency market in 2018, that is all the cryptocurrencies traded in 2018, with a novel method that involved the combination of three different clustering algorithms. Each method used a meaningful representation considering different aggregation or granularity level of the daily returns: from the yearly average return and volatility, the yearly distribution of returns, and finally the observed time series of daily returns. For each representation we used prototype based clustering methods, so the prototypes of each cluster are meaningful and make possible to interpret the result.

Furthermore, we enhanced our profiling of the cryptocurrency market with association tests that validate the potential relationship between the clustering results and other descriptive features of the cryptocurrencies (technological attributes, financial ratios and age). These tests make possible to determine whether some features are related with a particular financial performance detected by the clustering algorithms.

Our analysis confirmed that there is an underlying structure of the data , which also persisted when considering a longer time period. Each one of the clustering algorithms helped to reveal different aspects of the cryptocurrency market. Furthermore, the combination of the different clustering results proved valid to detect the main trends in the cryptocurrency market, but also particular behaviors beyond these trends.

Finally, the association tests served to better describe the resulting clusters by adding the significant relationships found with the financial ratios, technological attributes and age of the cryptocurrencies.

In summary, we believe that the methodology used provides a consistent and descriptive tool supported by modern clustering techniques that may be useful for investors that need to understand the cryptocurrency market, as it reduces the dimensionality of the data set and identify the main trends in a descriptive manner. . For further investigations, the associations of some of the key financial ratios and cluster associations could play an important role enhancing the performance of the algorithms for the asset selection and diversification of portfolios (Brauneis and Mestel, 2019; Liu, 2019; Platanakis et al., 2018)) or improving the forecasting performance (Mallikarjuna and Rao, 2019) to tackle the difficulty of a new market.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
The initial idea was conceived by JA. The experiments were designed by LL. The searching on the data bases, statistical analysis and software design was performed by LL. The work was drafted by LL and revised critically by JA. All authors read and approved the final manuscript.

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

Funding
This work was supported by the European Union's H2020 Coordination and Support Actions under Grant 825215.

Author details
[1]Complutense University, Faculty of Statistical Studies, Madrid, Spain. [2]Complutense University, Faculty of Computer Science, Madrid, Spain. [3]Complutense University, Institute of Knowledge Technology, Madrid, Spain.

References
Acharya, J.G.A.: Cluster ensembles. WIRES Data Mining and Knowledge discovery **1**(4), 305–315 (2011)
Aggarwal, C., C., Reddy, K., C.: Data Clustering: Algorithms and Applications, 1st edn. Chapman & Hall/CRC, ??? (2013)
Aghabozorgi, S., Teh, Y.W.: Stock market co-movement assessment using a three-phase clustering method. Expert Systems with Applications **41**(4, Part 1), 1301–1314 (2014). doi:10.1016/j.eswa.2013.08.028
Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y.: Time-series clustering–a decade review. Information Systems **53**, 16–38 (2015)
Agresti, A.: An Introduction to Categorical Data Analysis. John Wiley & Sons, ??? (2018)
Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recognition **46**, 243–256 (2013)
Arroyo, J., Maté, C.: Forecasting histogram time series with k-nearest neighbours methods. International Journal of Forecasting **25**(1), 192–207 (2009). doi:10.1016/j.ijforecast.2008.07.003
Arroyo, J., González-Rivera, G., Maté, C., San Roque, A.M.: Smoothing methods for histogram-valued time series: an application to value-at-risk. Statistical Analysis and Data Mining: The ASA Data Science Journal **4**(2), 216–228 (2011). doi:10.1002/sam.10114
Bacon, C.R.: Practical Portfolio Performance Measurement and Attribution. The Wiley Finance Series. John Wiley & Sons, ??? (2008)
Baek, C., Elbeck, M.: Bitcoins as an investment or speculative vehicle? a first look. Applied Economics Letters **22**(1), 30–34 (2015)
Banerjee, A., Dave, R.N.: Validating clusters using the hopkins statistic. 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542) **1**, 149–153 (2004)
Begum, N., Ulanova, L., Wang, J., Keogh, E.: 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 (2015). doi:10.1145/2783258.2783286. https://doi.org/10.1145/2783258.2783286
Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: KDD Workshop, vol. 10, pp. 359–370 (1994). Seattle, WA
Bojanowski, M., Edwards, R.: alluvial: R Package for Creating Alluvial Diagrams. (2016). R package version: 0.1-2. https://github.com/mbojan/alluvial
Bonanno, G., Caldarelli, G., Lillo, F., S., M., Vandewalle, N., Mantegna, R.N.: Networks of equities in financial markets. The European Physical Journal B - Condensed Matter **38**(2), 363–371 (2004). doi:10.1140/epjb/e2004-00129-6
Brauneis, A., Mestel, R.: Cryptocurrency-portfolios in a mean-variance framework. Finance Research Letters **28**, 259–264 (2019)
Brida, J., Risso, W.: Dynamics and structure of the 30 largest north american companies. Society for Computational Economics **35**(1), 85–99 (2009)
Burniske, C., Tatar, J.: Cryptoassets: The Innovative Investor's Guide to Bitcoin and Beyond. McGraw-Hill Education, ??? (2017). https://books.google.es/books?id=-5AtDwAAQBAJ
Chan, S., Chu, J., Nadarajah, S., Osterrieder, J.: A statistical analysis of cryptocurrencies. Journal of Risk and Financial Management **10**(2), 12 (2017)
Chaudhuri, T.D., Ghosh, I.: Using clustering method to understand Indian stock market volatility. Communications on Applied Electronics **2**(6), 35–44 (2015)
Corbet, S., Meegan, A., Larkin, C.J., Lucey, B., Yarovaya, L.: Exploring the dynamic relationships between cryptocurrencies and other financial assets. Economic Letters -(165), 28–34 (2018)
Corbet, S., Lucey, B., Urquhart, A., Yarovaya, L.: Cryptocurrencies as a financial asset: A systematic analysis. International Review of Financial Analysis **62**, 182–199 (2019)
Drozdz, S., Gebarowski, R., Minati, L., Oswiecimka, P., Watorek, M.: Bitcoin market route to maturity? evidence from return fluctuations, temporal correlations and multiscaling effects. Chaos: An Interdisciplinary Journal of Nonlinear Science **28**(7), 071101 (2018). doi:10.1063/1.5036517
Drozdz, S., Minati, L., Oswiecimka, P., Stanuszek, M., Watorek, M.: Signatures of the crypto-currency market decoupling from the forex. Future Internet **11**(7), 154 (2019). doi:10.3390/fi11070154
Drozdz, S., Minati, L., Oswiecimka, P., Stanuszek, M., Watorek, M.: Competition of noise and collectivity in global cryptocurrency trading: Route to a self-contained market. Chaos: An Interdisciplinary Journal of Nonlinear Science **30**(2), 023122 (2020a). doi:10.1063/1.5139634. https://doi.org/10.1063/1.5139634
Drozdz, S., Minati, L., Oswiecimka, P., Stanuszek, M., Watorek, M.: Complexity in economic and social systems: cryptocurrency market at around covid-19. Entropy **22**(9), 1043 (2020b)

D'Urso, P., De Giovanni, L., Massari, R.: Garch-based robust clustering of time series. Fuzzy Sets Syst. **305**(C), 1–28 (2016). doi:10.1016/j.fss.2016.01.010
D'Urso, P., De Giovanni, L., Massari, R.: Trimmed fuzzy clustering of financial time series based on dynamic time warping. Annals of operations research, 1–17 (2019)
D'Urso, P., Cappelli, C., Di Lallo, D., Massari, R.: Clustering of financial time series. Physica A: Statistical Mechanics and its Applications **392**(9), 2114–2129 (2013)
D'Urso, P., Giovanni, L.D., Massari, R., D'Ecclesia, R.L., Maharaj, E.A.: Cepstral-based clustering of financial time series. Expert Systems with Applications **161**, 113705 (2020). doi:10.1016/j.eswa.2020.113705
Fisher, R.A.: On the interpretation on teh x2 from contingency tables and the calculation of the p. Royal Statistical Society **85**(1), 87–94 (1922)
González-Rivera, G., Arroyo, J.: Time series modeling of histogram-valued data: The daily histogram time series of s&p500 intradaily returns. International Journal of Forecasting **28**(1), 20–33 (2012). doi:10.1016/j.ijforecast.2011.02.007
Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1), 100–108 (1979)
Henning, C., Meila, M., Murtagh, F., Rocci, R.: Handbook of Cluster Analysis. CRC Press, ??? (2016)
Hornik, K.: A CLUE for CLUster Ensembles. Journal of Statistical Software **14**(12) (2005). doi:10.18637/jss.v014.i12
Hornik, K.: Clue: Cluster Ensembles. (2019). R package version 0.3-57. https://CRAN.R-project.org/package=clue
Hu, A.S., Parlour, C.A., Rajan, U.: Cryptocurrencies: Stylized facts on a new investible instrument. Financial Management **48**(4), 1049–1068 (2019)
Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**(1), 193–218 (1985)
Irpino, A., Verde, R.: Dynamic clustering of histograms using wasserstein metric. In: COMPSTAT 2006, Proceedings in Computational Statistics, pp. 869–876. Physica-Verlag, Heidelberg (2006)
Irpino, A.: HistDAWass Package: An R Tool for Histograms-values Data. (2016). R package version 1.0.4. https://cran.r-project.org/package=HistDAWass
Irpino, A., Verde, R.: Basic statistics for distributional symbolic variables: a new metric-based approach. Advances in Data Analysis and Classification **9**, 143–175 (2015)
Irpino, A., Verde, R., De Carvalho, F.d.A.T.: Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Systems with Applications **41**(7), 3351–3366 (2014). doi:10.1016/j.eswa.2013.12.001
Kern, M., Lex, A., Gehlenborg, N., Johnson, C.R.: Interactive visual exploration and refinement of cluster assignments. BMC bioinformatics **18**(1), 1–13 (2017)
Lemire, D.: Faster retrieval with a two-pass dynamic-time-warping lower bound. CoRR **abs/0811.3301** (2008). 0811.3301
Letra, I.J.S.: What drives cryptocurrency value? a volatility and predictability analysis. PhD thesis, Instituto Superior de Economia e Gestão (2016)
Liao, S.-H.: Mining stock category association and cluster on Taiwan stock market. Expert Systems with Applications **35**, 19–29 (2007)
Liao, S.-H., Chou, S.-Y.: Data mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio. Expert Systems with Applications **40**(5), 1542–1554 (2013). doi:10.1016/j.eswa.2012.08.075
Liao, T.W.: Clustering of time series data-a survey. The journal of the pattern recognition society **1**(38), 1857–1874 (2005)
Liu, W.: Portfolio diversification across cryptocurrencies. Finance Research Letters **29**, 200–205 (2019)
L'Yi, S., Ko, B., Shin, D., Cho, Y.-J., Lee, J., Kim, B., Seo, J.: XCluSim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. BMC bioinformatics **16**(S11), 5 (2015)
MacQueen, J.: Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Sympposium on Mathematical Statistics and Probability **1**(-), 281–297 (1967)
Mallikarjuna, M., Rao, R.P.: Evaluation of forecasting methods from selected stock market returns. Financial Innovation **5**(1), 1–16 (2019). doi:10.1186/s40854-019-0157-x
Mantegna, R.: Hierarchical structure in financial markets. European Physical Journal B **11**(1), 193–197 (1999)
Marti, G., Nielsen, F., Bi'nkowski, M., Donnat, P.: A review of two decades of correlations, hierarchies, networks and clustering in financial markets. Papers 1703.00485, arXiv.org (March 2017). https://arxiv.org/abs/1703.00485
Mehta, C.R., Patel, N.R.: A network algorithm for performing fisher exact test in rxc contingency table. Journal of the American Statistical Association **78**(382), 427–434 (1983)
Mehta, C.R., Patel, N.R.: Exact tests^{TM}. SPSS exact tests **7**, 12 (1996)
Mizuno, T., Takayasu, H., Takayasu, M.: Correlation networks among currencies. Physica A: Statistical Mechanics and its Applications **364**, 336–342 (2006). doi:10.1016/j.physa.2005.08.079
Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2009). http://www.bitcoin.org/bitcoin.pdf
Nanda, S.R., Mahanty, B., Tiwari, M.K.: Clustering Indian stock market data for portfolio management. Expert System with Applications **37**, 8793–8798 (2010)
Newman, M.: Power laws, pareto distributions and zipf's law. Contemporary Physics **46**(5), 323–351 (2005). doi:10.1080/00107510500052444. http://www.tandfonline.com/doi/pdf/10.1080/00107510500052444
Noirhomme-Fraiture, M., Brito, P.: Far beyond the classical data models: symbolic data analysis. Statistical Analysis and Data Mining: The ASA Data Science Journal **4**(2), 157–170 (2011). doi:10.1002/sam.10112
Onnela, J.-P., Chakraborti, A., Kaski, K., Kertész, J., Kanto, A.: Dynamics of market correlations: Taxonomy and portfolio analysis. Physical Review E **68**(5) (2003). doi:10.1103/physreve.68.056110
Pele, D., Wesselhöfft, N., Härdle, W., Kolossiatis, M., Yannis, Y.: A Statistical Classification of Cryptocurrencies (2020). https://ssrn.com/abstract=3548462

Peterson, B.G., Carl, P., Boudt, K., Bennet, R., Ulrich, J., Zivot, E., Lestel, M., Balkissoon, K., Wuertz, D.: PerformanceAnlytics: Econometric Tools for Performance and Risk Analysis. (2018). R package version 1.5.2. https://cran.r-project.org/package=PerformanceAnalytics

Platanakis, E., Sutcliffe, C., Urquhart, A.: Optimal vs naïve diversification in cryptocurrencies. Economics Letters **171**, 93–96 (2018)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013). R Foundation for Statistical Computing. http://www.R-project.org/

Rani, S., Sikka, G.: Recent techniques of clustering of time series data: a survey. International Journal of Computer Applications **52**(15) (2012)

Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **344**(6191), 1492–1496 (2014). doi:10.1126/science.1242072. https://science.sciencemag.org/content/344/6191/1492.full.pdf

Rosvall, M., Bergstrom, C.T.: Mapping change in large networks. PLoS ONE **5** (2010)

Sarda-Espinosa, A.: Dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance. (2019). R package version 5.5.6. https://CRAN.R-project.org/package=dtwclust

Sardá-Espinosa, A.: Time-series clustering in R using the dtwclust package. The R Journal **11**(1), 22–43 (2019)

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal **8**(1), 289–317 (2016)

Sigaki, H.Y.D., Perc, M., Ribeiro, H.V.: Clustering patterns in efficiency and the coming-of-age of the cryptocurrency market. In: Scientific Reports (2019)

Song, J.Y., Chang, W., Song, J.W.: Cluster analysis on the structure of the cryptocurrency market via Bitcoin-Ethereum filtering. Physica A-Statistical Mechanics and its Applications **527** (2019)

Stosic, D., Stosic, D., Ludermir, T.B., Stosic, T.: Collective behavior of cryptocurrency price changes. Physica A: Statistical Mechanics and its Applications **507**, 499–509 (2018)

Watorek, M., Drozdz, S., Kwapien, J., Minati, L., Oswiecimka, P., Stanuszek, M.: Multiscale characteristics of the emerging global cryptocurrency market. Physics Reports (2020). doi:10.1016/j.physrep.2020.10.005

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y., *et al.*: Top 10 algorithms in data mining. Knowledge and information systems **14**(1), 1–37 (2008)

Yates, F.: Tests of significance for 2 x 2 contingency tables. Royal Statistical Society **147**(3), 426–463 (1984)

Yermack, D.: Is bitcoin a real currency? an economic appraisal. Working Paper 19747, National Bureau of Economic Research (December 2013). doi:10.3386/w19747. http://www.nber.org/papers/w19747

Zhang, W., Wang, P., Li, X., Shen, D.: Some stylized facts of the cryptocurrency market. Applied Economics **50**(55), 5950–5965 (2018)

Zieba, Damian, Kokoszczyski, Ryszard, Sledziewska, Katarzyna: Shock transmission in the cryptocurrency market. is bitcoin the most influential? International Review of Financial Analysis **64**, 102–125 (2019)

Figures

Tables

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, *P70* for volume values lower than the P_{70} percentile, *P80* for values higher than the P_{70} and lower than the P_{90}, and similarly *P90*, *P99* and *P100*.
MkCap	5	Percentiles of the market capitalization. Namely, *P70* for market cap values lower than the P_{70} percentile, *P80* for values higher than the P_{70} and lower than the P_{90}, and similarly *P90*, *P99* and *P100*.
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	6	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). We use the same partition than in the M2 ratio.
HeavyTail	2	Binary variable that take value 1 if the cryptocurrency has a heavy-tail behaviour or 0 if it does not.

Table 1 Categorical variables used on the association tests and values

Additional material

Additional file 2 — Additional figures
Some additional figures

	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			

Table 2 Cluster cardinality, mean value and standard deviation of the centroid or prototypes for the clustering methods. For Hist-DAWass and TADPole we compute the mean and standard deviation of the prototypes.

	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

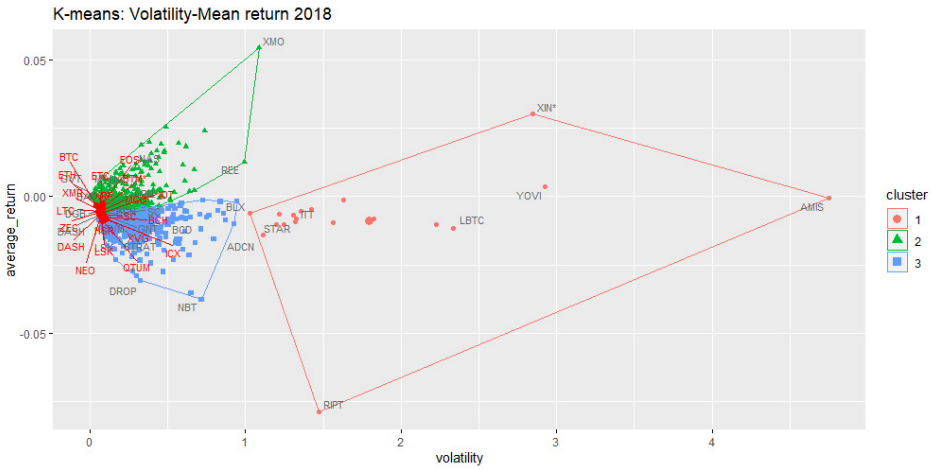
Table 3 Descriptive statistics for the prototypes of the Hist-DAWass clustering.

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

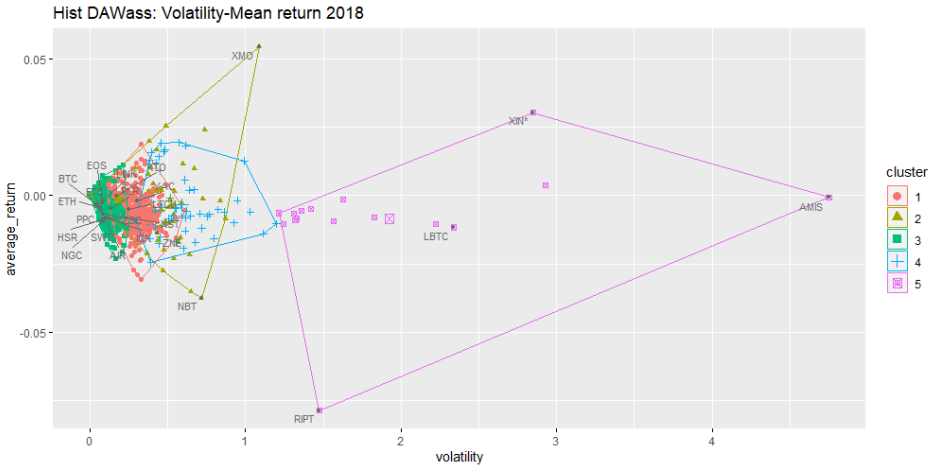
Table 4 Variability of TADPole clusters with the mean distance (Mean Dist.) to the centroid, standard deviation (Std. Dev.) and coefficient of variation (Coef. Var.).

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

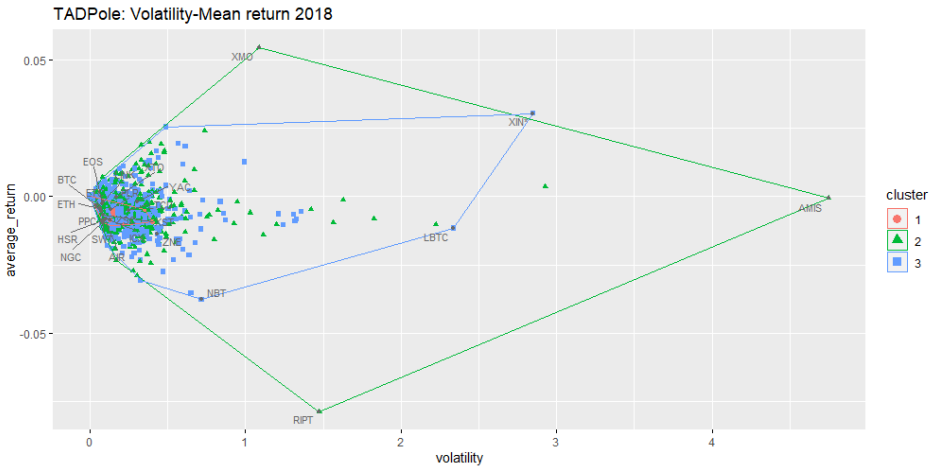
Table 5 Intersection of clusters across the different clustering algorithms, each column represent the cluster number. Intersections are sorted in inverse cardinality (N) order.



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

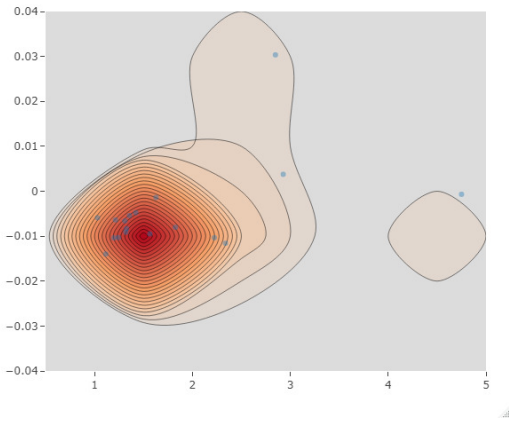


(b) Histogram DAWass clustering

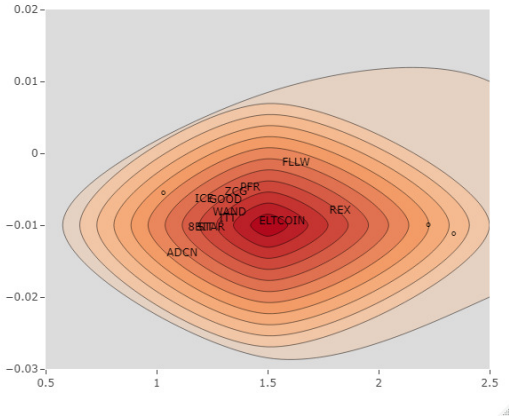


(c) TADPole clustering

Figure 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



(a) Cluster 1

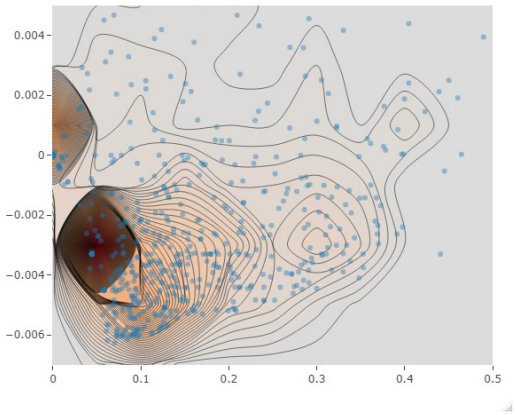


(b) Detail of Cluster 1

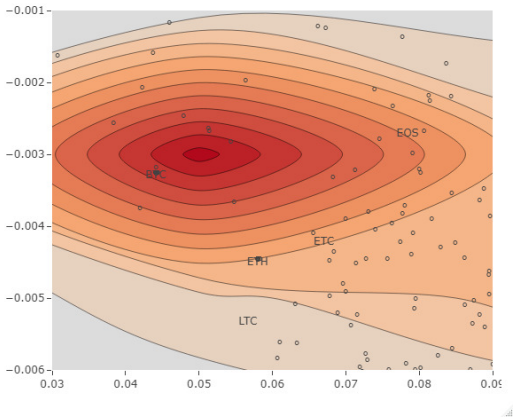
Figure 2 Cluster 1 represented in the bi-dimensional space $(\bar{r}, \sigma)$ by a 2D density contour plot.

		Volume				
Technique	Cluster/Intersection	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

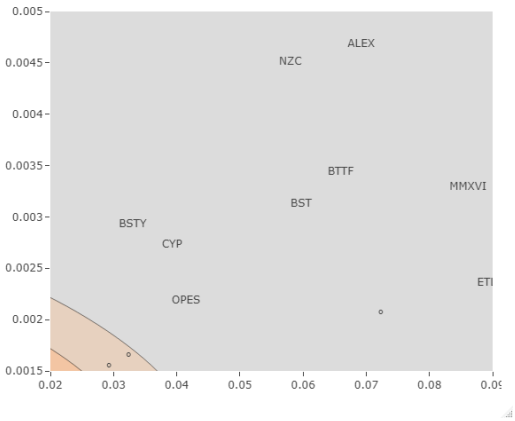
Table 6 *Volume* - Standardized Person's residuals



(a) Cluster 2

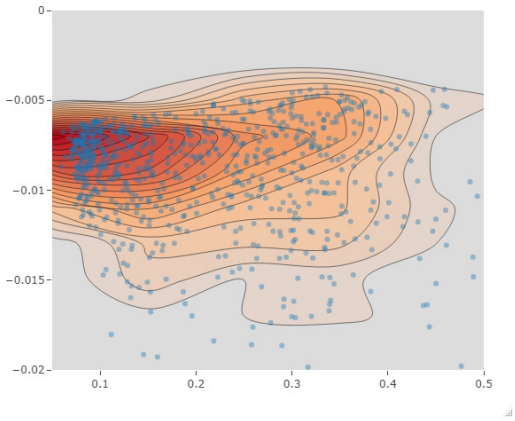


(b) Detail of Cluster 2 with the highest market cap cryptocurrencies

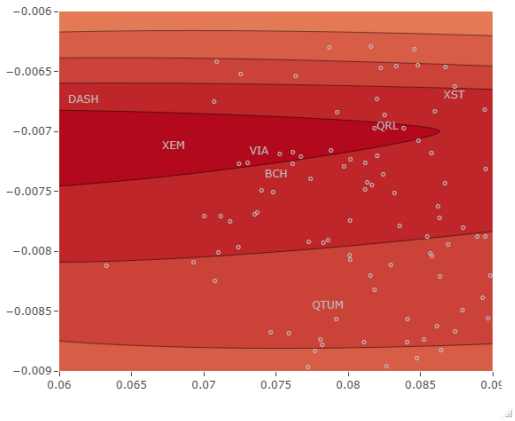


(c) Detail of cluster 2 for the high returns and low volatility cryptocurrencies

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



(a) Cluster 3

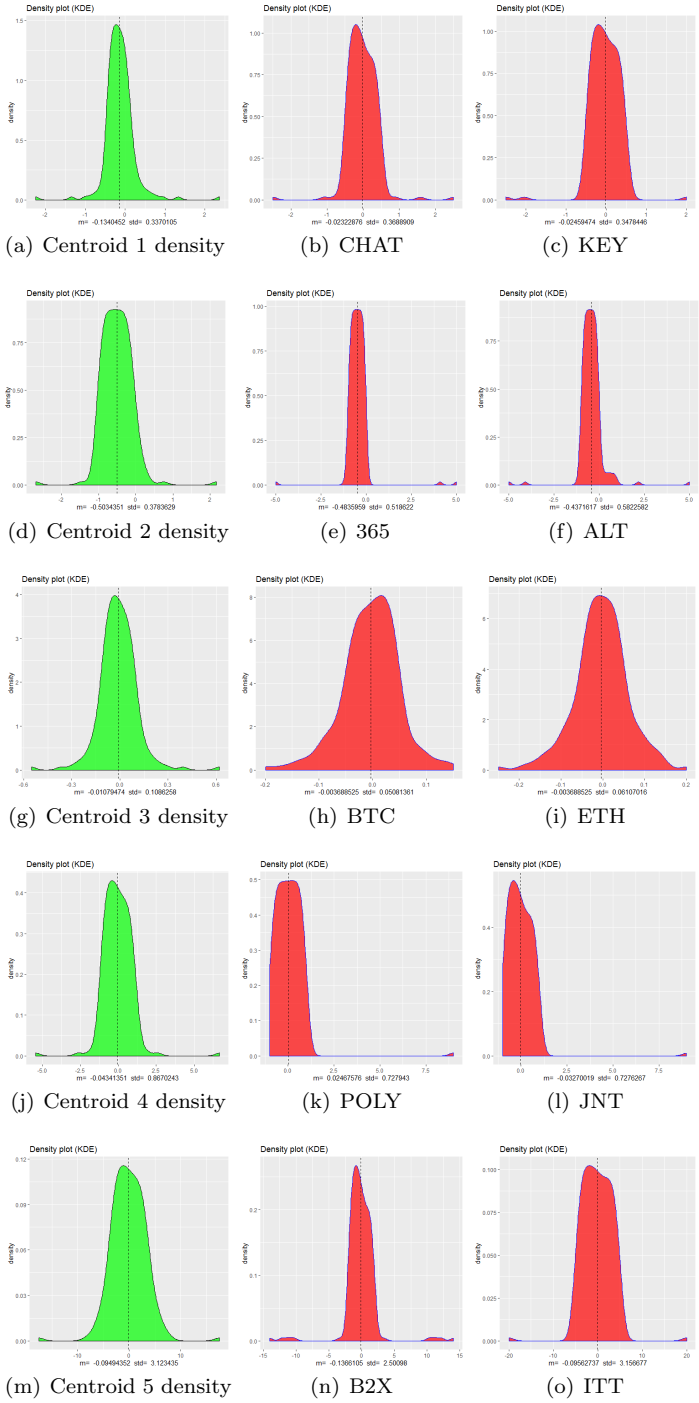


(b) Detail of Cluster 3 with the highest market cap cryptocurrencies)

Figure 4 Cluster 3 represented in the bi-dimensional space $(\bar{r}, \sigma)$ by a 2D density contour plot.

		Market cap (MKCap)				
Technique	Cluster/Intersection	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

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



(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

Figure 5 Density plot for prototypes (first column), and some representative cryptocurrencies of each cluster in terms of market capitalization (2nd and 3rd columns)

Figure 6 Medoids of the clustering results of the TADPole algorithm (time series of daily returns)

		Beta					
Technique	Cluster/Intersection	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	-

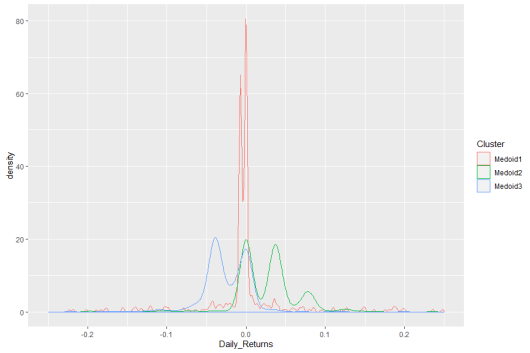
Table 8 *Beta* - Standardized Person's residuals

		Sharpe ratio		
Technique	Cluster/Intersection	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

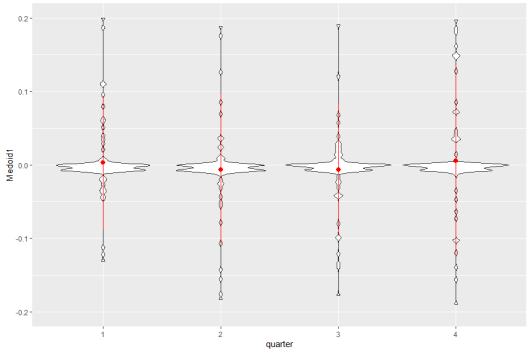
Table 9 *Sharpe ratio* - Standardized Person's residuals

	Market cap (MKCap)				
Encrypted algorithm (Algorithm)	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

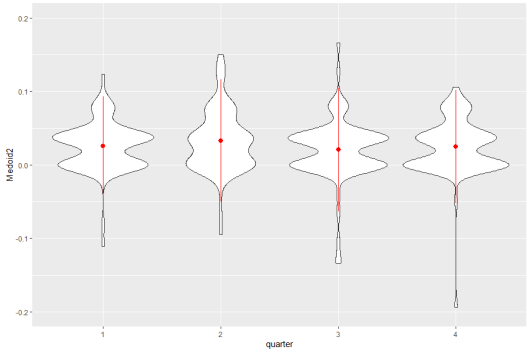
Table 10 Relevant associations between *Encrypted algorithm* - *Market cap* using the standardized Person's residuals



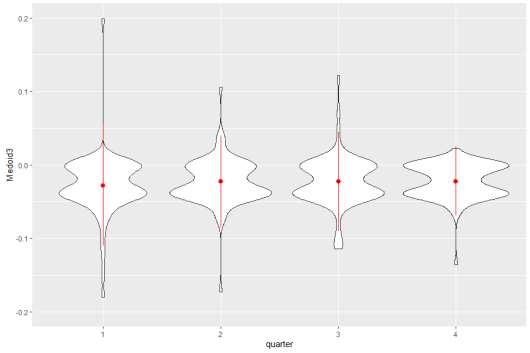
(a) Density plot of the three medoids



(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)

Figure 7 Density plot in Fig.a) and quaterly distribution of the TADPole medoids represented by violin graphs in Fig. b), c) and d) with the mean and standard deviation.

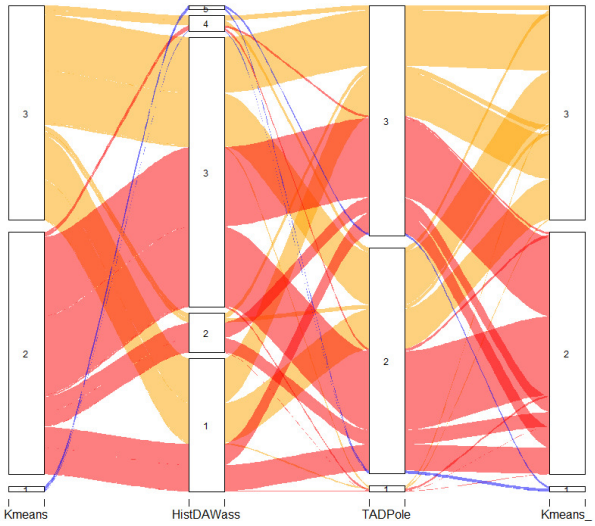


Figure 8 Alluvial plot showing the 'flows' of cryptocurrencies through the three clustering algorithms

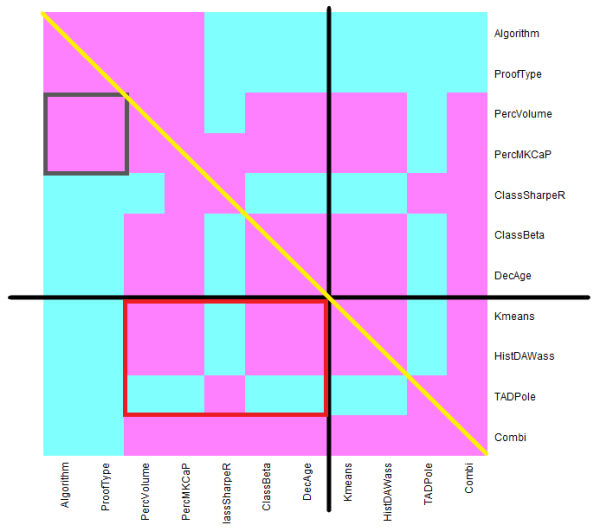


Figure 9 Matrix-type representation of the association tests in 2018 using the Fisher's exact test. Binary colored where pink color means significant association at p-values lower than 0.01. Red box for cluster-categorical variables and grey box focused on the particular associations with the technological variables. Yellow line represents the trivial association between a variable and its own

	Market cap (MKCap)				
Consensus algorithm (ProofType)	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

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

	Volume				
Consensus algorithm (ProofType)	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

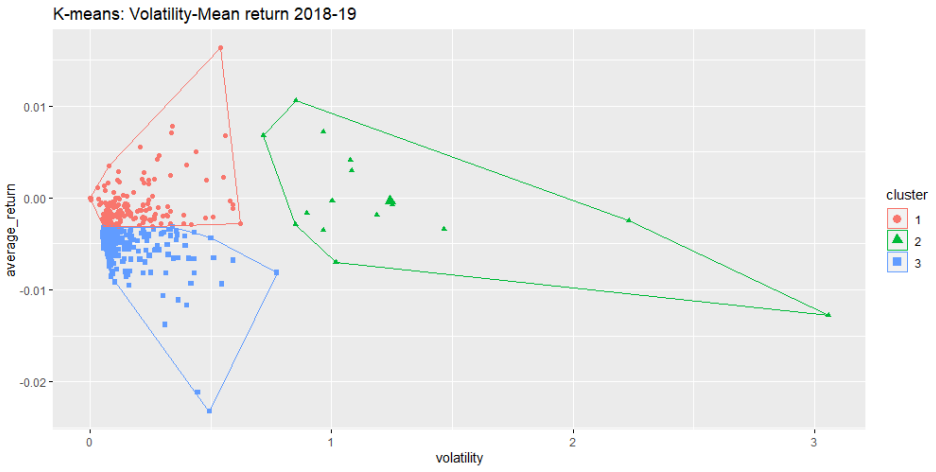
Table 12 Relevant associations between *Consensus algorithm* - *Volume* using the standardized Person's residuals

		Maturity						
Technique	Cluster/Intersection	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

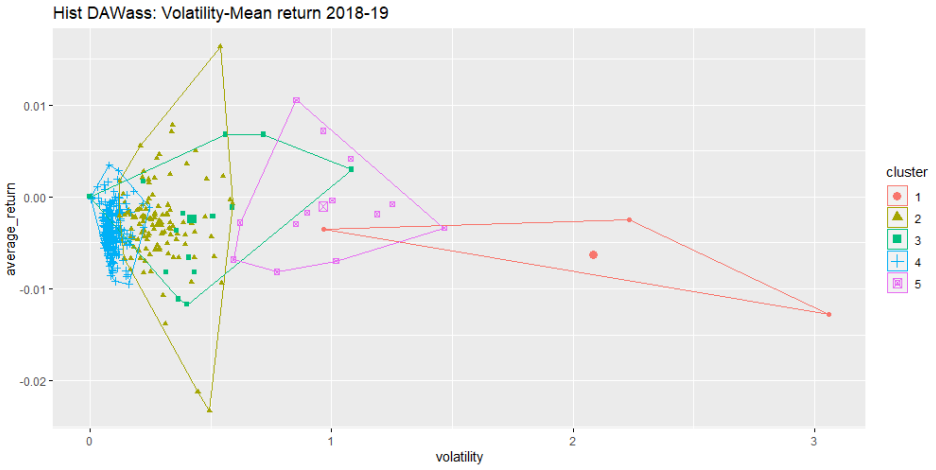
Table 13 *Maturity* - Standardized Person's residuals

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

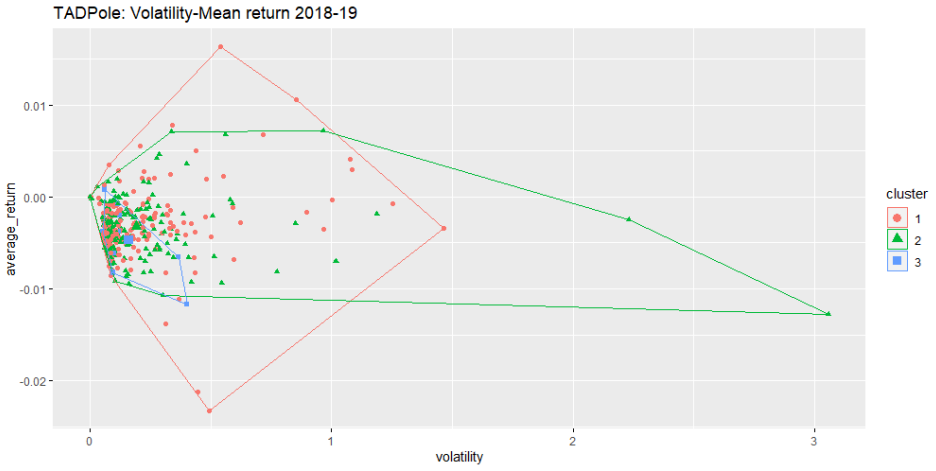
Table 14 Heavy-tail cryptocurrencies, Standardized Person's residuals for the association between the heavier tail distributions and clusters



(a) K-means clustering



(b) Histogram DAWass clustering



(c) TADPole clustering

Figure 10 Volatility-Average return plane in ordinary values for the different clustering techniques for 440 cryptocurrencies in 2018-19 time frame

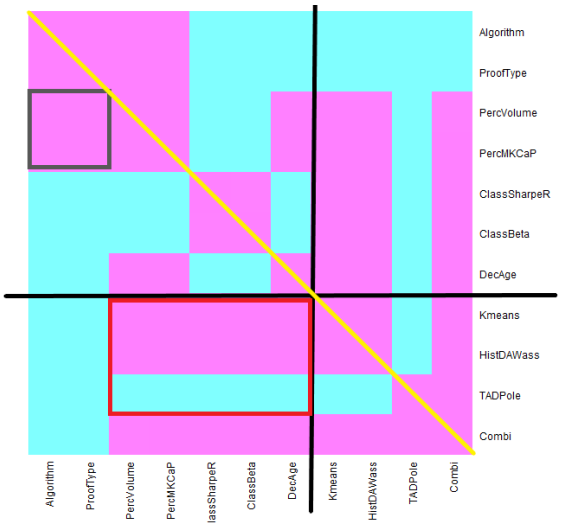


Figure 11 Matrix-type representation of the association tests in 2018-19 using the Fisher's exact test. Binary coloured where pink colour means significant association at p-values lower than 0.01. Red box for cluster-categorical variables and grey box focused on the particular associations with the technological variables. Yellow line marks the trivial maximum association for the same variables