Analysis of the cryptomarket applying different prototype-based clustering techniques

Luis Lorenzo & Javier Arroyo

Universidad Complutense de Madrid
Faculty of Statistic Studies
Faculty of Computer Sciences

luislore@ucm.es
javier.arroyo@fdi.ucm.es

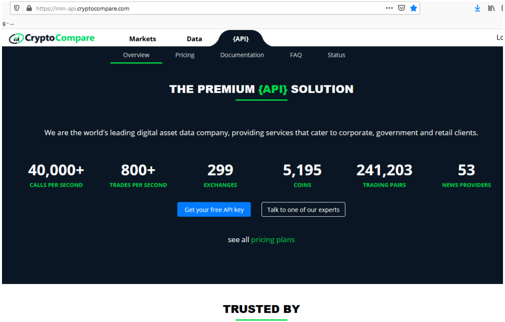
June 17, 2021

Overview

1. Motivation
   - What? Cryptomarket
   - Why? Goals
2. How? Methodology
3. Results
   - Clustering results
   - Persistence of the results
4. Conclusions

What? Cryptomarket

Since the appearance of Bitcoin, the cryptomarket has experienced an enormous growth not only in terms of capitalization but also in number of cryptos: over 5,000 cryptos with over 800 trades per second and almost 300 exchanges.
A **huge** and **heterogeneous** market difficult to manage and even understand for researches, investors and regulators.



Why? Goals

- A methodology that allow a quick exploration on the market and help us to **interprete** it
- Detection of trends and **underlying structures** that help us to **segment** and **organize** the market
- A methodology **scalable**, **intuitive** and **straightforward** for financial experts that require a **complementary view** of the market

Methodology

How? Methodology: Clustering

- Cluster analysis divides the dataset into groups (clusters) of cryptos with similar characteristics
- **Reduce the dimensionality of the problem**. We will use prototype-based clustering to have an object that describes each group
- Clustering provides a tool to describe the main **market trends**

How? Methodology: Clustering

We will use **prototype-based clustering** methods on **three different representations** of the cryptos:

- As the mean and standard deviation of the observed daily returns
- As a distribution of the observed daily returns
- As a time series of observed daily returns

We will use prototype-based clustering algorithms that work with such representations.

Clustering: k-means (1/3)

k-means

- One of the most widely used clustering algorithms
- The object to be clustered is a 2D representation with the standardized variables **yearly mean and standard deviation** of the daily log-returns (σ, μ)
- The representation is fairly typical in finance and it makes possible to easily display the whole cryptomarket in a 2D figure
- The distance considered is the *Euclidean distance*

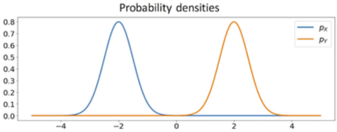
Clustering: HistDAWass (2/3)

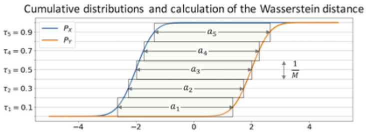
Dynamic Clustering for histogram data (Irpino and Verde, 2006)

- The object to be clustered will be a histogram representation of the observed distribution of daily log-returns
- The distance measure will be the l_2 *Wasserstein distance* that has a strong and intuitive meaning, as it can be decomposed as the addition of three elements: the histogram differences in terms of location, spread and shape.

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

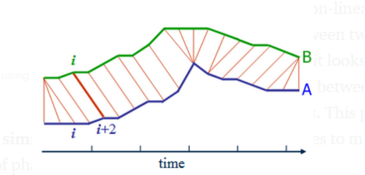
with h_1,h_2 as histograms.





Density based time series clustering (Begum, N,. et al., 2015)

- The object to be clustered will be the observed time series of daily log-returns
- TADPole is a **shape-based algorithm** that apply *Dynamic Time Warping* (DTW) as similarity measure which is considered one of the most popular and more accuarate than Euclidean measures.
- TADPole algorithm **speed-up the convergence** of clustering based on DTW distance implemented



Methodology: Association tests

Fisher's exact tests and Pearson's residual

- Association tests will help us to assess the relationship between the clustering results and some categorical variables not considered by the clustering algorithms
- The categorical variables analyzed will include: technological aspects of the crypto, the financial performance of the crypto, age of the crypto...
- We will apply the exact **Fisher's tests** and analyze the **Standardized Pearson's residuals** of the contingency tables ($r_{adj} > 3.5$ for significant associations)
- Fisher's exact tests is still valid when sample size is small

Methodology: data set

Data set

All cryptos traded during 95% of the time along 2018 (extended for comparison reasons up to 2019): 1,723 cryptos.

$$r_i(t) = ln(P_i(t)) - ln(P_i(t-1)) \tag{2}$$

Methodology: variables

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	4	Sharpe ratio divided into the following categories: *SRF* (Small Risk-free) for negative values *ERP* (Excess return positive) for positive values lower than 0.5 *ACC* (Acceptable) for values equal to or higher than 0.5 and lower than 1.0 *GOOD* for values equal to or higher than 1.0
Age	7	Deciles of the age variable (time on the market). 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

For the association tests, we will use heavy-tail significance tests for the existence of 1st and 2nd order moments (we will fiter out the cryptos that do not pass the test)

Results

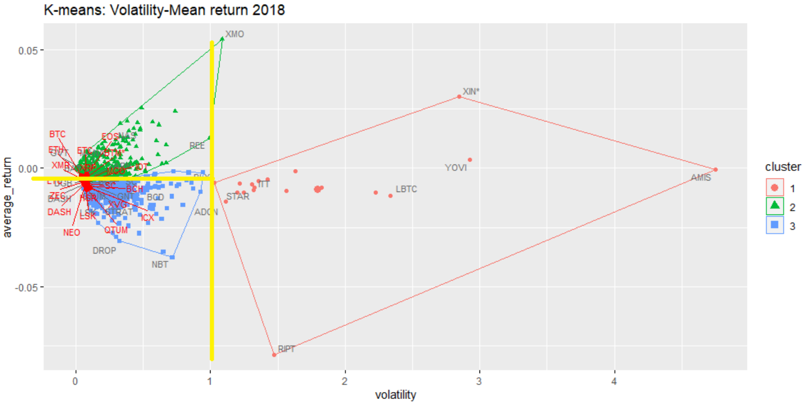
Clustering results

	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.

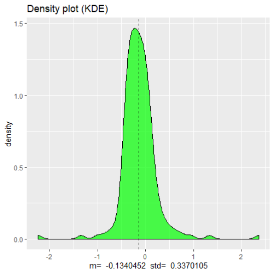
- Low number of clusters for all the methods
- All the methods identify 2 or 3 very big clusters, and 1 or 3 small ones
- According to the Adjusted Rand index, there is no agreement between the results of the three clustering methods. Thus, each method offer a complementary view on the market

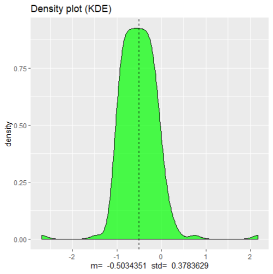
k-Means on the mean-volatility representation

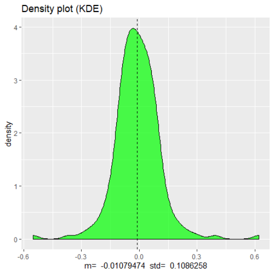


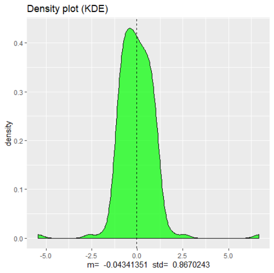
Dynamic clustering on the distribution of log returns

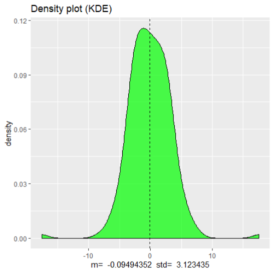
Density plots of the five prototypes











Dynamic clustering on the distribution of log returns

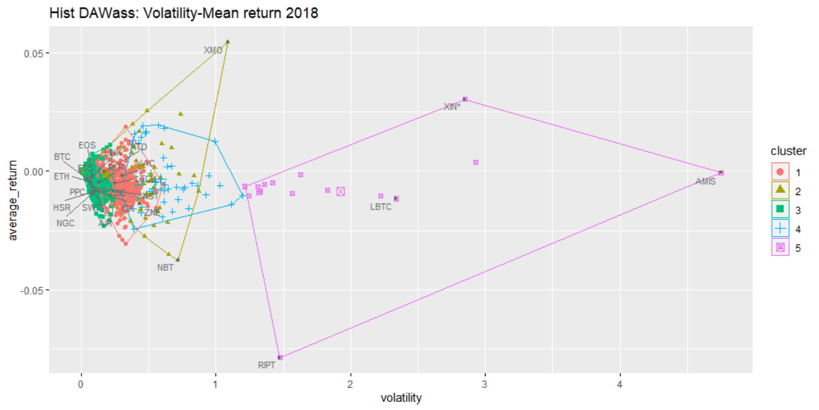
	Mean	Std. Dev.	Coef.Var.	Skew.	Kurt.	Med.	Min.	Max.	Var.Wass.
Clus. 1	-0.13	0.34	-2.51	**0.82**	**13.43**	-0.16	-2.24	2.36	0.025
Clus. 2	**-0.50**	0.38	**-0.75**	0.56	9.33	-0.51	-2.69	2.18	0.079
Clus. 3	**-0.01**	0.11	-10.06	0.28	7.10	-0.01	-0.55	0.62	0.005
Clus. 4	-0.04	0.87	-19.97	0.54	11.95	-0.08	-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.

Good separation of the prototypes by the coefficient of variation (σ/μ)

Dynamic clustering on the distribution of log returnsn

Results projection in the volatility-mean return plane



TADPole clustering on the time-series of log returns

TADPole		
Card.	*Mean*	*Std.Dev.*
22	-0.001	0.080
843	0.026	0.046
858	-0.028	0.047

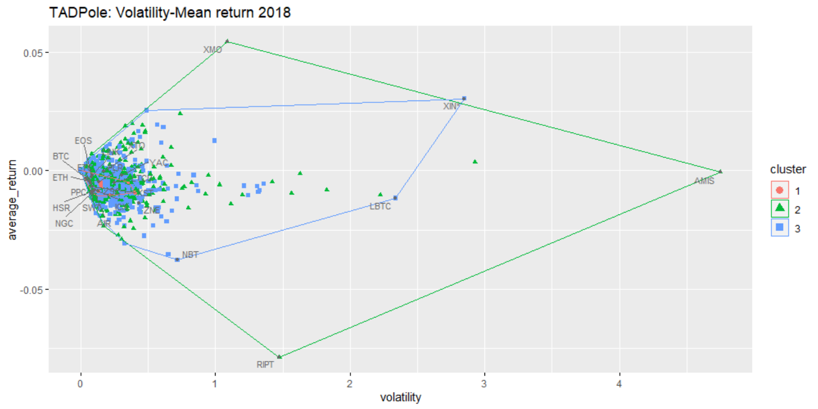
- The TADPOLE clustering identifies three clusters taking into account the time series **shapes** and **dispersion** over time.
- The centroids properly identify time series with mean log-returns below, over and around zero and with different standard deviations

TADPole clustering on the time-series of log returns

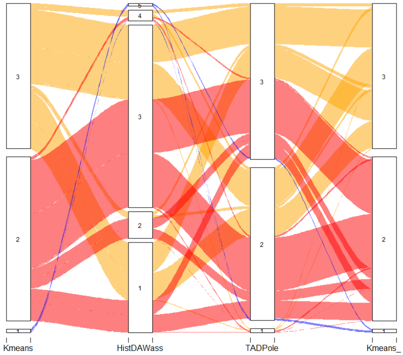
Time series prototypes

TADPole clustering on the time-series of log returns

Results projection in the volatility-mean return plane



Intersecction of the clustering results



- The alluvial plot makes possible to see the main trends of the market
- It is possible to find also minoritary, but interesting trends

Intersecction of the clustering results

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.

- Only 24 out of 45 possibles are populated
- The main 6 intersections (in number of cryptos) represent the 75% of the total market

Association tests

We run the tests only for the 1,262 cryptos which ensure the existence of first $\mathbb{E}\{x\}$ and second statistical $\mathbb{E}\{x^2\}$ moments.

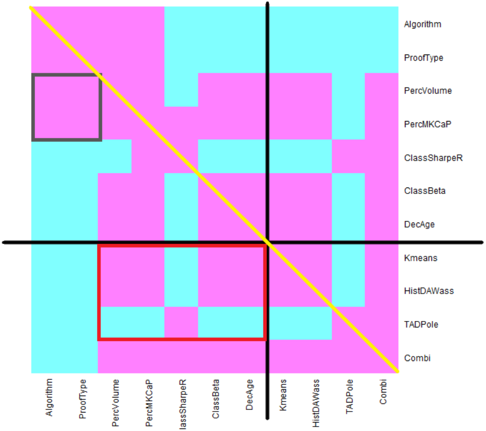


Figure: Fisher exact tests (p-values $\leq$ 0.01 in pink colour)

Results of the association tests

Standardized Pearson's residual: Volume

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

Table 6 *Volume* - Standardized Person's residuals

Standardized Pearson's residual: Market cap

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

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

Results of the association tests

Standardized Pearson's residual: Beta

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

Table 8 *Beta* - Standardized Person's residuals

Standardized Pearson's residual: Sharpe ratio

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

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

Results of the association tests

Standardized Pearson's residual: Maturity

Technique	Cluster/Intersection	D4	D5	D6	D7	D8	D9	D10
		Maturity						
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

Standardized Pearson's residual: Heavy-Tails

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

Persistence of the results

- We replicated the results for the 440 cryptos that traded both in 2018 and 2019 (730 days)
- The number and the shapes of the clusters in 2019 is quite similar to that in 2018
- The ARI index shows high agreement (¿0.3) for the mean-standard deviation and for the distributions, but null for the time series (It makes sense!)
- The association tests are also quite similar.

Conclusions

Conclusions

Take-aways

- High degree of homogeneity on the clusters despite of the high number of cryptos
- Significant associations detected between the clustering results and Volume, Market cap and financial ratios such as Beta and Sharpe
- The blockchain implementation could have an impact on the financial behaviour of the crypto.
- Younger and older cryptos have a particular and different financial behaviours detected by the clustering techniques.
- The TIME SERIES CLUSTERING shows a more unstable results, probably because the temporal similarities are difficult to hold.

References

Lorenzo, L., Arroyo, J. (2021)
Analysis of cryptomarket applying different prototype-based clustering techniques
Preprint, https://eprints.ucm.es/id/eprint/63821/

Mantegna, R. (1999)
Hierarchical structures in financial markets
European Physical Journal B,11(1), 193-197 (1999).

Aghabozorgi, S., (2014)
Time-series clustering -a decade review
Information Systems,53, 16-38 (2015).

Irpino, A., Verde, R. (2006)
Dynamic clustering of histogram using wasserstein metric
Proceedings in Computational Statistics,Physical-Verlag, 860-876 (2006).

Begum, N., et al. (2015)
Accelerating dynamic time warping clustering with novel admissible pruning strategy
Proceeding of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data mining,Association for Computing Machinery, New York, 49-58