REM WORKING PAPER SERIES

Reconstructing cryptocurrency processes via Markov chains

Tanya Araújo, Paulo Barbosa

REM Working Paper 0262-2023

February 2023

REM – Research in Economics and Mathematics
Rua Miguel Lúpi 20,
1249-078 Lisboa,
Portugal

ISSN 2184-108X

U LISBOA | UNIVERSIDADE DE LISBOA

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

REM RESEARCH IN ECONOMICS AND MATHEMATICS

REM – Research in Economics and Mathematics

Rua Miguel Lupi, 20
1249-078 LISBOA
Portugal

Telephone: +351 - 213 925 912
E-mail: rem@iseg.ulisboa.pt

https://rem.rc.iseg.ulisboa.pt/

https://twitter.com/ResearchRem

https://www.linkedin.com/company/researchrem/

https://www.facebook.com/researchrem/

Reconstructing cryptocurrency processes via Markov chains

Tanya Araújo[1,2*] and Paulo Barbosa[2]

[1*]Research Unit on Complexity and Economics, UECE-REM, R. Miguel Lupi, 20, Lisbon, Portugal.
[2]ISEG, University of Lisbon, R. Quelhas 6, Lisbon, 1200-781, Portugal.

*Corresponding author(s). E-mail(s): tanya@iseg.ulisboa.pt;
Contributing authors: paulofbarbosa5@gmail.com;

Abstract

The growing attention on cryptocurrencies has led to increasing research on digital stock markets. Approaches and tools usually applied to characterize standard stocks have been applied to the digital ones. Among these tools is the identification of processes of market fluctuations. Being interesting stochastic processes, the usual statistical methods are appropriate tools to their reconstruction. There, besides chance, the description of a behavioural component shall be present whenever a determinist pattern is ever found. Markov approaches are at the leading edge of this endeavour. In this paper, Markov chains of orders one to eight are considered as a way to forecast the dynamics of three major cryptocurrencies. It is accomplished using an empirical basis of intra-day returns. Besides forecasting, we investigate the existence of eventual long-memory components in each of those stochastic process. Results show that the average predictions obtained from using the empirical probabilities is better than random choices.

Keywords: Markov chains, Criptocurrency, Forecasting, Market Processes

JEL Classification: D8 , H51

1 Introduction

Since Bitcoin was introduced by Nakamoto, 2008, cryptocurrencies have received great attention from monetary authorities, firms and investors (John et al., 2022). Among the reasons for all this attention are the possibility of reducing risk management, improving portfolios and analysing consumer sentiment (Dyhrberg, 2016). Indeed, few financial innovations have drawn similar attention to regulators, investors and stakeholders.

Attention naturally led to the characterisation of some stylized facts of the digital market. Reference Cunha and Silva, 2020 shows that stylized facts of Bitcoin (BTC) data are similar to those of traditional financial assets. Namely, distributions of BTC one-day returns display fat tails, exhibits volatility clustering, and the correlation between its volume and volatility happens to be always positive. Similar results were found in reference Urquhart, 2017 where moments of high volumes were shown to correspond to higher risks.

Simultaneously, reference Bariviera, et al., 2017 focus on the study of BTC long-range memory of both daily and intra-day prices. There, it was shown that BTC data displays high volatility, long-range memory unrelated to market liquidity, and intra-day prices similar across different time scales (5 to 12 hours). Moreover, persistent behavior of daily prices from 2011 to 2014 were captured by the calculation of the Hurst exponent, showing that, after 2014, it decreased to near 0.5 as in a random process.

Likewise, references Dyhrberg, 2016 and Baur and Lucey, 2010 show that BTC has similar hedging capabilities to gold against the US dollar and the Stock Exchange Index. In the former, a threshold Garch model was applied to one-day BTC returns. In the latter, the investigation relies on hedging capabilities, concluding that this virtual currency can be used alongside gold to eliminate or minimize specific market risks. In addition, as BTC can be traded continuously, reference Dyhrberg, 2016 argues that this virtual currency has specific speed advantages and can be added to the list of hedging tools.

Stochastic partial differential equation is applied by reference Cheah and Fry, 2015 to model BTC behaviour. There, the authors argue that BTC behaves like a traditional asset being BTC prices dominated by highly speculative periods. Therefore, they conclude that the cryptocurrency market show great similarity, in terms of stylized empirical facts, when compared to traditional markets, and, more precisely, in what concerns vulnerability to speculative bubbles.

The study reported by reference Corbet et al., 2018 focus on the investigation of the stochastic properties of six major cryptocurrencies and their linkages with six standard stock market indices. The main findings show that the behaviour of cryptocurrency markets are highly connected to each other but decoupled from the main standard stock indexes. Therefore, digital stocks are seen as an important contribution since they offer diversification benefits to investors.

In the present paper, in order to estimate the behavior of some major cryptocurrencies, the processes of one-hour returns of BTC, Ethereum (ETH) and

Ripple (XRP) are reconstructed. It is done by using Markov chains of orders one to eight. The reconstruction of those cryptocurrencies processes starts with the identification of: *(i)* the allowed (markovian) transitions in the state space that corresponds to current orbits of the system, and *(ii)* the occurrence frequency of each orbit in typical samples. It is accomplished by taking the first half of each sample to the computation of the conditional probabilities of the allowed transitions. From this empirical base, each second half is estimated and compared with a random choice. Results show that the average predictions obtained from the empirical transitions probabilities outperform random forecasts.

The structure of this paper is as follows: the next section describes the methodology, Section 3 presents and discuss the results, while Section 4 presents some concluding remarks.

2 Data and Methods

The data used in the present study is sourced from coinmarketcap.com, where historical information about over one thousand cryptocurrencies is available. Three major cryptocurrencies according to their market capitalization: BTC, ETH and XRP were considered. Fig. 1 shows the behavior of hourly-price data for those three digital stocks.

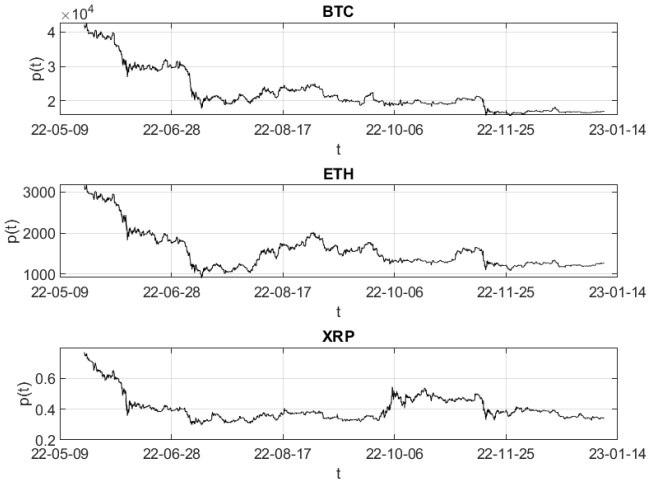

Fig. 1 Hourly-prices of the three cryptocurrencies, from May-20-2022 to Jan-08-2023.

The three plots in Fig.1 present the series of prices of BTC, ETH and XRP, from May-20-2022 to Jan-08-2023, comprising each of them 6306 observations.

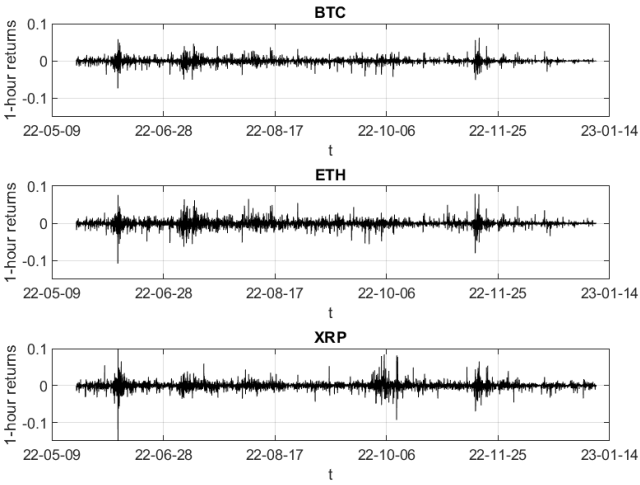

Fig. 2 Hourly-returns of the cryptocurrencies, from 2022-05-20 to 2023-01-08

From the plots in Fig.1, it is quite apparent that the data is very far from stationary, but a different situation comes out when, instead of prices, we considered the series of one-hour returns.

$$r_t = log(p_t) - log(p_{t-1}). \tag{1}$$

The three plots in Fig.2, show the series the one-hour returns, while the plots in Fig.3 illustrate their dynamics.

$$r\,(t,1) \rightarrow r\,(t+1,1) \tag{2}$$

These last three plots (Fig.3) show that, in the three cases, the bulk of the data consists of a central core of small fluctuations with a few large flights away from the core. This structure of the data will have influence on the results obtained in the next section.

2.1 Markov Chains

There is a huge literature applying Markov Chains to model the behaviour of financial time series. This approach is a fundamental tool in the study of stochastic processes. It has been used widely in many different disciplines, like weather, epidemic, land use, consumer behavior and even for identification of writers (Khmelev and Tweedie, 2001).

A sequence of random variables $Z_1, Z_2, \ldots, Z_t, \ldots$ with Markov characteristic is known as a process with first order dependence (as described in Equation 3, this process has no memory). The Markov characteristics means that the distribution of the future realization of Z_{n+1} depends on its immediately previous state (Z_n) and not on further previous states $(Z_{n-1}, Z_{n-2}, \ldots)$. Formally,

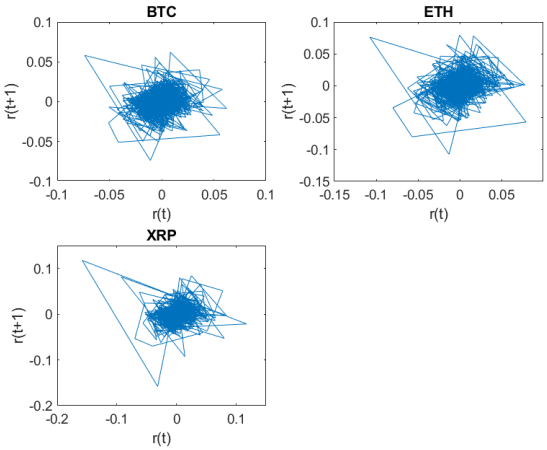

Fig. 3 The dynamics of hourly-returns of the three cryptocurrencies

$$\begin{aligned} Pr(Z_{n+1} = z_{n+1} \mid Z_1 = z_1, \ldots, Z_n = z_n) \\ = Pr(Z_{n+1} = z_{n+1} \mid Z_n = z_n) \end{aligned} \tag{3}$$

In a dynamical system, shifting from state i to state j has transition probability p_{ij}

$$p_{ij} = Pr(Z_1 = s_j \mid Z_0 = s_i) \tag{4}$$

Fortunately, Markov chains can be approached from a higher order perspective, being Markov chains of higher orders the processes in which the next state depends on two or more preceding ones. Here, Markov chains of orders one to eight are considered as a way to predict cryptocurrency returns and to investigate the existence of eventual long-memory components in that stochastic process.

2.2 Coding

We consider the dynamical system being coded by a finite alphabet Σ. Then, Ω, the space of orbits of the system are comprised of infinite sequences $\omega = i_1 i_2 \cdots i_k \cdots$, $i_k \in \Sigma$, with the dynamical law being a shift σ on these symbol sequences.

$$\sigma\omega = i_2 \cdots i_k \cdots \tag{5}$$

Depending on the dynamical law of the coded system, not all sequences will be allowed. The set of allowed sequences in Ω defines the *grammar* of the shift. The set of all sequences which coincide on the first n symbols is called a $n-$block

and is denoted $[i_1 i_2 \cdots i_n]$. The probability measures over the $n-$blocks constitute part of information that may be inferred from the data, being the main tool used to characterize the dynamical properties of the dynamical system.

To calculate the probability measures over the $n-$blocks the following computation is performed: for each series of one-hour returns (r_t) with mean $\hat{r}$ and standard deviation s, a five-symbols code Σ is defined.

$$\Sigma = \{-2, -1, 0, 1, 2\} \tag{6}$$

Then,

$$\begin{array}{rcl} \left(r\left(t\right) - \overline{r\left(t\right)}\right) > s & \Longleftrightarrow & 2 \\ s \geq \left(r\left(t\right) - \overline{r\left(t\right)}\right) > \frac{s}{3} & \Longleftrightarrow & 1 \\ \frac{s}{3} \geq \left(r\left(t\right) - \overline{r\left(t\right)}\right) > -\frac{s}{3} & \Longleftrightarrow & 0 \\ -\frac{s}{3} \geq \sigma\left(r\left(t\right) - \overline{r\left(t\right)}\right) > -s & \Longleftrightarrow & -1 \\ -s \geq \left(r\left(t\right) - \overline{r\left(t\right)}\right) & \Longleftrightarrow & -2 \end{array} \tag{7}$$

This coding is used and the empirical frequencies $\widetilde{\mu}\left([i_1 \cdots i_k]\right)$ for blocks of successively larger order k are found. Naturally, k cannot be arbitrarily large because of statistics. The reliability of results is threatened whenever 5^k is larger than the size N of the data sample. Therefore, statistical reliability may be directly tested by comparing the number of different occurring blocks and 5^k.

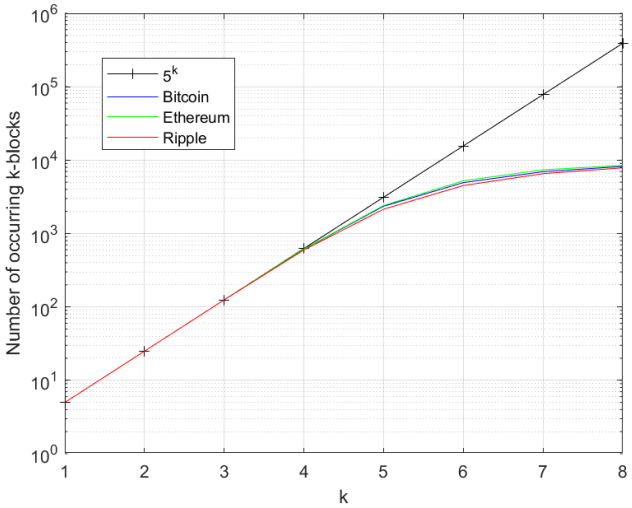

Fig. 4 Comparing the number of different occurring blocks and 5^k

Fig.4 shows the evolution of the number of occurring $k-$blocks and 5^k, where the number $p\left(k\right)$ of occurring blocks of size k in each data sample is

compared with the maximum possible number, 5^k. In all cases, after $k = 4$ the comparison shows the lack of statistics apparent in the comparison of $p\,(k)$ with 5^k. These results seem to suggest that the data is described by a short range memory.

2.3 Predicting

Prediction starts by taking the time series of each cryptocurrency and spliting the series into two halves, the first half is then used to predict the other half. The coding procedure presented in the last section is used and after the first half of each sample is coded into $\Sigma = \{-2, -1, 0, 1, 2\}$, we perform the computation of the conditional probabilities of the allowed transitions on these symbol sequences. The conditional probabilities are computed for blocks of successively larger order (k).

As an example, the conditional probabilities computed for blocks of size two are defined in the following Markov transition matrix.

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \\ p_{41} & p_{42} & p_{43} & p_{44} & p_{45} \\ p_{51} & p_{52} & p_{53} & p_{54} & p_{55} \end{bmatrix}$$

where each p_{ij} indicates the probability of shifting from state i to state j, as in Equation 3.

The transition matrices for blocks of orders up to eight are computed. From this empirical base, each second half is estimated and compared with a random number chosen at random.

However, when the conditional probabilities are inferred from limited experimental data an extended Markov approximation is more convenient, we used the Less-than-k-Markov approach as presented in reference Vilela et al., 2002.

2.3.1 The Less-than-*k*-Markov approach

In each simulation, with an approximation of order k, we look at the current block $(a_1 \cdots a_k)$ of order k and use the $k-$empirical probability to infer the next state a_0. If that block is not present in the data that was used to construct the empirical probabilities, then we look at the $k-1$ sized block $a_1 \cdots a_{k-1}$ and use the $k-1$ order empirical probabilities. If required, the process is repeated until an available empirical probability given by a $(k-i)-$order block with $2 < i < 8$ is found.

Such an extended Markov approach is applied to each $k-$order block in order to estimate the successor a_0 of each block $(a_1 \cdots a_k)$. In so doing, the successor a_0 is compared with a prediction $\widetilde{a}_0$ obtained by throwing a random number with the empirical probabilities $\widetilde{P}\,(a_0[a_1 \cdots a_k)$.

The Past predicting the future

Once the empirical probabilities are computed from the first half of each series ($t = 1, 2, ...n$) the second half is visited ($t = n + 1, n + 2, ...2n$) in order to quantify the magnitude of the error found when using each $k-$sized block, the quantity $e_k(t)$ is computed for each sample: `BTC`, `ETH` and `XRP`.

$$e_k(t) = [\widetilde{a}_0 - a_0(t)] \tag{8}$$

As half of each series comprises n observations, the averaged error for each $k - block$ is calculated

$$e_k = \frac{1}{n} \sum_{t=n+1}^{2n} [\widetilde{a}_0 - a_0(t)] \tag{9}$$

The average error of the forecast of the second half of each sample is therefore computed as the distance between the observed a_0 and the corresponding estimated state $\widetilde{a}_0$.

The same procedure is performed with the successor a_0 of each $k-$order block being estimated at random ($\widetilde{r_0}$). There, the error is given by the distance

$$eRand_k = \frac{1}{n} \sum_{t=n+1}^{2n} [\widetilde{r_0} - a_0(t)] \tag{10}$$

In the end, the errors e_k and $eRand_k$ are averaged over 50 different runs.

2.4 Method outline

In the following, a brief and summarized description of the algorithm used in the simulations is presented. The final results contain average values over 50 runs for each cryptocurrency.

3 Results and Discussion

The first three plots in Fig. 5, show the average error obtained with a 5-symbols alphabet for the three cryptocurrencies. The last plot shows the error obtained for each cryptocurrency and computed when the prediction is performed at random, i.e., from a surrogate matrix of probabilities.

These results are similar to those obtained by Vilela et al., 2002 where daily returns of three standard stocks and the NYSE index were analysed. Not surprisingly, in all cases, the average prediction obtained from using the empirical probabilities is better than a random choice. However, here, `ETH` and `XRP` data show even higher improvements coming from the four and five-symbol blocks, respectively.

The stocks `BTC` and `ETH` seem to share closer similarities than `XRP`. However, the one-symbol probabilities are similar in the three digital stocks. This suggests that the statistical short-memory component of the market process might

Outline of the algorithm

Take each cryptocurrency series of hourly-prices: `BTC`, `ETH` or `XRP`

1. Compute hourly-returns: $r_t = log(p_t) - log(p_{t-1})$ from the series of prices p_t
2. Split the series of returns into two halves of the same size n: H_1 and H_2
3. Code the two halves $H1$ and $H2$ into $\Sigma = \{-2, -1, 0, 1, 2\}$
4. Perform the steps (4.1 to 4.5) along $j = 1, 2, ..., 50$ simulations:
   For each $k - block$ successively larger $k = 2, ..., 8$
   4.1 Build the conditional probabilities $\widetilde{P}[(a_0[a_1 \cdots a_k])$
   4.2 Look at the k sized block $a_1 \cdots a_k$ and use the k order empirical probabilities to infer each next state a_0 with
   $\Diamond$ If the block $a_1 \cdots a_k$ is not found, repeat using blocks of size $(k - \tau)$ until the available empirical probability is found
   4.3 Visit each $a_0(t) \in H2$, $t = n + 1, n + 2, ...2n$
   4.4 Measure the error $e_k(j) = \frac{1}{n}(\sum_{t=n+1}^{2n}[a_0(t) - \widetilde{a}_0])$
   4.5 Compute $eRand_k(j) = \frac{1}{n}(\sum_{t=n+1}^{2n}[a_0(t) - r_0]\ )$ being r_0 chosen at random
5. Compute the average errors e_k and $eRand_k$ over $j = 1, 2, ..., 50$ simulations
   - $e_k = \frac{1}{50}\sum_{j=1}^{50} e_k(j)$
   - $eRand_k = \frac{1}{50}(\sum_{j=1}^{50} eRand_k(j))$

be similar for many different stocks, whether for the long-memory component such a similarity might be lost.

As already mentioned in the previous section, statistical limitations underlie the improvements obtained by using higher order blocks, showing much larger fluctuations.

Results previously obtained (Vilela et al., 2002) and the need to further characterize the presence of small and large fluctuations, led to the application of the same method with same data samples being coded by 3-symbol alphabet. As before, s is the standard deviation of the hourly-returns samples.

$$\Sigma = \{-1, 0, 1\} \tag{11}$$

Then,

$$\begin{aligned} \left(r(t) - \overline{r(t)}\right) > s &\iff 1 \\ s \geq \left(r(t) - \overline{r(t)}\right) > s &\iff 0 \\ -s \geq \left(r(t) - \overline{r(t)}\right) &\iff -1 \end{aligned} \tag{12}$$

When this shorter code is adopted, the number of large events is the same as before being the statistics of small fluctuations improved. The method is the same with the single replacement of the 5-symbol alphabet by the new one $\Sigma = \{-1, 0, 1\}$.

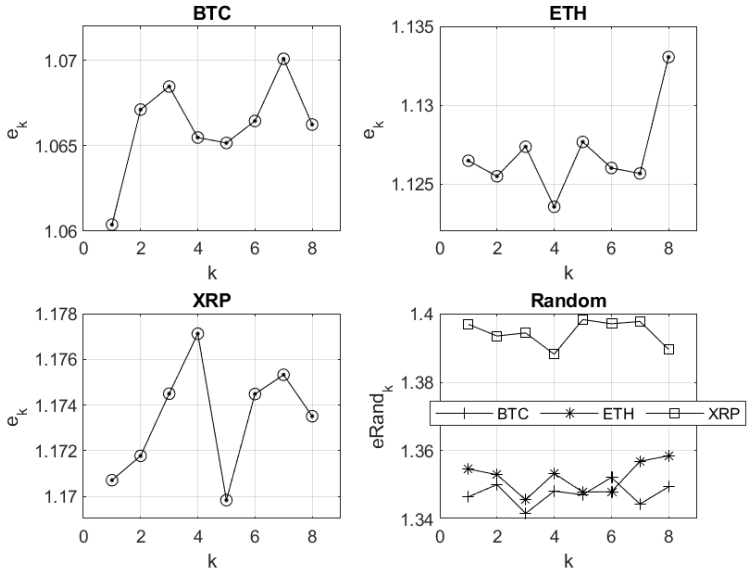

Fig. 5 The past predicting the future, 5-symbols alphabet.

The first three plots in Fig.6 show the average error obtained with a 3-symbols alphabet for the three cryptocurrencies. The last plot shows the error obtained for each cryptocurrency and computed when the prediction is performed at random, i.e., from a surrogate matrix of probabilities.

Results show a prediction improvement extending to block sizes larger than before (with the 5-symbol alphabet). Because small fluctuation errors are decreased by better statistics, the persistence of the improvement for larger blocks seems to highlight the presence of a long-memory component.

Again, the stocks `BTC` and `ETH` seem to share closer similarities than any of them with `XRP`.

The first two plots Fig.6 show that e_k computed for `BTC` and `ETH` is equally ranged in the y-axis. On the contrary, e_k computed for `XRP` displays quite different limits. A closer correlation between the first two stocks is also present in the values of $eRand_k$ as the last plots in Fig.5 and Fig.6 show.

These difference displayed by `XRP` fluctuations is certainly related to the much larger flights observed in the dynamics of `XRP` hourly returns presented in Fig.3. The improvement in the predictions obtained for small-order blocks is similar to those presented in reference Vilela et al., 2002, where the dynamics of standard stocks were analyzed. A similar approach has also analyzed the dependence of memory on the dynamics of processes of cryptocurrency daily-returns. There, Nascimento et al., 2022 - also looking at Bitcoin, Ethereum and Ripple - found the occurrence of long-range memory up to 7-order Markov chains.

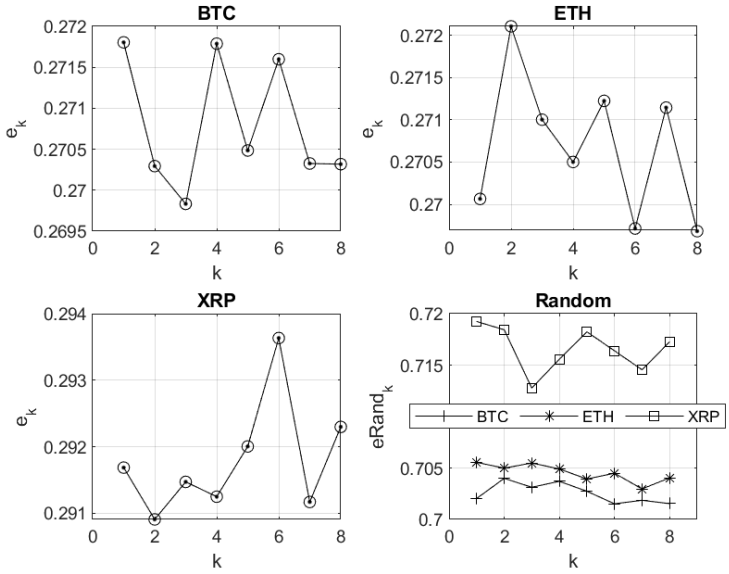

Fig. 6 The past predicting the future, 3-symbols alphabet.

4 Conclusions

In this paper, a Markov approach is used to model the fluctuations of the hourly-returns of three cryptocurrencies: Bitcoin, Ethereum and Ripple.

Markov chains of orders one to eight were considered as a way to predict cryptocurrency returns and to investigate the occurrence of eventual long-memory components in those stochastic processes.

Since conditional probabilities are inferred from limited experimental data, an extended Markov approximation seems to be advantageous. Here, we used the Less-than-k-Markov approximation as presented in reference Vilela et al., 2002.

The most important result is that the average prediction obtained from using the empirical probabilities outperform a random choice.

The main contributions rely on a predictive approach not yet used for series of cryptocurrencies. Moreover, using hourly data we benefit from better statistics when compared with daily ones but avoiding the inconvenient of high-frequency data (i.e. minute observations) since it involves the interplay of many more reaction time scales and market compositions in the trading process. Therefore, the choice of series of hourly observations seems to be an appropriate way to understand the stochastic process that underlies the market mechanism.

Notice, however, the trade-off between higher order approximations and lack of statistics, the main limitation of our approach. Future work is planned to apply the same approach to explore the use of the empirical probabilities of

one cryptocurrency to predict the behavior of the others. In so doing, we would be able to understand the strength of connectivity between digital stocks in the behaviour of the cryptocurrency market.

Acknowledgments. The authors acknowledge financial Support from FCT – Fundação para a Ciência e Tecnologia (Portugal). This article is part of the Strategic Project UIDB/05069/2020. The authors acknowledge financial Support from FCT – Fundação para a Ciência e Tecnologia (Portugal).

References

Bariviera, A., Basgall, M. Hasperué, W. & Naiouf, M. (2017). Some stylized facts of the bitcoin market, *Physica A: Statistical Mechanics and its Applications* 484, 82–90. https://doi.org/10.1016/j.econlet.2015.02.029

Baur, D. & Lucey, B. (2010). Is gold a hedge or a safe haven? an analysis of stocks, bonds and gold, *Financial Review* 45(2), 217–229. https://doi.org/10.1111/j.1540-6288.2010.00244.x

Charfeddine, L. Benlagha, N. & Maouchi, Y. (2020), Investigating the dynamic relationship between cryptocurrencies and conventional assets: Implications for financial investors. *Economic Modelling*, 85, 198-217. https://doi.org/10.1016/j.econmod.2019.05.016

Cheah, E-T. & Fry, J. (2015). Speculative bubbles in bitcoin markets? an empirical investigation into the fundamental value of bitcoin, *Economics letters* 130, 32–36. https://doi.org/10.1016/j.econlet.2015.02.029

Coinmarketcap.com at https://coinmarketcap.com.

Corbet, S., Meegan, A., Larkin, C., Lucey, B., & Yarovaya, L. (2018). Exploring the dynamic relationships between cryptocurrencies and other financial assets, *Economics Letters* 165, 28-34. https://doi.org/10.1016/j.econlet.2018.01.004

Cunha, C. & Silva, R. (2020). Relevant stylized facts about bitcoin: Fluctuations, first return probability, and natural phenomena, *Physica A: Statistical Mechanics and its applications* 550, 124155 https://doi.org/10.1016/j.physa.2020.124155

Dyhrberg, A. (2016). Hedging capabilities of bitcoin. is it the virtual gold?, *Finance Research Letters* 16, 139–144. https://doi.org/10.1016/j.frl.2015.10.025

John, K., OHara, M. & Saleh, F. (2022). Bitcoin and beyond, *Annual Review of Financial Economics* 14(1), 95–115. https://doi.org/10.1146/annurev-financial-111620-011240

Khmelev, D. & Tweedie, F. (2001). Using Markov Chains for Identification of Writer, *Literary and Linguistic Computing* 16(3), 299–307. https://doi.org/10.1093/llc/16.3.299

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system, *Decentralized Business Review*, 21260

Nascimento, F. Santos, S., Jale, A. Júnior, X. & Ferreira, T. (2022).Extracting rules via markov chains for cryptocurrencies returns forecasting. *Computational Economics*, 1-20. https://doi.org/10.1007/s10614-022-10237-7

Urquhart, A.(2017). Price clustering in bitcoin, *Economics letters* 159, 145–148. https://doi.org/10.1016/j.econlet.2017.07.035

Vilela Mendes, R. Lima, R. & Araújo, T. (2002). A process-reconstruction analysis of market fluctuations, *International Journal of Theoretical Applied Finance* 5(08), 797–821. https://doi.org/10.1142/S0219024902001730