The future of Bitcoin: a Synchrosqueezing Wavelet Transform to predict search engine query trends

Contributions to KDWEB Conference, a.d. 2016

Marco Stocchi*, Ilaria Lunesu, Simona Ibba,
Gavina Baralla, and Michele Marchesi

Dept. of Electric and Electronics Engineering,
University of Cagliari, Italy,

{marco.stocchi,ilaria.lunesu,simona.ibba,
gavina.baralla,michele.marchesi}@diee.unica.it

Abstract. In recent years search engines have become the go-to methods for achieving many types of knowledge, spanning from detailed descriptions or general information interesting to the user. Likewise several reassignment techniques are capturing the attention of researchers in the field of signal analysis. Particularly, the Synchrosqueezing Wavelet Transform - SST allows signal decomposition and instantaneous frequency extrusion, at the same time promising consistent reconstruction capabilities, hence the possibility to contrive an SST assisted inference engine. We are going to test it using datasets extracted from search engine trends, using a cloud of keywords related to the *Bitcoin* topic. This could be useful to study the evolution of the cryptocurrency both in time and geographical terms, and to estimate the future number of queries. The importance of Bitcoin queries prediction goes beyond the academic and research environments and, as such, it could lead to valuable commercial applications, such as financial recommender systems or blockchain-based transaction managers development.

Keywords: search engine query trends, bitcoin, machine learning, synchrosqueezing wavelets

1 Introduction

Search engines have an increasing impact in people's behavior. They are no longer simple instruments of information, but real action drivers. According to [2] Google is the most complete repository of human activities about the past,

present, and future. It has an enormous impact on marketing, culture, business and it is the interpreter of the thoughts, expectations, wishes of the people. Therefore Google is currently the main mean of accessing knowledge. Google Trends is the tool that Google (prominent search engine operator) provides to understand how the user searches evolve in time. It analyzes the keywords, lays down a search index on the time axis and it allows to discover in which geographic region a search is more popular. This index is calculated with a ratio between the query volume for a particular keyword divided by the total number of queries. According to [4] we analyse these query indices because they are correlated with economic indicators that we can use to infer a short-term economic prediction. Queries are important markers to understand subsequent consumer purchases in a particular geographic region and they raise many economic questions. Our goal in this paper is to introduce an analysis and prediction of Google Trends datasets, using a forecasting method featured by a novel preprocessing module based on the SST.

Recently the state of the art relative to the prediction of time series is pointing towards novel approaches based on signal decomposition or time-frequency analysis. To this purpose, the application of wavelet transforms (either in their discrete or continuous versions) constitutes an immediate method to explore the timewise development of a series spectrum and, as such, it has been tested in conjunction of established predictor systems in order to evaluate the achievable forecasting accuracy [8] [16] [18] [19].

In the late 90, Huang N. [11] proposed to decompose a signal in a fixed number of intrinsic mode functions, called IMFs. Such system was named Empirical Mode Decomposition - EMD and it captured the attention of researchers since, by then, there were no existing theoretical guarantees on the possibility to decompose a signal in components featured by both amplitude and phase modulation. Also, in 2005 an Ensemble Empirical Mode Decomposition - EEMD [23] was proposed. It takes advantage of a noise assisted procedure to improve the decomposition results obtained by the original EMD, collating portions of signal of comparable scales into the same respecting modes.

The empirical results obtained with the EMD and EEMD approaches remained beyond the researchers mathematical comprehension until 2011, when Daubechies, Lu, Wu [6], starting from the definition of the Continuous Wavelet Transform - CWT and after having provided the algorithm to perform its reassignment, (the Synchrosqueezed Wavelet Transform - SST), proved that the latter can be viewed as an adaptive time-frequency decomposition whose intent is the same as the EMD. This approach was first applied in audio processing research [7] and, recently, to surface electromyography and electrocardiogram data [6], after having formulated the necessary theoretical guarantees that a Synchrosqueezing Wavelet Transform provides an adaptive time-frequency decomposition comparable to the aforementioned EMD. More recently, Thakur, Brevdo et al. [20] successfully applied SST to paleoclimate time series and proved the stability of the method. Li and Liang [15] used SST to vibration monitor-

ing signal; while Herrera et al. [10] used the same SST approach to identify instantaneous frequencies in seismic signals.

Our interest in the SST approach is motivated by the possibility to contrive a prediction system making use of the Synchrosqueezing Wavelet Transform as a dataset preprocessing module, as well as by new possible developments in the machine learning field. In our work we plan to project and develop a prediction system suitable to the analysis of streaming datasets such as the search engine query trends, whose prediction importance goes beyond the academic environment and, as such, could lead to valuable commercial or industrial applications.

The rest of the paper is organized as follows: Section 2 introduces the Bitcoin and Google Trends with a description of time series extraction procedures, in Section 3, the related works are presented drawing upon relevant literature about the Synchrosqueezed Wavelet Transform, Google trends and Bitcoin. Section 4 depicts plannings for the most important machine learning features of the system and an analysis of the risks connected with the prediction procedures. Section 5 portrays conclusions and an outline of the future research directions.

2 Bitcoin in Google Trends

We focus our attention on the analysis of a cloud of keywords associated to the main keyword "bitcoin". We start from the basic idea that a high number of queries executed on Google corresponds to an expression of high interest of the users community, and that a large number of transactions, either purchases, sales or simple monetary exchanges, are made on the side of the Bitcoin currency. Hence analysing the data trends we intend to predict the future interest on the bitcoins of such users, in order to discover their potential expectations and commitment to keep performing transactions with the same cryptocurrency. Bitcoin is a complete form of digital money. It is the first experimental peer-to-peer payment network operated by users without a central authority or intermediaries, and it allows to send digital cash through the Internet in a quick, cryptographically safe way, and, above all, at no forced cost. In the original manifesto, the anonymous inventor S.Nakamoto described the Bitcoin as a system for electronic transactions without relying on trust, through the use of cryptographic proof [1]. It is not controlled or monitored by any government or central bank, and it stands for an evolutionary phenomenon in the financial markets; however, its exchange rates experienced dramatic high volatility in the past few years. Our intentions to analyse the search engine query trends related to the bitcoins are motivated by the possibility to predict the evolution of future searches related to the same topic; such interest is strictly related to two specific reasons: - the Bitcoin system is totally decentralized, open source and absolute transparent; it is the very first time that crowds can observe a worldwide transaction flux on a public ledger; - it is a system that allows the reduction of transaction costs and related financial risks. The benefits of such innovations to communities cannot be overestimated. These are the main factors that let us infer that the use of bitcoin will increase in time; also, as recently showed by [13], the Bitcoin network is

exhibiting exponential growth. We are going to study a prediction system based on Google Trends in order to test the validity of such hypothesis.

3 Related Works

In recent years search engines have become the go-to methods for achieving many types of knowledge, either attaining detailed topic descriptions or surface information of any kind or interest. Analysing the search terms used, and their frequency, a first indicator of what people are interested in, and how they might use this information for, can be obtained. It is therefore possible to question whether such data contains valuable predictive power that researchers can exploit to build forecasting models. In literature there exist several examples on the use of search queries to predict different real phenomena.

Yang et al. [24] discussed the use of web search query volumes to predict the visitors number of a popular touristic destination in China, comparing the results obtained by search datasets extracted from two different search engines (Google and Baidu). Wu et al. [22] proved that data from search engines (such as Google) provide a highly accurate yet simple way to predict future business activities. They apply a specific methodology suitable to predict the housing market trends. Choi et al. [4] showed the use of search engine's data to forecast near-term values of economic indicators. They present results related to automobile sales, unemployment claims, travel destination planning, and consumer confidence datasets. More recently, as digital money emerged as a new intriguing fact in the financial markets, the behaviour of Bitcoin - the most widespread digital currency, rose questions about the behavior of its exchange rates, at the same time offering a field to study the dynamics of the market, including the effects related to highly speculative investment and trading. In the paper [14], digital currencies and search queries on Google Trends and Wikipedia are connected, and their relationship is studied. Results show that there exist a strong asymmetry between the effects of an increased interest in the currency whenever its price oscillates around its trend values. Yelowitz et al. [25] use Google Trends data to study the driving interest in the Bitcoin, with the caution that search query interest does not imply active participation. Based on informal evidence about Bitcoin users, authors construct proxies for four possible clienteles.

If we can mention other specific cases we can consider the work presented by Matta et al. [17] in which the existing relationship between Bitcoins trading volumes and the queries volumes of Google search engine is studied. Particularly, they found evidence of significant cross correlation values, demonstrating that search volumes power can anticipate changes in trading volumes of the Bitcoin.

In the present work we associated these topics to the use of Synchrosqueezed Wavelet Transform - SST, in order to forecast search engines query trends. The SST is a useful tool for studying multicomponent signals with oscillating modes and processing non stationary signals. After the seminal work of Daubechies, Lu, Wu [6], Thakur, Brevdo et al. [20] provide insights on the stability of Synchrosqueezing, and implementation aspects related to the decomposition and

reconstruction of sampled series via the SST. Recently the SST has gained the attention of other researchers, especially in order to study the effectiveness of its forecasting capabilities. Specifically, H.-T. Wu et al. [21] use the SST as predictor of ventilator weaning. The signal time-frequency analysis, performed with this type of tool, allows authors to have a very good prediction with only 3 minutes of respiration data, wheres traditional methods need about 20 minutes to guarantee a safe forecast. Hazra et al. [9] apply the SST transform on particular signals coming from rotating machinery. The decomposition of the signals allows to estimate *Condition Indicators* CI. The CI are used for a novelty detection technique based on Self Organizing Maps (SOM).

4 Method

As outlined in the seminal works of Daubechies [6] and [20] a synchrosqueezing wavelet transform can be performed in a two steps fashion. The first step is to perform the Continuouns Wavelet Transform - CWT of the signal of interest, the second step is to perform the synchrosqueezing algorithm - SST. We are now going to describe such two steps. The CWT transformation, used to analyze a signal $f(t)$, is naturally endowed with a reconstruction algorithm. In order to recall the CWT algorithm and its reconstruction procedures, let us denote a the scaling parameter and b the translation one; $\psi(t)$ a mother wavelet function whose shifted and scaled versions are denoted $\psi(\frac{t-b}{a})$, and finally write:

$$W_f(a,b) = \int_{-\infty}^{\infty} f(t)\, a^{-1/2}\, \overline{\psi\Big(\frac{t-b}{a}\Big)}\, \mathrm{d}t. \tag{1}$$

$$f(t) = C_\psi \int_{-\infty}^{\infty} \int_{0}^{\infty} W_f(a,b)\, a^{-5/2}\, \psi\Big(\frac{t-b}{a}\Big)\, \mathrm{d}a\, \mathrm{d}b, \tag{2}$$

where C_ψ is constant and depends only on the wavelet. Eq. 1 differs from the Discrete Wavelet Transform - DWT, in which the parameters a, b are selected from a discrete sublattice [5]. Performing the CWT of a pure tone it can be observed that the frequency localization is spread out [6] along the scale axis, and that such effect cannot be avoided. Specifically, considering a tone $f(t) = Acos(\omega t)$ and supposing that the choosen wavelet has a spectrum $\hat{\psi}(\xi)$ concentrated in proximity of $\xi = \omega_0$, the above mentioned spreading effect would be around the scale point $a = \omega_0/\omega$. In order to provide a much more definite instantaneous frequency detection of a signal, however, it can be noted that if the choosen wavelet is complex, the real and imaginary components of the CWT contain enough phase information to pinpoint the oscillatory behaviour of $f(t)$ in the b direction. Hence the intuition [7] to retrieve a matrix $\omega_f(a,b)$, associated to the $W_f(a,b)$, containing the instantaneous frequencies extracted from the CWT via phase transformation:

$$\omega_f(a,b) = -i \frac{\partial/\partial b\, W_f(a,b)}{W_f(a,b)} \tag{3}$$

Where $^{\partial}/_{\partial b}\, W_f(a,b)$ can be calculated using the time derivative property of the Fourier transform. Possessing both W_f and ω_f matrices it is now possibile to reassign the wavelet transform (second step), in order to pass from a time-scale representation to a time-frequency plane. Let us denote $S(W,\omega)$ the Synchrosqueezing operator, and $T_f(\omega,b)$ the Synchrosqueezed Wavelet Tranform of $f(t)$, such that:

$$\begin{gathered} S(W_{a,b},\omega_{a,b}) : (a,b) \to (\omega_{a,b},b) \\ T_f(\omega,b) = \int W_f(a,b)\, a^{-3/2}\, \delta(\omega(a,b)-\omega)\, \mathrm{d}a, \quad a : W_s(a,b) \neq 0 \end{gathered} \tag{4}$$

In a discrete environment, ω spaces linearly or logarithmicly from the fundamental frequency ω_0 to the Nyquist frequency ω_N of a sampled series. If one chooses a linear frequency scale, as we do in the present work, having a vector of frequencies $\boldsymbol{\omega} = \{\omega_0, \omega_i, ..., \omega_N\}$, the operator $S(W,\omega)$ can be contrived simply using a standardized lower bound algorithm to search the ω_i, (center frequency of a "bin" gathering a group of frequencies $[\omega_i - \Omega, \omega_i + \Omega]$, $\Omega = (\omega_i - \omega_{i-1})/2$), nearest to an instantaneous frequency $\omega_f(a,b)$. Once the destination bin is found, the contribuition of the $W_f(a,b)$ can be summed into the right destination position in matrix $\boldsymbol{T}_f$:

$$T_f(\omega,b) = {}^{1}/_{2\Omega} \sum_k W_f(a,b)\, {a_k}^{-3/2}\, (\Delta a)_k, \quad a_k : |\omega(a_k,b) - \omega| \leqslant \Omega \tag{5}$$

The Synchrosqueezing Wavelet Transform possesses a corresponding reconstruction algorithm. Let us denote $f(b)$ the reconstructed signal. In [6] it is proved that the integral form $\int_0^\infty W_f(a,b) a^{-3/2} \mathrm{d}a$ is proportional to $f(b)$. The discrete version of the reconstruction algorithm can then be written from eq.(5):

$$f(b) \approx \Re\Big\{ C_\psi^{-1} \sum_i T_f(\omega_i,b)(2\Omega) \Big\} \tag{6}$$

The above descriptions highlight the foremost theoretical aspects related to the SST, retained as foundation to the development of an advanced Synchrosqueezed Transform implementation.

5 Prediction using SST

Referring to [6] [3] we can list some of the principal SST properties: SST is robust to several types of noise, and it is able to evaluate instantaneous frequency and amplitude modulation. SST is an invertible transformation, that allows to reconstruct the signal accurately. Whenever the analysing signal exhibits seasonality features, seasonal oscillations can be detected, and the signal can be effectively rebuilt. Since the procedure is local in nature, the dynamic evolution of the signals amplitude and frequency can be pointwise detected; also the SST can be used to analyse time series of any length. The SST is an adaptive method,

in fact the choice of the mother wavelet, used to perform a CWT, is not essential to ensure the adherence to the above mentioned properties. Finally, SST removes the energy of the noise out of the range of interest. After the mapping of the CWT into the time-frequency plane, we can partition the frequency axis in a fixed number of sets having the same size, depending on the number of modes $m_i(t)$ we intend to use to decompose the signal. Having created a number N_m of intrinsic modes $m_i(t), 0 \leqslant i \leqslant N_m$, our goal is to perform a prediction of each $m_i(t)$ one-step ahead. If the IMFs could be used as a representative training set, suitable to be input to our inference modules, then reconstructing the one-step ahead forecast would simply be obtained by summing the single mode estimations: [1]:

$$\widehat{f}(t) = \sum_k \widehat{m}_k(t), \quad 0 \leqslant k \leq N_m \tag{7}$$

The prediction could be performed using neural regression, for example by means of a backpropagation multilayer perceptron - BPMLP network plugged to each $m_i(t)$. However, designing the prediction system in such way, there would be risks associated to the accuracy that we set to achieve, since the $m_i(t)$ would exhibit variable oscillatory behaviour. Hence in order to train adequately each BPMLP, the input size should be greater or at least equal to one full oscillation of the mode. Note that this means that, in some cases, the input size of the network would be huge, causing lengthy training (and retraining) operations and degrading the whole systems efficiency, or worse, jeopardizing the feasibility of an effective prediction. Thus, in order to anticipate such issues, the number of hidden BPMLP layers should be initially kept low. This should not represent a limitation since, generally speaking, (and given the experience of the authors on the development and testing of artificial neural networks) the number of hidden perceptron layers must be necessarily increased only if the MLP is used for classification purposes (i.e. the network must learn to classify a great quantity of differently labeled patterns); whereas the MLP is herein going to be used for regression purposes, hence once the features of a mode $m_i(t)$ have changed, the machine could forget the previous internal configuration and adapt to the changing statistical properties of $m_i(t)$. Also, in order to preserve systems long memory, experience on financial series testing taught us to effectively employ self organizing layers (derived by Kohonen's original formulation of Self Organizing Maps [12]) to store time series behaviours in prototypes structured databases.

6 Conclusions

Search engine trends, through the insertion of keywords, allow us to understand the search volume evolution over time. Therefore, a huge amount of interesting time series exists, and it could be analysed for several purposes. After having investigated the literature related to both the Google Trends and the Bitcoin

[1] let us denote in this case the estimations using the *hat* accent hoping that no confusion with the classical Fourier transform's notation occurs.

network, a novel forecasting approach is attracting our attention, as well as the possibility to use it for predicting the aforementioned data series. We start from the basic idea that a time series can be sampled in patterns of fixed size; such patterns can be preprocessed in order to produce inputs to a predictor system. Our purpose is to perform an accurate forecast of the Bitcoin search volumes. The approach we are going to propose is based on the original series decomposition (time-frequency analysis), based on instantaneous frequency extrusion operations performed via the SST. Such preprocessing step is meant to assist a group of cooperating statistic predictors or neural networks; our implementation experience will let us face neural regression issues more rapidly, and for such reason we will privilege backpropagation multilayer perceptrons tests first, in order to evaluate the difficulties related to the forecast of oscillatory signals in a reasonable amount of time. Eventually, we may formulate a novel machine learning technique for real-valued forecast, tailored to predict amplitude and frequency modulated signals (such as the intrinsic modes extracted via the SST) in an efficient manner. We think that the success of the proposed approach is key to machine-learning innovations suitable to predict several categories of time series (such as the search engine trends); also, the extensibility of the model to the analysis and prediction of general time series cannot be overruled at this time. Even if our approach could reveal fallacies, however, the possibility of predicting the Bitcoin trends is of great interest to users of the cryptocurrencies, its importance being not only confined to the academic and research fields, but also crucial to understand how timely and geographically can a blockchain evolve. Whether the Bitcoin (or a next-generation cryptocurrency) is committed to eventually replace the traditional fiat currencies is still an open issue, but we consider that modeling its evolution, in terms of consensus trends, is a first approach to our empirical understanding of future developments in the field of financial transactions.

References

1. G. S. Atsalakis and K. P. Valavanis. Surveying stock market forecasting techniques part ii: Soft computing methods. *Expert Systems with Applications*, 36:5932–5941, 2009.
2. J. Battelle. *The search: how Google and its rivals rewrote the rules of business and transformed our culture.* Nicholas Brearley Publishing, 2005.
3. Yu-Chun Chen, Ming-Yen Cheng, and Hau-Tieng Wu. Non-parametric and adaptive modelling of dynamic periodicity and trend with heteroscedastic and dependent errors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):651–682, 2014.
4. Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9, 2012.
5. I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 1988.
6. I. Daubechies, J. Lu, and Hau-Tieng W. Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. *Applied and Computational Harmonic Analysis*, 30:243–261, 2011.

7. Ingrid Daubechies and Stephane Maes. A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models. *Wavelets in medicine and biology*, pages 527–546, 1996.
8. I. Francis and K. Sangbae. *An introduction to Wavelet Theory in Finance: A Wavelet Multiscale Approach*. World Scientific, 2012.
9. Budhaditya Hazra, Shilpa Pantula, and Sriram Narasimhan. Novelty detection in airport baggage conveyor gear-motors using synchro-squeezing transform and self-organizing maps. In *PHM Society Conference*, volume 4, page 060, 2013.
10. Roberto H Herrera, Jiajun Han, and Mirko van der Baan. Applications of the synchrosqueezing transform in seismic time-frequency analysis. *Geophysics*, 79(3):V55–V64, 2014.
11. N. Huang. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond.*, 454:903–995, 1998.
12. Teuvo Kohonen. The self-organizing map. *Neurocomputing*, 21(13):1 – 6, 1998.
13. D. Kondor, M. Psfai, I. Csabai, and G. Vattay. Do the rich get richer? an empirical analysis of the bitcoin transaction network. *PLoS ONE*, 9(2): e86197. doi:10.1371/journal.pone.0086197, 2014.
14. Ladislav Kristoufek. Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific reports*, 3, 2013.
15. Chuan Li and Ming Liang. A generalized synchrosqueezing transform for enhancing signal time–frequency representation. *Signal Processing*, 92(9):2264–2274, 2012.
16. P. Masset. Analysis of financial time-series using fourier and wavelet methods. *SSRN*, page http://ssrn.com/abstract=1289420, 2008.
17. M. Matta, I. Lunesu, and M. Marchesi. The predictor impact of web search media on bitcoin trading volumes. *Information Filtering and Retrieval - DART 2015*, 2015.
18. J.B. Ramsey. Wavelets in economics and finance: Past and future. *Studies in Nonlinear Dynamics and Econometrics*, 6(3), 2002.
19. J.C. Reboredo and M.A. Rivera-Castro. Wavelet-based evidence of the impact of oil prices on stock returns. *International Review of Economics and Finance*, 29:145–176, 2014.
20. G. Thakur, E. Brevdo, N.S. Fuckar, and H.T. Wu. The synchrosqueezing algorithm for time-varying spectral analysis: Robustness properties and new paleoclimate applications. *Signal Processing*, 93(5):1079–1094, 2013.
21. Hau-Tieng Wu, Shu-Shua Hseu, Mauo-Ying Bien, Yu Ru Kou, and Ingrid Daubechies. Evaluating physiological dynamics via synchrosqueezing: Prediction of ventilator weaning. *Biomedical Engineering, IEEE Transactions on*, 61(3):736–744, 2014.
22. L. Wu and E. Brynjolfsson. The future of prediction: How google searches foreshadow housing prices and sales. *Social Science Research Network*, 2013.
23. Z. Wu and N.E. Huang. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, 1:1–41, 2009.
24. Xin Yang, Bing Pan, James A. Evans, and Benfu Lv. Forecasting chinese tourist volume with search engine data. *Tourism Management*, 46:386 – 397, 2015.
25. Aaron Yelowitz and Matthew Wilson. Characteristics of bitcoin users: an analysis of google search data. *Applied Economics Letters*, 22(13):1030–1036, 2015.