Towards Cryptocurrency Index - Analysis of the market

Master's Thesis submitted

to

Prof. Dr. Wolfgang Karl Härdle

Humboldt-Universität zu Berlin
School of Business and Economics
Institute for Statistics and Econometrics
Ladislaus von Bortkiewicz Chair of Statistics

by

Simon Trimborn
553486

in partial fulfillment of the requirements
for the degree of
Master of Science
Berlin, March 30, 2015

Abstract

A market index is constructed for the crypto currency market by using newly developed methods for such a task. The choice of the number of index constituents is performed with the AIC and BIC criterion and the liquidity rule is set by taking into account the BIS survey. This newly created index, CRIX, is then used to compare the market against Bitcoin and other markets. It is found that the crypto market is much riskier than other known markets.
It is also created a minimum variance CRIX and an optimal forecasting model is found by taking into account social media data.

Keywords: Index construction, Cryptocurrencies, Bitcoin, Forecasting, Minimum variance, CRIX

Acknowledgements

I'd like to thank Prof. Wolfgang Härdle for his advice and support in writing this thesis. I also appreciate the valuable talks with the members of the Ladislaus von Bortkiewicz Chair of Statistics.

I further want to thank Prof. David Lee and Dr. Ernie Teo of the Sim Kee Boon Institute for Financial Economics at the Singapore Management University for their help and support.

A special gratitude is dedicated to my parents, Werner and Susanne Trimborn, for their support while my studies.

List of Figures

List of Tables

Contents

1 Introduction

In 2009 a new kind of currency came on the market. It was revolutionary in many directions. It has a limited amount, it works without a central bank, it relies on cryptography, the users make changes to its structure and the whole transaction history is public, just to name a few. It is completely different to all the currencies known so far. This is Bitcoin, the first cryptocurrency. Since the founding this market saw a success story. Coinmarketcap, `https://coinmarketcap.com/all/views/all/`, tracks at the 2015-03-24 595 cryptocurrencies (cryptos) with a combined market capitalization of round about 4 billion USD. Some of these cryptos brought again a revolutionary approach into the market, others copied just the code of Bitcoin. Perhaps because it is the first one in the market and hadn't a competitor for around 2 years, it became the most important crypto. It showed over time a huge increase in price and trading volume, see figure 1.

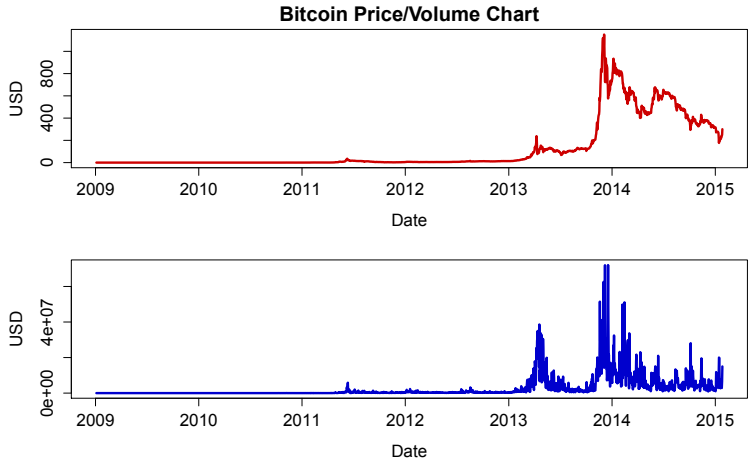


Figure 1: Price, Volume, source: www.blockchain.info

But also the trading volume of the entire market increased in the recent time. This is visualized with the figure 1, while the 2014-07-14 is excluded because the value, $2,651,869,935$ USD, is that high that the impression of the image would be distorted.

The public interest in this market is increasing. Many people don't know about cryptos, but many already heard about Bitcoin. A huge number of start-ups came up which are working on Bitcoin and/or cryptos. Some offer exchanges, some work on new kinds of cryptos, others follow completely different ideas. But still is the possibility missing to get an idea about the performance of this market. Here can catch in a market index.

Market Indeces are a widely applied instrument to get information

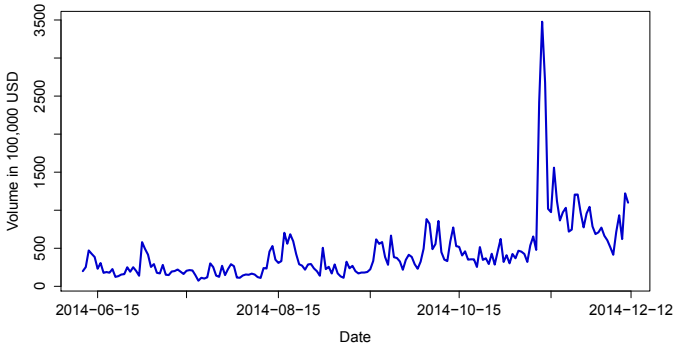


Figure 2: All Coins in CoinGecko database on 2014-12-12

about a specified market or asset universe. A very famous example is the S&P500 which is often interpreted as a market proxy for the US bluechips stock market. The very same is the DAX30 for the german stock market. These indeces are used e.g. to find out if a stock performed well compared to its market. They are also used to track the performance of a market. But for the rather new market of cryptocurrencies is this comparability still not possible, since a trustful index does not exist. This makes it very hard to compare this market against another one. Of course, it would be easy to look just at Bitcoin. This is till now the crypto which is the most famous and attracts more people than other cryptos. But there are other cryptos which are promising and/or still play an important role in the market. Therefore it is worth to look not just at Bitcoin when talking about the crypto market.

In the following it will be first described how to construct such an index and what methods are applied here. For the number of index constituents is used a AIC and BIC comparison of indeces with a different amount of constituents to find the one which fits the best and has at the same time as less index members as possible. Due to the behavior of the residuals it is necessary to apply in this step stable distributions. The newly created index (CRIX) is afterwards used to compare Bitcoin against the market. Also it is created a minimum volatility index to test if this kind of optimization brings value. On top of that it is checked how risky that new market is, compared to known markets to give the reader an idea about what level of risk to think when working with the crypto market. After finding the risk level of the CRIX it will be checked if it is possible to forecast the index in order to eliminate uncertainty from an investment. Since the community of some cryptos is very active in social media platforms, it suggests that social media data could improve the forecast, what is indeed possible.

2 Index construction

Indices are very important in the financial industry to measure the performance of a market or a sector. They are dedicated to give any interested person an overview how this market is performing at the moment or performed in the past. They are often used to compare investments from inside of this market against the index. By doing this can be concluded if this investment performed better or worse than the market measure (benchmark). Index theory is therefore a very up-to-date topic and institutes like the MSCI or S&P offer many indices to give market participants a good insight into a defined market. But index theory isn't a rather new topic, indeed it is a concept from the 19th century which was especially impelled by Laspeyres and Paasche. According to Lippe (2013), the Laspeyres index is given by the formula

$$P_{0t}^{L} = \frac{\sum_i P_{it} Q_{i0}}{\sum_i P_{i0} Q_{i0}}$$

with P_{it} the price of item i at time t and Q_{i0} the quantity of item i at time 0.

The price index of Paasche is defined by

$$P_{0t}^{P} = \frac{\sum_i P_{it} Q_{it}}{\sum_i P_{i0} Q_{it}}$$

with P_{it} the price of item i at time t and Q_{it} the quantity of item i at time t.

Both indices have in common that they compute the change in prices relative to a basis year. In the case of Laspeyres, the quantities from the basis year are taken to compare the changes in prices. For Paasches index, the quantities of the underlying basket from the current period are used to compare the evolution of the prices. This leads of course to different results. According to Lippe (ibid.), there are sources stating that Laspeyres would have used the weighting of Paasche when he had better access to such data.

Modern price indices like the $S\&P500$ use often a modified version of P_{0t}^{L}, see Indices (2014). The S&P500 index formula is

$$Index\ value = \frac{\sum_i P_i Q_i}{Divisor} \tag{1}$$

with P, Q and i defined as before. The $Divisor$ represents the basis year and ensures that the $Index\ value$ has on the starting date a predefined value. It will be updated later in case that the quantities change over

time. This is absolutely mandatory because otherwise any change in the number of shares or a change in the index constituents would change the value of the index. This would distort the picture of the movement of the market, therefore it must be ensured, that this does not happen.

2.1 Properties of an Index

The CRyptocurrency IndeX (CRIX) is dedicated to be a benchmark for the cryptocurrency market, therefore it is necessary that every interested person, especially investors, understand its construction and be able to build a portfolio which tracks the index. Also the CRIX should give investors an appropriate benchmark for investments, therefore the construction of the CRIX is based on the four principles from Szado (2012).

1. Transparent and Unambiguous
2. Frame-able and Customize-able
3. Appropriateness and Coverage
4. Invest-able

The first principle is about the components, prices and the methodology. It must be clear which constituents shall be part of an index at all time to everyone. Also there shouldn't exist any ambiguity in the choice of the constituents or the prices. When an asset is traded on several exchanges than it must be made clear from which exchange the price will be taken. To afford this goal it is convenient to publish the methodology of the index and announce any changes in the rules publicly.

The Frame-ability of the second principle means that the interpretation and the message behind the index should be understandable. It should be always clear what kind of market the index covers, e.g. the US share market, and how the index shall be interpreted regarding its rules. The last point relies e.g. on the way how the weights are chosen. An equally weighted index has a different interpretation than one for which the weights are optimized regarding minimum variance of the entire index.

Customize-able refers to the possibility to create sub-portfolios out of the index to catch better the investment preferences of an investor or to create a benchmark which gives insight into a part of the whole index. For example, a subindex can be created for a big market index which just tracks companies with yearly revenues below an assigned value or which companies are from a specific industry sector.

Appropriateness and Coverage refer entirely to the needs of investors. The first means that the CRIX is an useful benchmark for an investment portfolio. Because the goal is to cover the entire market the CRIX shall be an appropriate benchmark for portfolios which reflect the entire crypto market. Coverage means that the index should cover the investment horizon which an investor can access or is allowed to access. Thinking about the whole market, the CRIX should really cover as much cryptos, so that this goal is achieved.

The last property, Invest-ability, refers to the possibility to actually buy the CRIX to low costs. The trading fees at exchanges are currently low, therefore it must be just ensured that the index constituents are tradable. For this property it is necessary to construct a proper liquidity rule, which excludes cryptos with too less trading from the CRIX because these ones wouldn't be buy-able.

2.2 Difficulties of Index construction

While the construction of an index occur several difficulties:

- Weighting
- Number of index members
- Reallocation period
- Cap
- Liquidity

An absolutely essential decision to make is the weighting of the index constituents. There exist very different types of weighting. The S&P500 uses an approach based on capitalization weighting, other indices like the MSCI World Minimum Volatility Index rely on a volatility optimal weighting in the sense of Markowitz (1952). The choice of the weighting clearly defines how the index can be used. Capitalization weighted indices are often designed to be a market proxy, while optimization weightings are more frequently used to beat a market benchmark and rely on the constituents of a market index.

The number of index members is also very important but it exists to my knowledge no clear rule to find the number of index constituents. The amount of index constituents directly refers to the points made in section 2.1, because it defines the coverage of the CRIX. The goal is to construct a market benchmark, therefore it must be found a method which ensures that the CRIX really covers the market. Here it will be used a statistical

approach based on the AIC and BIC criterion to find the optimal amount of constituents.

The reallocation period is of importance since the more often the list of constituents is overlooked, the better reflect the index the current market situation. But regarding the understandability the constituents shouldn't change too often. A daily reallocation would result into a very good proxy for the current market situation but for anyone using the CRIX would this procedure result into unclearness about the index constituents and current weights assigned to the members. The CRIX would also be inconvenient for investments in such a case because an investor has to pay with every change in his tracking portfolio transaction fees. Therefore, it must be found a middle ground which takes into account both matters.

Some indices also have a cap. E.g. the DAX30 from Germany has a 10% cap on a single stock, see AG (2013), and the IPC35 from Mexico has a 25% cap to cut down the incluence of the company of Carlos Slim Helú on the index value, see MEXBOL (2013). While constructing an index like the CRIX it is absolutely necessary to think about the necessity of a cap to ensure that a single crypto or a group of cryptos doesn't rule the market movement.

Very important is that it is ensured, that an asset in an index is traded and therefore can be bought by investors. This is meant with Liquidity. The S&P500 uses the public float rate, the IPC35 relies on the turnover factor. Public float rate means how many stocks of the total amount are available for trade. In the case of the S&P500, it must be at least 50%. The turnover factor is defined by MEXBOL (ibid.) as

$$TO_{it} = \frac{Vol_{it}}{AF_{it}}$$

with Vol as the volume, AF as the number of floating shares and TO as the turnover factor. i stands for stock i and t for time. Therefore, there is again no clear rule which approach to use. Due to the lack of information for the usually used rules, a different approach, described in section 5.5, will be used here.

3 Methodology

3.1 Index Mathematics

The index mathematic rely on the one used for the S&P500, see Indices (2014). The basis for the S&P500 is the formula, already given in 1,

$$Index\ value = \frac{\sum_{i=1}^{n} MV_i}{Divisor}$$

with MV_i as the market capitalization of the index constituent i. The $Divisor$ is a crucial instrument in the construction because it ensures that the index value displays just when the price changes relative to the base value. On the start of the index will the $Divisor$ be chosen such that the base value is reached, e.g.

$$Divisor = \frac{\sum_{i=1}^{n} MV_i}{1000}$$

with $base\ value = 1000$. This is used later to make the index invariant against changes in the index constituents. At the reallocation date the index constituents can change due to a gain or loss in market capitalization of an asset. They can change also between two reallocation dates due to the failure of a constituent. Both shouldn't affect the value of the index. The $Divisor$ has to change then in the following sense:

$$\frac{\sum_{i=1}^{n} MV_{i,t-1}}{Divisor_{t-1}} = Index\ value = \frac{\sum_{j=1}^{n} MV_{j,t}}{Divisor_t} \tag{2}$$

with $MV_{i,t-1}$, $MV_{j,t}$ the market capitalization of index constituent i respectively j at time $t-1$ (right before a change in the constituent list) and at time t (right after the change) and n as the amount of constituents.

If the index shall have a cap, a modification of the formulas is necessary. Following Indices (ibid.) it is then necessary to compute a capping factor with this formula for every index constituent i on the reallocation date t:

$$AWF_{it} = \frac{CW_{it}}{W_{it}} \tag{3}$$

with CW_i the capped weight and W_i the weight the asset i would normally have in the index. The formula for the index calculation is then

$$Index\ value = \frac{\sum_{i=1}^{n} MV_i \cdot AWF_i}{Divisor} \tag{4}$$

with AWF_i the current adjusted weighting factor of constituent i.

3.2 Missing data

Cryptocurrencies are traded every day, including weekends. This is not the case for financial data from standard exchanges. But it would be inappropriate to exclude the weekend information, instead an approach is necessary to find datapoints for the weekends, so that an analysis becomes possible. It can also happen that data are missing in some of the time series. To circumvent these challenges there will be applied two approaches.

3.2.1 Last Observation Carried Forward

To find values for the weekends of the financial time series data like S&P500 the 'Last Observation Carried Forward' (LOCF) approach will be used, see e.g. Enders (2010). LOCF simply assigns the last observed information to the next day. Since the financial time series which are used in this analysis aren't traded on weekends and holidays, it can be assumed that the value doesn't change. This makes this approach appropriate for the analysis.

3.2.2 Multiple Imputation with Bootstrapping

When data are missing due to other events, e.g. a technical issue of the data provider, LOCF is not appropriate since the data exist, they were simply not stored. Enders (ibid.) shows several empirical studies in medicine in which this approach doesn't work properly in such cases and that it is often recommended a multiple imputation procedure. The applied approach here is 'Multiple Imputation with Bootstrapping' (MIB), documented in Honaker, King, and Blackwell (2011). The authors published their approach in the R-package 'Amelia II', which will be used to overcome the missing data problem. The approach works as follows:

1. Generate m samples out of the dataset with the observed values.
2. Run bootstrapping procedure to replace unobserved values with B loops per sample.
3. Obtain distribution from bootstrapped samples and find imputed values with 'Expectation-Maximization' (EM).
4. Perform analysis on the m samples.
5. Combine results with arithmetic mean.

The MIB method assumes multivariate normality for the dataset $\mathcal{X}$, $(n \times d)$, consisting of $\mathcal{X}^{obs}$ and $\mathcal{X}^{mis}$ which represent the observed and unobserved data respectively. Formally written this is

$$\mathcal{X} \sim \mathrm{N}_d(\mu, \Sigma),$$

with d as dimension, μ the vector of expected values and Σ the covariance matrix.

The normality assumption is not always fulfilled. To overcome this drawback the Box-Cox-Power-Transformation will be used to seek for a transformation factor λ, which ensures normality. The transformation was introduced in Box and Cox (1964) and is given by the following formula

$$\mathcal{X}^{(\lambda)} = \begin{cases} \frac{(\mathcal{X}+C)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(\mathcal{X}+C) & \text{if } \lambda = 0 \end{cases}$$

with C a vector of constants. It will be used a grid searching algorithm to find the optimal λ. The Shapiro-Wilk test, introduced in Shapiro and Wilk (1965), is used to test for normality. Since a multivariate normality test is necessary the expansion of the Shapiro-Wilk test from J. P. Royston (1983) will be taken. First, the formula for the original test of Shapiro-Wilk is

$$W = \frac{(\sum_{i=1}^{n} A_i X_{1,i})^2}{\sum_{i=1}^{n}(X_{1,i} - \bar{X_1})^2}$$

with $X_{1,1} < X_{1,2} < \cdots < X_{1,n}$ where X_1 is one variable in $\mathcal{X}$ and $A^{\top} = (A_1, \ldots, A_n) = \frac{K^{\top}\mathcal{V}^{-1}}{(K^{\top}\mathcal{V}^{-1}\mathcal{V}^{-1}K)^{1/2}}$, where $K = (K_1, \ldots, K_n)^{\top}$ are expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and $\mathcal{V}$ is the covariance matrix of those order statistics, as stated in Razali and Wah (2011). With the extension of P. Royston (1992) the test can be used for $3 \leq n \leq 5000$. Defining for the multivariate version under the assumption that W follows a normal distribution

$$Z = ((1 - W)^{\kappa} - \mu)/\sigma.$$

μ is the mean, σ the standard deviation of the normal distribution and κ an adjustment parameter. Z follows in this case a standard normal distribution. Assume now that Z_i with $i = 1, \ldots, d$ were obtained, then

$$V_i = \left[\Phi^{-1}\{\frac{1}{2}\Phi(-Z_i)\}\right]^2$$

with Φ the cdf of the standard normal distribution.

Following J. P. Royston (1983) $V_i \sim \chi_1^2$ does hold. By defining $\mathcal{C} = corr(\mathcal{V})$ as the correlation matrix of $\mathcal{V}$ consisting of the V_i, showed J. P. Royston (ibid.) that

$$H = \frac{e}{d} \sum_{i=1}^{d} V_i \overset{a}{\sim} \chi_e^2$$

with

$$e = \frac{d}{1 + (d-1)\bar{C}}$$

where $\bar{C} = \sum_i \sum_j C_{ij}/(d^2 - d)$. H serves now as the test statistic which follows approximately a χ_e^2 distribution.

The reason for using this test and not a Kolmogorov-Smirnov test or Anderson-Darling test is, that Razali and Wah (2011) showed that the Shapiro-Wilk test is the most powerful in identifying normally distributed data.

The second important assumption is, that the data are 'Missing At Random' (MAR). Following Honaker, King, and Blackwell (2011), this means that the pattern of missingness just depends on $\mathcal{X}^{obs}$. Specify $\mathcal{M}$ as the missingness matrix with

$$M_{ij} = \begin{cases} 1, & \text{if } X_{ij} \in \mathcal{X}^{mis} \\ 0, & \text{otherwise,} \end{cases}$$

the MAR assumption is then defined as

$$P(\mathcal{M}|\mathcal{X}) = P(\mathcal{M}|\mathcal{X}^{obs}).$$

The algorithm works as following. First a bootstrapping approach with replacement will be used to create m datasets. After it, the EM algorithm is used to find proper values for the missing data.

The Algorithm. Due to the multivariate normality assumption, it follows that the marginals are also normally distributed. The likelihood function is known to depend on the mean and variance and as Honaker and King (2010) noted, this assumption implies that the data can be imputed by a linear regression model. The sufficient statistics for this regression are therefore the mean and variance. As stated in Honaker and King (ibid.), the matrix $\mathcal{Q} = \mathcal{X}^\top \mathcal{X}$ summarizes the sufficient statistics

because of the joint normality. It holds

$$\mathcal{Q} = \sum_i \begin{pmatrix} n & X_{i1} & \dots & X_{ik} \\ X_{i1} & X_{i1}^2 & \dots & X_{i1}X_{ik} \\ \vdots & & \ddots & \\ X_{ik} & \dots & & X_{ik}^2 \end{pmatrix},$$

because the first column of $\mathcal{X}$ is a constant. That is according to Honaker and King (2010) a crucial assumption. The variable which indicates the date is here the constant term.

Honaker and King (ibid.) carry out that by using the sweep operator $\mathcal{Q}$ will be transformed to the parameters of the conditional mean and the unconditional covariance matrix. They further define S as a binary vector which indicates with $S_i = 1$ the rows and columns to sweep and with $S_i = 0$ the ones not to sweep. The resulting matrix shall be termed $\theta(S)$. The example from Honaker and King (ibid.) is performed on the first row and column and gives

$$\theta\{S = (1, 0, \dots, 0)\} = \begin{pmatrix} -1 & \mu \\ \mu^\top & \Sigma \end{pmatrix}.$$

After constructing the $\mathcal{Q}$-matrix follows the **E-step** of the EM algorithm. In this step the expectation of the quantities will be computed where necessary. As Honaker and King (ibid.) stated by treating observed values as known holds the following:

$$\mathrm{E}[X_{ij}X_{ik}] = \begin{cases} x_{ij}x_{ik} & \text{if } M_{ij} = M_{ik} = 0 \\ \mathrm{E}[X_{ij}]x_{ik} & \text{if } M_{ij} = 1, M_{ik} = 0 \\ \mathrm{E}[X_{ij}X_{ik}] & \text{if } M_{ij} = M_{ik} = 1. \end{cases}$$

The expectations can be computed by the following relations:

$$\begin{aligned} \mathrm{E}[X_{ij}] &= x_i^{obs}\theta\{1 - M_i\}_j^b \\ \mathrm{E}[X_{ij}X_{ik}] &= \mathrm{E}[X_{ij}]\,\mathrm{E}[X_{ik}] + \theta\{1 - M_i\}_{jk}^b \end{aligned}$$

with b denoting the iteration round of the EM algorithm. With this formulas at hand a new dataset, $\hat{\mathcal{X}}$, which consists of the observed values and the expected values for the missing data can be constructed, derived by the formula

$$\hat{X}_i^{b+1} = x_i^{obs} + M_i * (x_i^{obs}\theta\{1 - M_i\}^b)$$

where $*$ is the operator for element wise multiplication.

Honaker and King (2010) state that the missing values within any observation have a covariance matrix which can be extracted as submatrix of θ as

$$\Sigma^{b+1}_{i|x_i^{obs}} = M_i^\top M_i * \theta\{1 - M_i\}^b.$$

The latter equation will be $\sigma^2_{ij} = 0$ for all covariances unless i and j are both missing in this observation.

In the **M-step** the formulas from the **E-step** will be used to update $\mathcal{Q}$, which results to

$$\mathcal{Q}^{b+1} = \sum_i \left(\hat{X}_i^{b+1\top} \hat{X}_i^{b+1} + \Sigma^{b+1}_{i|x_i^{obs}} \right).$$

Regarding the convergence, Honaker and King (ibid.) note that the values of the observed data are constant throughout the entire iteration process. The missing data are filled in with the current estimates of the sufficient statistics. Since in each iteration the last estimate with partial weight and the observed data are been taken into account, the next estimate will be closer to the true values by construction. The iteration procedure stops, when the new values don't change too much, so that they are assumed to be close to the optimum. Honaker and King (ibid.) state that convergence to at least a local optimum is guaranteed under simple regularity conditions.

By this procedure are m datasets obtained and the analysis is then performed on each of this sets. Finally, it is necessary to combine the results. Honaker, King, and Blackwell (2011) suggest to bring the results simply with the arithmetic mean together:

$$\bar{Q} = \frac{1}{m} \sum_{j=1}^{m} Q_j$$

with Q_j as the statistic from dataset j, which can be e.g. the regression coefficient or Expected Shortfall. The variance of the point estimate $\bar{Q}$ is now the average of the variances inside of each dataset plus the variance across the datasets, see Honaker, King, and Blackwell (ibid.). A correction term for the bias is necessary for the latter part of the variance because $m < \infty$. The corresponding formula is then

$$\sigma^2_Q = \frac{1}{m} \sum_{j=1}^{m} \sigma^2_{Q_j} + \sum_{i=1}^{m} \frac{(Q_i - \bar{Q})^2}{m-1} \left(1 + \frac{1}{m}\right) \tag{5}$$

where $\sigma^2_{Q_j}$ is the variance within the dataset j.

3.3 Stable distributions

To find the optimal number of index constituents, a model comparison based on Akaikes Information Criterion (AIC) and Schwartz Information Criterion (BIC) will be used, see Akaike (1998) and Schwarz (1978), defined as

$$AIC = -2\ln L_\theta(y) + 2k \tag{6}$$

and

$$BIC = -2\ln L_\theta(y) + k\ln(n) \tag{7}$$

with k the number of parameters, θ the parameters, y the realizations of a random variable Y and L the likelihood function. For the likelihood function a distribution is necessary, therefore we need to make at least an assumption about the distribution family. The used class of distributions will be the stable distributions due to their more flexible form in modeling the returns of a model.

Following the definition in Nolan (2015) stable variables are defined as following:

Definition 1. *A random variable Y is stable if for Y_1 and Y_2 independent copies of Y and any positive constants a and b,*

$$aY_1 + bY_2 \overset{\mathcal{L}}{=} cY + d$$

holds for some positive c and some $d \in \mathbb{R}$.

To cite Nolan (ibid.): 'The most concrete way to describe all possible stable distributions is through the characteristic function or Fourier transform.' He gives the following alternative definition for the $0-$parameterization, using the sign function defined as

$$sign(u) = \begin{cases} -1 & u < 0 \\ 0 & u = 0 \\ 1 & u > 0. \end{cases}$$

Definition 2. *A random variable Y is stable if*

$$Y \overset{\mathcal{L}}{=} \begin{cases} \gamma(Z - \beta\tan\frac{\pi\alpha}{2}) + \delta & \alpha \neq 1 \\ \gamma Z + \delta & \alpha = 1, \end{cases}$$

where Z is a random variable with characteristic function

$$\mathrm{E}\left[\exp(iuZ)\right] = \begin{cases} \exp\left(-|u|^{\alpha}\left[1 - i\beta\tan\frac{\pi\alpha}{2}\left(sign(u)\right)\right]\right) & \alpha \neq 1 \\ \exp\left(-|u|\left[1 + i\beta\frac{2}{\pi}\left(sign(u)\right)\ln|u|\right]\right) & \alpha = 1 \end{cases}$$

and the four parameters are defined in the range $\alpha \in (0.2]$, $\beta \in [-1, 1]$, $\gamma \leq 0$ and $\delta \in \mathbb{R}$.

The characteristic function of Y is given by

$$\mathrm{E}\left[\exp(iuY)\right] =$$
$$\begin{cases} \exp\left(-\gamma^{\alpha}|u|^{\alpha}\left[1 + i\beta(\tan\frac{\pi\alpha}{2})(sign(u))(|\gamma u|^{1-\alpha} - 1)\right] + i\delta u\right) & \alpha \neq 1 \\ \exp\left(-\gamma|u|\left[1 + i\beta\frac{2}{\pi}(sign(u))\ln(\gamma|u|)\right] + i\delta u\right) & \alpha = 1. \end{cases}$$

Cizek, Härdle, and Weron (2011) call α the index of stability, β the skewness parameter, γ and δ the scale and location parameter.

Cizek, Härdle, and Weron (ibid.) state that the Maximum Likelihood (ML) approach is the most accurate but slowest estimation method. They mention that for a vector of observations, $y = (y_1, \ldots, y_n)$, and the parameters $\theta = (\alpha, \beta, \gamma, \delta)$ the maximum log likelihood function is

$$L_{\theta}(y) = \sum_{i=1}^{n} \ln \hat{f}(y_i; \theta)$$

where $\hat{f}(\cdot; \theta)$ is the approximated stable pdf. Because the pdf is in general not known, it is necessary to approximate it numerically. This formula will be then used to estimate the parameters so that the resulting distributions are as accurate as possible.

To check afterwards if the resulting distribution is reasonable will be tested with the Kolmogorov-Smirnov test if the residuals fit the ecdf of simulated observations of the estimated stable distribution. The Null-hypothesis is given by

$$H_0 : F_n(y) = \hat{F}(y)$$

and following Cizek, Härdle, and Weron (ibid.) the test statistic is obtained by

$$D = \sup_{y} |F_n(y) - \hat{F}(y)|$$

where $\hat{F}(y)$ is the approximated stable distribution and $F_n(y)$ is the empirical cdf of the residuals. The critical values are tabulated.

3.4 Analysis tools

To analyze the performance of Bitcoin and other markets against the CRIX different key indicators will be compared. These figures will be presented in this section.

3.4.1 Return

The first indicator is the return. In the rest of this paper the log returns are always used for the analysis, but for the market comparison the absolute return will be taken. This one is given by the formula

$$r_a = \frac{y_T}{y_1}$$

with y_T the realized value at the last day in the time series and y_1 the corresponding value from the first day.

3.4.2 Volatility

Two approaches are very common to measure the volatility of a stock series, the realized volatility and GARCH models. E.g. Liu and TSE (2013) show in their Monte Carlo study that GARCH estimates outperform the estimates of the realized volatility for estimating the daily volatility. A GARCH model will be used here, namely a GARCH(1,1), defined as

$$\sigma_t^2 = \alpha_0 + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

with $\alpha_0 > 0$, $\alpha, \beta \geq 0$ and $\varepsilon_t|(\varepsilon_{t-1}, \sigma_{t-1}^2, \dots) \sim \mathrm{N}(0, \sigma_t^2)$. This model will be applied to the daily data and to get afterwards the volatility of the whole time period, the following formula will be applied, see Liu and TSE (ibid.):

$$V = \sum_{t=1}^{T} \sigma_t^2$$

where T indicates the last day of the time period. For volatility estimations where missing data are part of the dataset formula 5 has to be applied to account for the variance across the m datasets. In this case $\sigma_{q_j}^2 = V_j$, with $V_j = V$ for every dataset j.

The estimation of the GARCH model is performed with the Pseudo Maximum Likelihood approach. Following Bollerslev (1986) the estimation is then performed by applying

$$L = \frac{1}{T}\sum_{t=1}^{T}\left(-\frac{1}{2}\ln(\sigma_t^2) - \frac{1}{2}\frac{\varepsilon_t^2}{\sigma_t^2}\right).$$

3.4.3 Sharpe Ratio

A further measure of comparison is the Sharpe Ratio, introduced in Sharpe (1966). This is a measure which compares the excess return of an asset with the standard deviation of the excess return and is derived by the formula

$$SR = \frac{\varepsilon_{r,T}}{\sigma_{r,T}}$$

with

$$\begin{aligned}
\varepsilon_{r,T} &= \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{r,t} = \frac{1}{T}\sum_{t=1}^{T}(\varepsilon_{a,t} - \varepsilon_{f,t}),\\
\sigma_{r,t} &= \sqrt{\alpha_0 + \alpha\varepsilon_{r,t-1}^2 + \beta\sigma_{r,t-1}^2},\\
\sigma_{r,T} &= \sqrt{\sum_{t=1}^{T}\sigma_{r,t}^2}
\end{aligned}$$

where $\varepsilon_{a,t}$ is the return of an asset and $\varepsilon_{f,t}$ the return of a secure asset, both at time point t.

Like for the variance measure, described in 3.4.2, it is necessary to take into account the variance across the m datasets when missing data are present. Again formula 5 shall be applied.

The Sharpe Ratio has the big advantage that it brings excess return and the corresponding variance together. A low variance and also a high return boost the SR, therefore come risk, measured by the variance, and gain together into one measure.

3.4.4 Expected Shortfall - Extreme Value Theory

To compare the cryptocurrency market against others the 'Expected Shortfall' (ES) will be compared. To define ES, it is useful to determine first the 'Value-at-Risk' (VaR). Following Artzner et al. (1999) and Franke, Härdle, and Hafner (2008), the VaR is defined as

Definition 3. *Given $\alpha \in (0,1)$ is the VaR_α for a random variable X with distribution function F determined as*

$$VaR_\alpha(X) = \inf\{x|F(x) \leq \alpha\}.$$

The ES is then determined as

$$\mathrm{E}\left[X|X > VaR_\alpha\right]$$

Gschöpf (2014) uses different approaches to find the ES in a small sample environment. Two of this methods will be used in this analysis.

The first one is the approach of McNeil and Frey (2000), the second one is presented in Taylor (2008). Gschöpf (2014) found the first approach to be significantly unbiased but inefficient in small samples. The second one is more efficient but significantly biased. Both will be applied to the different time series to overcome the drawbacks of both approaches and ensure that the comparison is qualified.

McNeil and Frey (2000) defined $\{\varepsilon_t^{neg}\}_{t\in\mathbb{Z}}$ as a strictly stationary time series which represents the negative log returns of the underlying. It is assumed that the negative log returns follow the process

$$\varepsilon_t^{neg} = \mu_t + \sigma_t Z_t \tag{8}$$

with Z_t as a strict white noise process. They proposed an ARMA-GARCH approach to obtain the realizations of Z_t. A pseudo ML approach is used, so that no assumption on the distribution of Z, $F_Z(z)$, is made. Afterwards a threshold u is chosen and a 'General Pareto Distribution' (GPD) is fitted to the data beyond this threshold. McNeil and Frey (ibid.) state that it is assumed that the tails begin with the threshold u. Therefore the choice of u is crucial for the analysis.

The GPD has the following distribution function, as given in McNeil and Frey (ibid.):

$$G_{\xi,\zeta}(z_t) = \begin{cases} 1 - (1 + \xi\frac{z_t}{\zeta})^{-1/\xi} & \xi \neq 0 \\ 1 - \exp(-\frac{z_t}{\zeta}) & \xi = 0 \end{cases}$$

where $\zeta > 0$, the support is $z_t \leq 0$ when $\xi \leq 0$ and $0 \geq z_t \geq -\frac{\zeta}{\xi}$ when $\xi < 0$. McNeil and Frey (ibid.) further carry out that for a random variable W with an exact GPD distribution with parameter $\xi < 1$ and ζ can be shown that

$$\mathrm{E}\left[W|W > w\right] = \frac{w + \zeta}{1 - \xi},$$

where $\zeta + w\xi > 0$.

It is shown in McNeil and Frey (ibid.) that in case that the excesses of the threshold have exactly this distribution, then it follows that

$$\mathrm{E}\left[Z_t|Z_t > z_{t,\alpha}\right] = z_{t,\alpha}\left(\frac{1}{1-\xi} + \frac{\zeta - \xi u}{(1-\xi)z_{t,\alpha}}\right)$$

with $z_{t,\alpha}$ as the $VaR_{t,\alpha}$, where the t indicates the dependence on time.

3.4.5 Expected Shortfall with Expectiles

The approach to estimate ES with expectiles comes from Taylor (2008). First, it will be introduced what expectiles are and afterwards the connection to ES will be made clear. Differently to Taylor (ibid.) all the expressions will be given for the upper tail of the distribution since the random variable ε_t^{neg} is defined as the negative log returns.

Taylor (ibid.) states that the $(1-\alpha)-$quantile of a random variable ε_t^{neg} is the parameter θ_t that solves the function

$$\min_{\theta_t} \mathrm{E}\left[((1-\alpha) - \mathrm{I}(\varepsilon_t^{neg} > \theta_t))(\varepsilon_t^{neg} - \theta_t)\right].$$

where I is the indicator function.

For the $(1-\tau)$ expectile of ε_t^{neg} θ_t is again the parameter which solves the function

$$\min_{\theta_t} \mathrm{E}\left[|(1-\tau) - \mathrm{I}(\varepsilon_t^{neg} > \theta_t)|(\varepsilon_t^{neg} - \theta_t)^2\right] \tag{9}$$

Following Taylor (ibid.) the parameter θ_t will be defined here as a condition model $\mu_t(1-\tau)$ which shall be estimated by 'asymmetric least squares' (ALS), which serves as the least squares analogue of quantile regression, see Taylor (ibid.). Taylor (ibid.) refers to Newey and Powell (1987) to affirm that the solution $\theta_t = \mu_t(1-\tau)$ from 9 gives us the following equation:

$$\left(\frac{1-2(1-\tau)}{1-\tau}\right) \mathrm{E}\left[(\varepsilon_t^{neg} - \mu_t(1-\tau))\,\mathrm{I}(\varepsilon_t^{neg} > \mu_t(1-\tau))\right] = \mu_t(1-\tau) - \mathrm{E}(\varepsilon_t^{neg})$$

He further states that this expression can be rewritten as

$$E\left[\varepsilon_t^{neg} | \varepsilon_t^{neg} > \varepsilon_{t,1-\alpha}^{neg}\right] = \left(1 + \frac{1-\tau}{(1-2(1-\tau))(1-\alpha)}\right)\mu_t(1-\tau).$$

by defining $F(\mu_t(1-\tau)) = 1-\alpha$ with F as the cdf of ε_t^{neg} and using the fact that $\varepsilon_t^{neg} \sim (0, \sigma_t^2)$.

For the derivation of ES i.i.d. data are assumed because this is implied by a strictly white noise process. To ensure this, it is possible to use a GARCH model to account for the time dependent volatility, as already stated in section 3.4.4. For the definition, see section 3.4.2.

3.4.6 Forecasting

In the end it is also a goal to forecast the CRIX, so that uncertainty about the future can be eliminated. It will be used the class of ARMA(p, q)

models. Following Franke, Härdle, and Hafner (2008), an ARMA(p, q) model is defined as

$$Y_t = \eta + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} + \varepsilon_t$$

with Y_t a random variable, η the intercept, ε the error term, $\alpha_1, \ldots, \alpha_p$ the parameters of the AR part and $\beta_1, \ldots, \beta_q$ the parameters of the MA part.

The optimal model will be found by interpreting the ACF, PACF, looking at the BIC, as defined in 7, and the AIC, see 6.

To compare the forecast, the measures Mean Squared Error (MSE) and Mean Directional Accuracy (MDA) will be used. Chatfield (2001) defines the MSE as

$$MSE = \mathrm{E}[(Y_{t+h} - \mathrm{E}[Y_{t+h}|y_t, y_{t-1}, \ldots])^2] \tag{10}$$

with h as the forecasting period and y as the realizations of Y.

The second one is the Mean Directional Accuracy (MDA). It is defined by Blaskowitz and Herwartz (2011) as

$$MDA = \sum_{i=1}^{n} I(r_y > r_{yh}) \tag{11}$$

with r_y the returns in the forecasting period and r_{yh} the returns of the forecasts.

4 Data

The dataset of the cryptocurrencies is kindly provided by CoinGecko, `https://www.coingecko.com/en`. It consists of 88 cryptos in the time period 2014-06-09 till 2014-12-12. Available are pricing data, market capitalization and trading volume in USD for every crypto. There are also available social media data from `www.reddit.com`. These are counting data for the average number per hour of new posts and comments on the front page in the last 48 hours. In the analysis index data of the S&P500 and the RTSI, euro (USD/EUR) and russian rouble (USD/RUB) pricing data in USD and the US treasury yield data with 1 month to maturity are also used. The last data are all obtained via Datastream with the friendly support of the RDC of the SFB649.

As already noted cryptos are also traded on the weekends and the data from reddit are also available for the weekends. Since this is not the case for the datasets from Datastream, the price for this data from Friday is

	p-value	Shapiro-Wilk distance	λ
1	0.0000	0.4050	0.1000
2	0.0000	0.7228	0.2000
3	0.0000	0.8884	0.3000
4	0.0019	0.9650	0.4000
5	0.0000	0.9396	0.5000
6	0.0000	0.9117	0.6000
7	0.0000	0.8828	0.7000
8	0.0000	0.8534	0.8000

Table 1: Grid searching algorithm for $\lambda \in [0.1, 0.8]$, steps $= 0.1$

set as the value for Saturday and Sunday too. This method is called 'Last Observation Carried Forward', explained in 3.2.1. For one of the cryptos the data for two days are missing. To complete the data the Multiple Imputation method with Bootstrapping is used, see 3.2.2. Since just two data of 152 are missing per time series (pricing data, market capitalization, trading volume, posts and comments) the rate of missingness is rather low. A web search didn't give any reasons to think that these 2 missing data points were related to a special event and therefore not random. The MAR condition is therefore treated as fulfilled. As Honaker, King, and Blackwell (2011) state 5 imputed datasets should be enough to come up with adequate results. But the analysis shows that the variance is very high with just $m = 5$. Increasing the number of imputed datasets to 20 lowered the variance and the results of the measured statistics were very steady. Therefore 20 imputed datasets are used.

As already described in section 3.2.2 the Box-Cox-transformation is used to fulfill normality. The results of the grid searching algorithm are given in the tables 1, 2 and 3. The p-value is the result of the Shapiro-Wilk normality test. The tables show that $\lambda = 0.389$ give the highest p-value but it is still not significant on a level of $\alpha = 0.01$. Anyway is the Box-Cox-Power-Transformation with this value used to make the assumption of multivariate normality as plausible as possible. To calculate the Shapiro-Wilk statistic the function 'mshapiro.test' from the R-package 'mvnormtest' is used. It was also checked if the ln-operator would help to fulfill the normality assumption ($\lambda = 0$) but this was not the case.

5 The Index

In this section is described, how the index is constructed and how the difficulties from section 2.2 are circumvented.

	p-value	Shapiro-Wilk distance	λ
1	0.0000	0.8884	0.3000
2	0.0000	0.8975	0.3100
3	0.0000	0.9058	0.3200
4	0.0000	0.9132	0.3300
5	0.0000	0.9198	0.3400
6	0.0000	0.9257	0.3500
7	0.0000	0.9310	0.3600
8	0.0000	0.9357	0.3700
9	0.0000	0.9399	0.3800
10	0.0030	0.9671	0.3900
11	0.0019	0.9650	0.4000
12	0.0012	0.9627	0.4100
13	0.0008	0.9604	0.4200
14	0.0005	0.9580	0.4300
15	0.0003	0.9554	0.4400
16	0.0002	0.9529	0.4500
17	0.0001	0.9503	0.4600
18	0.0001	0.9476	0.4700
19	0.0000	0.9450	0.4800
20	0.0000	0.9423	0.4900
21	0.0000	0.9396	0.5000
22	0.0000	0.9368	0.5100
23	0.0000	0.9340	0.5200
24	0.0000	0.9311	0.5300
25	0.0000	0.9282	0.5400
26	0.0000	0.9255	0.5500
27	0.0000	0.9229	0.5600
28	0.0000	0.9201	0.5700
29	0.0000	0.9173	0.5800
30	0.0000	0.9145	0.5900
31	0.0000	0.9117	0.6000

Table 2: Grid searching algorithm for $\lambda \in [0.3, 0.6]$, steps $= 0.01$

	p-value	Shapiro-Wilk distance	λ
1	0.0000	0.9399	0.3800
2	0.0000	0.9447	0.3810
3	0.0000	0.9452	0.3820
4	0.0001	0.9456	0.3830
5	0.0001	0.9461	0.3840
6	0.0001	0.9466	0.3850
7	0.0001	0.9470	0.3860
8	0.0001	0.9475	0.3870
9	0.0001	0.9479	0.3880
10	0.0032	0.9674	0.3890
11	0.0030	0.9671	0.3900
12	0.0029	0.9669	0.3910
13	0.0028	0.9667	0.3920
14	0.0027	0.9665	0.3930
15	0.0026	0.9663	0.3940
16	0.0024	0.9661	0.3950
17	0.0023	0.9658	0.3960
18	0.0022	0.9656	0.3970
19	0.0021	0.9654	0.3980
20	0.0020	0.9652	0.3990
21	0.0019	0.9650	0.4000
22	0.0019	0.9647	0.4010
23	0.0018	0.9645	0.4020
24	0.0017	0.9643	0.4030
25	0.0016	0.9641	0.4040
26	0.0015	0.9638	0.4050
27	0.0015	0.9636	0.4060
28	0.0014	0.9634	0.4070
29	0.0013	0.9632	0.4080
30	0.0013	0.9629	0.4090
31	0.0012	0.9627	0.4100
32	0.0012	0.9625	0.4110
33	0.0011	0.9622	0.4120
34	0.0011	0.9620	0.4130
35	0.0010	0.9618	0.4140
36	0.0010	0.9615	0.4150
37	0.0009	0.9613	0.4160
38	0.0009	0.9611	0.4170
39	0.0009	0.9608	0.4180
40	0.0008	0.9606	0.4190
41	0.0008	0.9604	0.4200

Table 3: Grid searching algorithm for $\lambda \in [0.38, 0.42]$, steps $= 0.001$

5.1 Index weighting

Big market indices like $S\&P500$ or $DAX30$ are weighted by market capitalization, see Indices (2014) and AG (2013). It will be taken here the same approach to ensure that the CRIX can be treated as a benchmark for the market in the sense of Laspeyres.

5.2 Reallocation period

The updating of changes in an index happens typically at specific dates. This happens often quarterly. Here the update of the constituent list is performed monthly because the cryptocurrency market is very young and still in a building phase. New cryptos come up weekly and some also vanish fast. It is absolutely logical, that a more frequent updating of the constituent list will help to better track the development of the market. It is then possible to react faster to changes in the market structure.

5.3 Starting date

Due to the available data the starting date of the CRIX is the 2014-07-15 because the first month is necessary to evaluate which cryptos are part of the index and to compute the weightings for the following month.

5.4 Cap

Because the data show that Bitcoin is really a market ruler regarding the market capitalization alternative cryptos won't be adequately represented in the CRIX. Because the trading volume of the cryptos is much higher than their importance due to market capitalization, see e.g. Ong et al. (2015), this would be an underrepresentation. But at the same time it must be ensured that the influence of Bitcoin stays high in the CRIX to not underweight this crypto. Unfortunately doesn't exist to my knowledge a test or method to find the optimal cap.

I decided in the end to choose a cap of 50% for a single crypto to make the CRIX representative for the market.

5.5 Liquidity rule

As described in section 2.2 it is very common to use the free floating assets in the underlying asset universe to define which assets shall be eligible to participate in the index. This approach is not applicable for cryptos because the number of free floating coins of a crypto is not known and also difficult to define. On Coindesk is an article from Tim Swanson online

which shows with charts from John Ratcliff that most of the Bitcoins haven't moved since the origin of Bitcoin, see Swanson (2014). Figure 3 shows a plot which relies on an updated dataset to which the article refers. It is visible that the most of all existing Bitcoins haven't been used frequently. Therefore it can't be assumed that all coins of Bitcoin are free floating and it can be assumed that other cryptos have similar properties. It stays also the question after what time a coin is no longer free floating. Due to the anonymity of cryptocurrencies, it is not clear if a participant wants to hold the cryptos as an investment or if he just hasn't got the possibility to spend them till now. Probably, he will do this the day after the reallocation and his coins were falsely considered as not free floating. Therefore is another but simple procedure necessary. This approach is the following:

$$\frac{1}{T}\sum_{t=1}^{T} MV_{it} \cdot 0.001 \leq \frac{1}{T}\sum_{t=1}^{T} Vol_{it}$$

where MV_{it} indicates the market capitalization of crypto i at timepoint t in a period with length T and Vol stands for trading volume with the same interpretation for i, t and p.

The idea behind this approach is that a large trader shall be able to buy or sell a crypto on an average day. Since the crypto market is much smaller than the large currency markets or share markets, will the definition of a large trader from this well known markets be related to the crypto market. A large trader is defined by the U.S. Securities and Exchange Commission (SEC) regarding volume as a market participant with trading volume of 20 million USD on a calendar day, see `https://www.sec.gov/divisions/marketreg/large-trader-faqs.htm`. In combination with the latest survey report of the Bank for International Settlements (BIS) (2013) result the 0.1% rule. In the BIS survey is the smallest single reported currency the hungarian forint. This survey is some kind of benchmark for the FX market, therefore shall this currency as the smallest important one be used for comparison. It is taken the smallest one because the cryptos are also small currencies compared to the large known markets. Next will be checked how many hungarian forint a large trader would control if he buys a value of 20 million USD in forint. It will be just taken into account the monetary base because for most of the cryptos don't exist any financial vehicles. The BIS survey was published in September 2013, therefore will be taken the monetary base of Hungary for this month converted with the exchange rate 0.0044 HUF/USD, as of 2013-09-04. The monetary base is then 16.13656 billion USD. With the 20 million

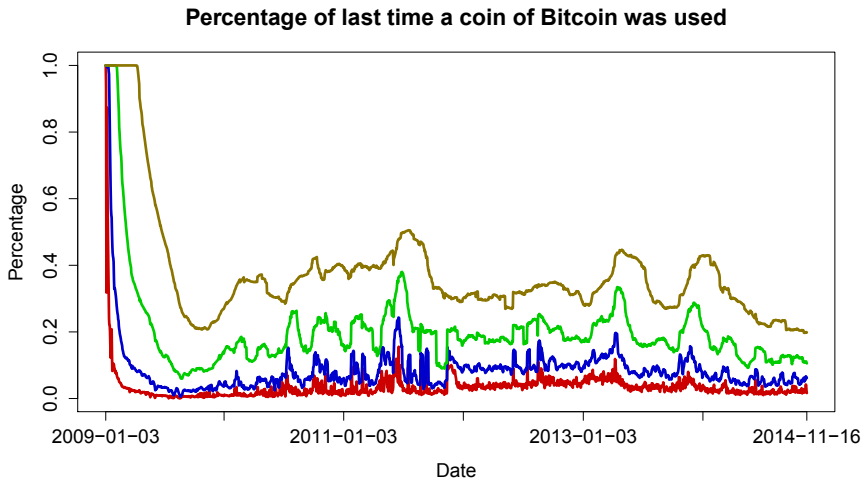


Figure 3: Last use of Bitcoins - datasource: John Ratcliff, one day, one week, one month, 1-3 month

CRIXusagebtc

USD gives this a percentage rate of approximately 0.1%. Therefore would a large trader control 0.1% of the hungarian forint, the smallest single reported currency in this benchmark report. A large trader shall have at least a similar influence in a crypto in the CRIX. The definition of a large trader is changed for this market to a relative one to take into account that this market can grow in value and bigger market players could catch in. Also shall this definition give weight to the reasonable assumption that in bigger cryptos like Bitcoin would more money be invested than in a small one. Therefore, for a small crypto with less market value would be a trader or investor much faster be a large one, regarding investment amount, as it would be for the large cryptos.

The used percentage rate will be therefore 0.1%. This approach ensures that it was possible in the last period for a trader or investor to buy or sell on an average day 0.1% of the entire market value of a crypto.

This is a very crucial restriction for the CRIX. A crypto which doesn't fulfill the requirement in a time period is not eligible to be part of the index and will be excluded from determining the index constituents.

5.6 Number of Index constituents

For the number of index constituents will be used, as already noted in section 2.2, an approach based on comparing the AIC and BIC of indices with different amounts of constituents against an index without any liquidity rule or restriction to the number of constituents. This approach is chosen because an index is dedicated to be a market proxy. It should rebuild the market movement as good as possible, but including all the

market participants into the index would cause it to become too complex. AIC and BIC were developed to reach the goal of finding a model which fits the data as good as possible while using as few parameters as possible by adding a penalty term. To reach this goal for the CRIX, indices with a number of 10, 15, 20, 25 cryptos will be builded and tested against the full market. The full market are all cryptos in the dataset which are combined by market capitalization weighting and with the cap rule from section 5.4. The following model will be used to compare the indices:

$$Y = X + \varepsilon$$

with Y the full market, X the index and ε the residual term. X can be interpreted as a proxy for Y because Y includes by definition all the parameters in X. Therefore would enhance including more index constituents (parameters) into X the fit to Y. Here the AIC and BIC catch in which are dedicated to find an index for which the ε will be minimized while as less index constituents as possible are included. Elsewhere all cryptos would be the perfect model and the index wouldn't reduce the dimensionality.

By analyzing the different time series of the cryptos it became obvious that the value of the crypto Ripple (xrp) really went places while the observed time period, see plot 5.6. This massive increase in value can distort the result. Therefore, the difference between the entire market and the index twice will be analyzed. One time for the full time period and the second time without the last 3 weeks to kick out the effect of the xrps value gaining.

The figure 5.6 indicates that increasing the number of constituents, the index approaches better to whole market. For the distribution of the residuals between the two time series the wide class of stable distributions, described in section 3.3, will be taken into account. As already stated, the maximum likelihood method gives the best results in estimating the parameters of a stable distribution. Here the function 'stableFit' from the R-package 'fBasics' will be used, see Wuertz et al. (2014). The corresponding results for the indices are displayed in table 5.6 and the corresponding density plots are given in figure 5.6. For the comparison, random variables with the estimated stable parameters were created. One time as much as the sample has residuals ($N = 152$) and one time $N = 1000$. The plots show that the estimated distribution fits the observed density quite well. The bandwidth was automatically calculated for each curve with a rule-of-thumb method which uses a gaussian kernel. The same analysis was performed for the shorter time series. The corresponding

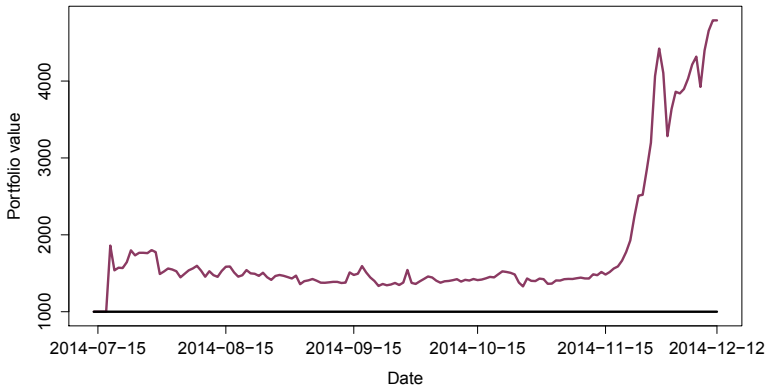


Figure 4: xrp

const.	alpha	beta	gamma	delta
10	1.0802	-0.5294	1.8325	-6.5880
15	1.0603	-0.6694	1.1667	-5.7474
20	0.9882	-0.1968	0.8273	-4.9448
25	1.1703	-0.3657	0.9562	-4.5690

Table 4: Parameters of the stable distributions, const. = number of constituents, full time series

results are displayed in plot 5.6 and table 5.6. This time is $N = 131$. To test if the estimated models fit the data, the Kolmogorov-Smirnov test, described in section 3.3, is used. The resulting p-values are given in table 5.6. They are computed based on a Monte-Carlo simulation with 1000 replications. The results show that the distributions fit the data.

The estimated distributions were then used to compute the AIC and BIC of the 8 models to decide in the end how many index constituents to use. As stated in Feigelson and Babu (2012), the AIC can be used for different families of probability distributions, what can be transfered to the BIC because the BIC is nothing else than the AIC with a stronger penalty for higher parameterized models if $\ln(n) > 2$. Therefore, different distributions can be applied to the 8 indices. The AICs of the indices against the entire market are given in table 5.6 and the corresponding BICs in table 5.6. In case of the full time series both information criterion are minimized for the model with 25 members. But this are clearly unnormal market conditions. The higher difference between the time series comes from the early explained high returns of xrp. It can be assumed that this won't happen every day. The AICs and BICs for the shorter time series show that the model with 20 parameters shall be preferred. Therefore, the model with 20 parameters will be chosen so that the CRIX won't overfit most of the time the market.

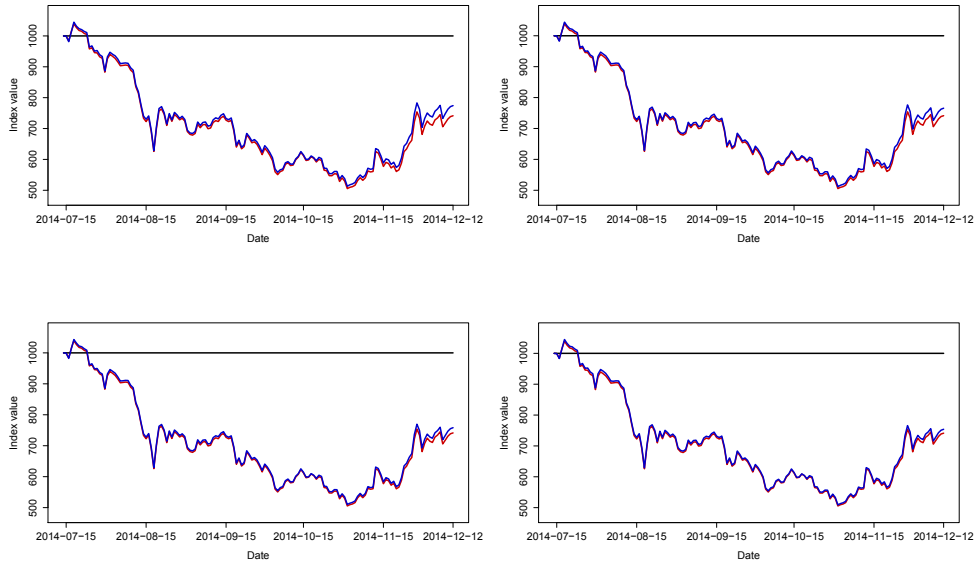


Figure 5: 4 times indices with 10, 15, 20 and 25 constituents against the full market, from upper left to right bottom, whole market, index

CRIXindicests

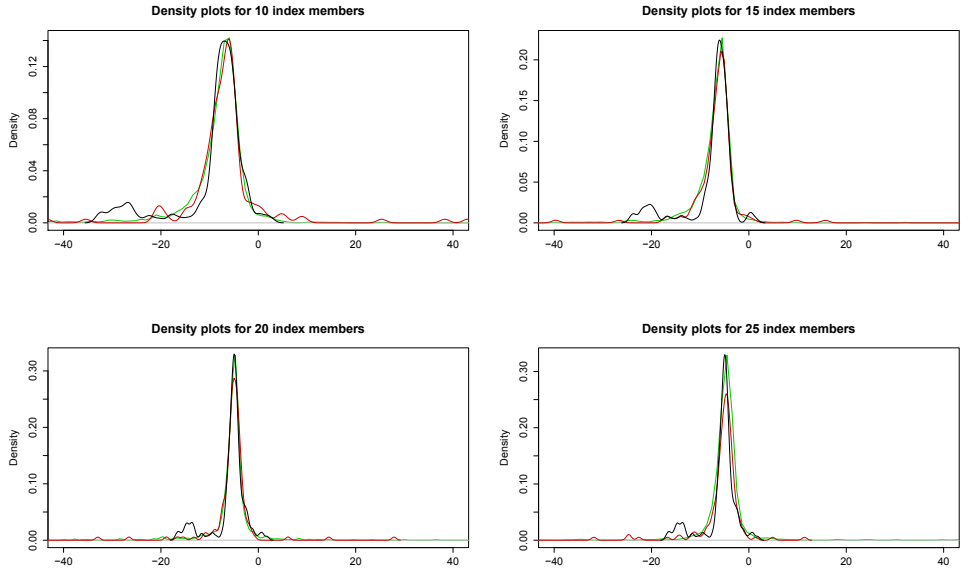


Figure 6: 4 density plots with 10, 15, 20 and 25 constituents against the full market, from upper left to right bottom, N = 1000 random stable observations, N = 152 random stable observations, N = 152 residuals from model

CRIXdensityfts

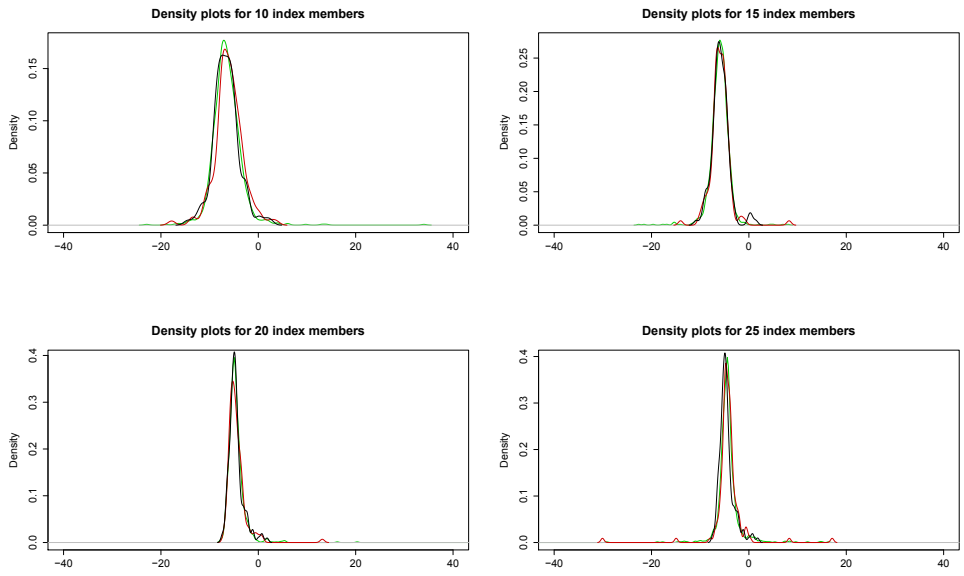


Figure 7: 4 density plots with 10, 15, 20 and 25 constituents against the full market, from upper left to right bottom, N = 1000 random stable observations, N = 131 random stable observations, N = 131 residuals from model

CRIXdensitysts

const.	alpha	beta	gamma	delta
10	1.6904	0.2848	1.5523	-6.7731
15	1.6482	0.0348	0.9682	-5.9355
20	1.5914	0.9999	0.7702	-5.0395
25	1.4130	0.3261	0.7524	-4.5789

Table 5: Parameters of the stable distributions, const. = number of constituents, short time series

number const.	10	15	20	25
long ts	958.77	852.69	802.45	802.10
short ts	666.74	578.58	540.57	594.00

Table 6: BIC for index (long ts) and without last 3 weeks (short ts), different number of constituents

number const.	10	15	20	25
long ts	928.53	807.33	741.97	726.50
short ts	637.98	535.45	483.07	522.12

Table 7: AIC for index (long ts) and without last 3 weeks (short ts), different number of constituents

	10	15	20	25
full ts	0.62	0.63	0.54	0.79
short ts	0.97	0.86	0.78	0.80

Table 8: Kolmogorov-Smirnov test for the stable distributions of the indeces with the full time series and without the last 3 weeks (short ts), different number of constituents

5.7 Market Capitalization rule

Cryptos which fulfill the foregoing liquidity rule, see section 5.5, are eligible and they will be ordered by the value of their market capitalization. It will be chosen the 20 cryptos with the highest market capitalization.

5.8 Special events

1. In case that a crypto stops being traded on all exchanges while it is part of the index, it will be replaced with the next crypto depending on the ordering.

2. If a crypto doesn't exist any longer, it will be also replaced with the next crypto from the ordered list.

5.9 The base value and divisor

The base value of the CRIX will be set to 1000. To achieve this value the formula with the related AWF_i for each constituent i will be used, see equation 3

$$Index\ value = \frac{\sum_{i=1}^{20} MV_i \cdot AWF_i}{Divisor}$$

with $Index\ value = 1000$ on the starting date.

Because the amount of coins of a crypto changes every day while the mining process is still active, it is necessary to adjust the divisor every day to account for this change in the index value. For this, the formula will be used 2.

5.10 Index constituents

The index members over the 5 periods are given in the table 9. It is shown e.g. the gain in influence of xrp or doge (Dogecoin) and how other cryptos like drk (Darkcoin) lose in importance in the CRIX. Bitcoin is of course over the entire time the most important crypto. The reason why bts (BitShares) suddenly is the fourth important crypto, is that there doesn't exist any data before this date for this crypto.

	1	2	3	4	5	6	7	8	9	10
1	btc	ltc	nxt	drk	ppc	xrp	doge	nmc	bc	xcp
2	btc	ltc	nxt	xrp	drk	ppc	doge	nmc	bc	xcp
3	btc	ltc	xrp	bts	nxt	ppc	doge	drk	nmc	xmr
4	btc	xrp	ltc	bts	doge	nxt	ppc	drk	nmc	xcp
5	btc	xrp	ltc	bts	doge	nxt	ppc	xcp	drk	nmc
	11	12	13	14	15	16	17	18	19	20
1	xc	xmr	zet	qrk	vtc	xpm	ftc	mec	ifc	pot
2	xmr	xc	rdd	vrc	qrk	zet	xpm	ftc	ifc	mec
3	xcp	xc	bc	bcn	str	mona	rdd	qrk	zet	vrc
4	xmr	bc	str	xc	bcn	qrk	xpm	ftc	rdd	zet
5	str	xmr	bc	bcn	ftc	cann	qrk	xc	pnd	zet

Table 9: Index members in the 5 periods, ordered by influence

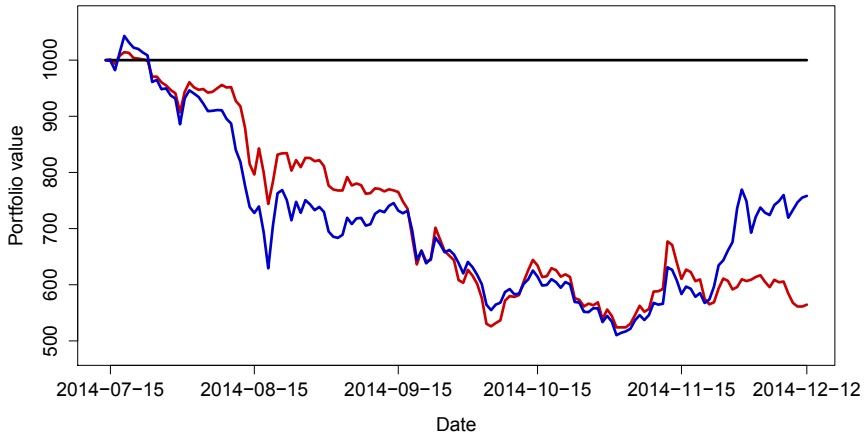


Figure 8: Performance of a btc portfolio btc against CRIX

CRIXinbtc

6 Bitcoin against CRIX

After constructing the CRIX it is now possible to compare Bitcoin (btc) against the market, namely CRIX. This one is picked because it is still the most important one in the entire crypto universe. The market capitalization is the highest and the trading volume also as the data from CoinGecko showed and Ong et al. (2015). A portfolio with the investment value of 1000 USD is constructed for btc to compare it against CRIX. The performance is visualized in figure 8.

The figure 8 - the black line is the 1000-value line - shows that the btc moves with the CRIX. This is partly not surprising since btc has an influence of 50% in the CRIX because of its very high market capitalization. While August till mid of September the btc portfolio had a higher value which shows that btc outperformed other cryptos. In the end period, CRIX was much higher than the btc portfolio. This movement is most probably caused by the massive gain in value of Ripple.

An analysis of the returns, variance and sharpe ratio, see table 6, shows that the CRIX lost in the end less value, which was already obvious from the figure 8. But the variance of the CRIX is much higher than that for btc, which is partly caused by the effect due to the MIB algorithm. Due to the additional variance between the 20 datasets, the variance is much higher, therefore, the variance is also given without the 2 days when missing values occurred. But the variance of the CRIX does still higher. By taking into account the US treasury rate with 1 month to maturity as secure investment is the Sharpe Ratio given as last indicator in table 6. It shows that btc was in this time period the less risky investment compared to CRIX. Even when more value was gained in the end with CRIX, an risk-averse investor should have chosen the btc portfolio instead of an investment into the CRIX.

	CRIX	btc
Return	0.7582	0.5645
Var	24.7681	4.5643
Var wo. missing value effect	4.6470	4.4993
Sharpe Ratio	-0.0032	-0.0018

Table 10: Comparison of CRIX and btc

7 CRIX against other markets

Cryptos are a new type of investment but it is unclear how risky this market really is. In this section, the risk shall be compared by looking at the tails of the distribution of the CRIX log returns and the log returns of other investments. Table 7 shows the Expected Shortfall derived with the Extreme Value (EVT) and TERES approach as described in the sections 3.4.4 and 3.4.5 for $\alpha = 0.01$ and a threshold of 0.1 for the full time period. Both approaches are used because EVT is significantly unbiased for small samples and TERES is significantly biased but more efficient. The results for the two other currencies show that the CRIX market is much more risky. The exchange rate USD/EUR is chosen as a relatively stable one and the results show that this exchange rate is in the case of TERES up to 7 times less risky than CRIX. Even for a riskier currency rate like USD/RUB the crypto market is up to 4 times riskier. Both approaches point into this direction, therefore can be concluded that CRIX is much more riskier than well known currencies. Therefore, it is kind of impossible to think about cryptos as some kind of currency in the risk perspective. The comparison with 2 share indices showed that CRIX is also riskier than them but the difference is smaller. By looking

at TERES, it is obvious that S&P500 and RTSI are round about half as risky as CRIX. For the EVT approach the difference is less large but CRIX still being riskier. It can be concluded, that CRIX is even riskier than this 2 share indices.

When investing into the crypto market, it must be taken into account that we deal here with a market which is much riskier than other investment classes. So thinking about it like a currency in the risk manner gives a wrong picture of this market.

	EVT	TERES
CRIX	-0.0702	-0.0931
USD/EUR	-0.0186	-0.0126
USD/RUB	-0.0285	-0.0230
S&P500	-0.0501	-0.0454
RTSI	-0.0404	-0.0421

Table 11: Expected Shortfall of different investments, full time period, $\alpha = 0.01$, threshold $= 0.1$

8 The investor view

This section is dedicated to analyze possibilities to invest better into the crypto market e.g. with an optimized portfolio or with forecasting the CRIX to decrease uncertainty about the future.

8.1 MVCRIX

As stated in MSCI (2012) minimum-variance strategies gained popularity in the last years. This approach uses the idea of Markowitz to optimize a portfolio by assigning weights to the constituents in a way that the overall variance of the entire portfolio is minimized, see Markowitz (1952).

The used procedure is the one from MSCI (2012). The constituents of CRIX are taken and just the weights are optimized so that the overall variance is minimized. This procedure is performed always for the last period and the weights are then used for the next period, like it is the case for the CRIX. Different to the rules in MSCI (ibid.) aren't used any borders for the weights of the Minimum Variance CRIX (MVCRIX). The variance is measured with a GARCH(1,1) model, as it is described in section 3.4.2, for the log returns of the index. Because there are too less observations to build a reliable covariance matrix for all the index constituents, the optimization is performed numerically with the optimizer 'solnp' from the R-package 'Rsolnp'. It's a function for nonlinear

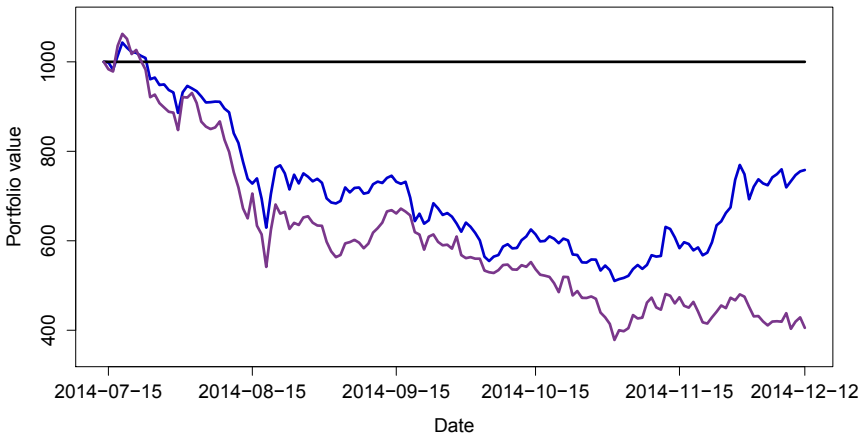


Figure 9: Performance of CRIX and MVCRIX

CRIXmvcrix

	CRIX	MVCRIX
Return	0.7582	0.4053
Var	24.7681	5.6955
Var wo. missing value effect	4.6470	5.6216
Sharpe Ratio	-0.0032	-0.0087

Table 12: Comparison of CRIX and MVCRIX

optimization using the augmented Lagrange method. The formula for the MVCRIX is again 4 but CW_i from equation 3 is replaced with the weights from the minimization of the variance.

The plot 9 visualizes the performance of the MVCRIX against the CRIX. It is obvious that the MVCRIX follows the direction of its parent index, the CRIX, but with a higher loss. The table 8.1 shows the return, variance and sharpe ratio for the 2 time series. The absolute return for the CRIX is much higher as for the optimized portfolio, the MVCRIX. The variance is lower but just the one with the effect due to the MIB approach, see formula 5. By excluding this effect (excluding variance from days with missing values) the MVCRIX shows a higher variance even it is dedicated to have a lower variance. Of course this can happen when the past volatility doesn't tell us much about future volatility. Following this findings the sharpe ratio for the CRIX is better than for the MVCRIX, taking into account the US Treasury rate with 1 month to maturity. This result shows us, that it would be better for an investor to invest into the market instead of an optimized portfolio. If the uncertainty in this investment can be decreased shall be checked in the following section by trying to forecast the movement of the CRIX.

8.2 Forecasting

To check if the CRIX can be well forecasted, the last 2 weeks will be taken for an out-of-sample forecast. The remaining data are used for training the model. To find a proper model, the ACF and PACF will be taken into account. The ACF, given in plot 8.2, shows that the data has an autoregressive structure. The PACF, see plot 8.2, makes obvious that there is a strong relationship in the first lag and insignificant relations to the remaining lags. Following the visual analysis an AR(1) model would be a proper choice for a time series model. Due to this result models around an AR(1) model were estimated using the function *arma* from the R-package *tseries*. The results are given in the tables 13, 14, 15 and 16. The last table shows that the AIC suggests an ARMA$(1, 1)$ model but the BIC advises an ARMA$(1, 0)$. The significance of the parameters show that the p-value of the MA parameter is at a significance level of 0.05 and including an intercept into the AR(1) model gives a p-value, which is insignificant on a 5% level. Therefore and because the PACF tends to an AR(1) model and the ACF shows no signs of a MA part an AR(1) model will be chosen.

Like e.g. the work of Zhang et al. (2015) showed, news do have an influence on the returns of stocks. Therefore, it can also be, that cryptocurrencies react to news. Counting data are obtained from the social media platform reddit, `http://www.reddit.com/`, for posts and comments about cryptocurrencies. The available data are in both cases counting data which show how many post/comments appeared on the front page of reddit in the last 48 hours on a hourly average. To get the posts for the CRIX, the weights of the index constituents were taken and applied to the corresponding posts and comments. The models for forecasting are regression models, one with the posts and one with the comments. The resulting models are

$$y_t = \beta_{posts} x_{1,t} + \varepsilon_t \tag{12}$$

and

$$y_t = \beta_{comments} x_{2,t} + \varepsilon_t$$

with y_t the CRIX value at time point t, $x_{1,t}$ and $x_{2,t}$ the weighted data for posts and comments respectively, β_{posts} and $\beta_{comments}$ the corresponding parameters and $\varepsilon_t \sim \mathrm{N}(\mu, \sigma^2)$.

To compare the forecasting models the MSE and MDA criterion will be used, as given in 10 and 11, and the results are given in table 17. It is clearly shown, that the AR(1) model is much better than the other two

	Estimate	Std. Error	t value	Pr(>|t|)
ar1	1.00	0.00	378.11	0.00

Table 13: Summary AR(1)

	Estimate	Std. Error	t value	Pr(>|t|)
ar1	0.97	0.01	74.59	0.00
intercept	17.14	9.40	1.82	0.07

Table 14: Summary AR(1) with intercept

possible models. The MSE is smaller and the MDA is the best together with the one for the model 12. A good sign is, that the MDA is for all three models above 0.5. This shows that the models all point often in the correct direction. But AR(1) has the lowest MSE, therefore this is the best in this comparison.

Next, it shall be checked if combined forecasting models give a more accurate forecast. For this, the following three models will be applied:

$$y_t = \alpha y_{t-1} + \beta_{posts} x_{1,t} + \varepsilon_t, \tag{13}$$

$$y_t = \alpha y_{t-1} + \beta_{comments} x_{2,t} + \varepsilon_t \tag{14}$$

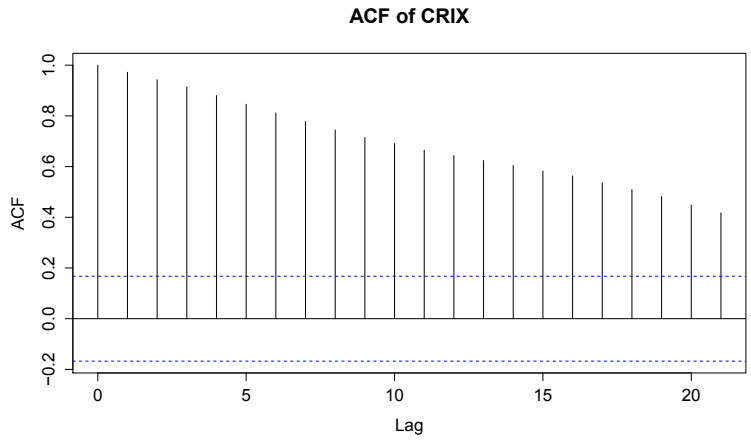
and

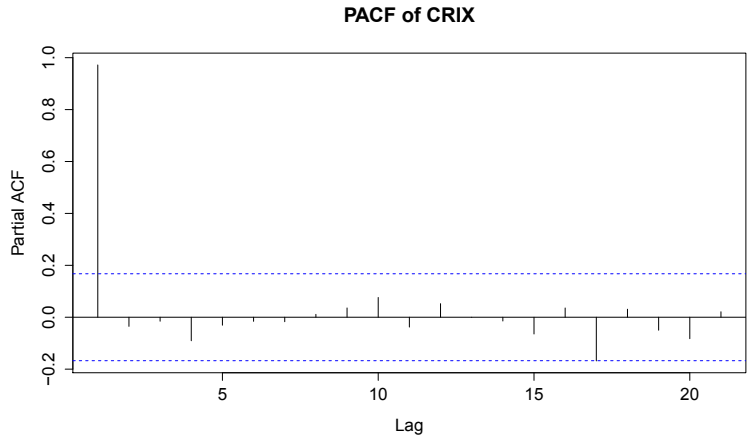
$$y_t = \alpha y_{t-1} + \beta_{posts} x_{1,t} + \beta_{comments} x_{2,t} + \varepsilon_t \tag{15}$$

with y_t the CRIX value at time point t, $x_{1,t}$ and $x_{2,t}$ the weighted data for posts and comments respectively, β_{posts} and $\beta_{comments}$ the corresponding parameters, α the AR parameter and $\varepsilon_t \sim \mathrm{N}(\mu, \sigma^2)$.

The results of the estimation for the models 13 and 14 are given in the tables 18 and 19. The summaries show that the AR parameter remains significant and in both regressions the parameters for the posts and comments are significant at a 5% level. The third model, given in 15, shows that combining the AR part with both social media measures gives insignificant parameter values for the posts and comments. Therefore, this model is not useful for the comparison.

The results of the forecasting are given in table 21. Obviously accuracy is gained by combining the models. In both cases the MSE is much lower and the MDA stays the same. The best model in this comparison is the one with the AR(1) parameter and the posts parameter. This result shows that it is possible for the cryptocurrency market to gain forecasting value by adding social media data. Interesting would be if the forecasts can be improved by having more accurate data at hand.





	Estimate	Std. Error	t value	Pr(>|t|)
ar1	1.00	0.00	323.67	0.00
ma1	0.19	0.09	1.98	0.05

Table 15: Summary ARMA(1,1)

	AR(1)	AR(1) & intercept	ARMA(1,1)
AIC	1241.28	1240.12	1239.56
BIC	1244.20	1245.96	1245.4

Table 16: AIC and BIC for different forecasting models

	MSE	MDA
AR(1)	523.21	0.62
Regression posts	4970.24	0.62
Regression comments	6931.54	0.54

Table 17: MSE and MDA of different models in forecasting comparison

	Estimate	Std. Error	t value	Pr(>|t|)
α	0.97	0.01	69.53	0.00
β_{posts}	15.70	7.61	2.06	0.04

Table 18: Summary Regression AR(1) & posts

	Estimate	Std. Error	t value	Pr(>|t|)
α	0.98	0.01	112.45	0.00
$\beta_{comments}$	0.16	0.06	2.43	0.02

Table 19: Summary Regression AR(1) & comments

	Estimate	Std. Error	t value	Pr(>|t|)
α	0.97	0.01	69.74	0.00
β_{posts}	6.35	10.06	0.63	0.53
$\beta_{comments}$	0.12	0.09	1.41	0.16

Table 20: Summary Regression AR(1) & posts & comments

	MSE	MDA
AR(1)	523.21	0.62
AR(1) & posts	506.03	0.62
AR(1) & comments	775.70	0.62

Table 21: MSE and MDA of the combined models in forecasting comparison

9 Conclusion

This master thesis showed the construction of a market index for the cryptocurrency market, the CRIX. A new approach was used to find the optimal number of index constituents which showed that this index is indeed useful to track the crypto market. The CRIX was then used for a market comparison. The results of the risk analysis showed that the crypto market is much more risky than known markets, especially it is impossible to see it as a currency when taking into account the risk behavior. It was also shown that the market performed in the considered time period better than Bitcoin, which is still the most important crypto. The forecasting comparison in the end showed that social media data can be used to forecast the price movements of the CRIX. It would be interesting to have better data at hand to analyze which information in detail push the market.

References

AG, Deutsche Börse (2013). *Guide to the Equity Indizes of Deutsche Boerse*.

Akaike, Hirotogu (1998). *Information Theory and an Extension of the Maximum Likelihood Principle*. English. Ed. by Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa. Springer Series in Statistics. Springer New York, pp. 199–213. ISBN: 978-1-4612-7248-9. DOI: 10.1007/978-1-4612-1694-0_15. URL: http://dx.doi.org/10.1007/978-1-4612-1694-0_15.

Artzner, Philippe et al. (1999). "Coherent Measures of Risk". In: *Mathematical Finance* 9.3, pp. 203–228. ISSN: 1467-9965. DOI: 10.1111/1467-9965.00068. URL: http://dx.doi.org/10.1111/1467-9965.00068.

Blaskowitz, Oliver and Helmut Herwartz (2011). "On economic evaluation of directional forecasts". In: *International Journal of Forecasting* 27.4, pp. 1058–1065. ISSN: 0169-2070. DOI: http://dx.doi.org/10.1016/j.ijforecast.2010.07.002. URL: http://www.sciencedirect.com/science/article/pii/S0169207011000057.

Bollerslev, Tim (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3, pp. 307–327.

Box, George EP and David R Cox (1964). "An analysis of transformations". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 211–252.

Chatfield, Chris (2001). *Time-series forecasting*. CRC Press.

Cizek, Pavel, Wolfgang Karl Härdle, and Rafał Weron (2011). *Statistical tools for finance and insurance*. Springer Science & Business Media.

Enders, Craig K (2010). *Applied missing data analysis*. Guilford Publications.

Feigelson, Eric D and G Jogesh Babu (2012). *Modern statistical methods for astronomy: with R applications*. Cambridge University Press, p. 62.

Franke, Jürgen, Wolfgang Karl Härdle, and Christian Matthias Hafner (2008). *Statistics of financial markets: an introduction*. Springer Science & Business Media.

Gschöpf, Philipp (2014). "Measuring risk with expectile based expected shortfall estimates". MA thesis. URL: %5Curl%7Bhttp://edoc.hu-berlin.de/docviews/abstract.php?id=40856%7D.

Honaker, James and Gary King (2010). "What to do About Missing Values in Time Series Cross-Section Data". In: *American Journal of Political Science* 54, pp. 561–581. URL: http://gking.harvard.edu/files/abs/pr-abs.shtml.

Honaker, James, Gary King, and Matthew Blackwell (2011). "Amelia II: A Program for Missing Data". In: *Journal of Statistical Software* 45.7, pp. 1–47. ISSN: 1548-7660. URL: http://www.jstatsoft.org/v45/i07.

Indices, S&P Dow Jones (2014). *Index Mathematics - Methodology*.

Lippe, Peter von der (2013). "Recurrent Price Index Problems and Some Early German Papers on Index Numbers". In: *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)* 233.3, pp. 336–366.

Liu, Shouwei and Yiu Kuen TSE (2013). "Estimation of monthly volatility: An empirical comparison of realized volatility, GARCH and ACD-ICV methods". In:

Markowitz, Harry (1952). "PORTFOLIO SELECTION*". In: *The Journal of Finance* 7.1, pp. 77–91. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.1952.tb01525.x. URL: http://dx.doi.org/10.1111/j.1540-6261.1952.tb01525.x.

McNeil, Alexander J and Rüdiger Frey (2000). "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach". In: *Journal of empirical finance* 7.3, pp. 271–300.

MEXBOL (2013). *Prices and Quotations Index (MEXBOL) - Methodology Note*.

MSCI (2012). "MSCI Global Minimum Volatility Indices Methodology". In:

Newey, Whitney K. and James L. Powell (1987). "Asymmetric Least Squares Estimation and Testing". English. In: *Econometrica* 55.4, ISSN: 00129682. URL: http://www.jstor.org/stable/1911031.

Nolan, J. P. (2015). *Stable Distributions - Models for Heavy Tailed Data*. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan. Boston: Birkhauser.

Ong, Bobby et al. (2015). "Evaluating the Potential of Alternative Cryptocurrencies". In: *Handbook of Digital Currency*.

Razali, Nornadiah Mohd and Yap Bee Wah (2011). "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests". In: *Journal of Statistical Modeling and Analytics* 2.1, pp. 21–33.

Royston, J. P. (1983). "Some Techniques for Assessing Multivarate Normality Based on the Shapiro- Wilk W". English. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 32.2, ISSN: 00359254. URL: http://www.jstor.org/stable/2347291.

Royston, Patrick (1992). "Approximating the Shapiro-Wilk W-Test for non-normality". In: *Statistics and Computing* 2.3, pp. 117–119.

Schwarz, Gideon (1978). "Estimating the Dimension of a Model". In: *Ann. Statist.* 6.2, pp. 461–464. DOI: 10.1214/aos/1176344136. URL: http://dx.doi.org/10.1214/aos/1176344136.

Shapiro, S. S. and M. B. Wilk (1965). "An Analysis of Variance Test for Normality (Complete Samples)". English. In: *Biometrika* 52.3/4, ISSN: 00063444. URL: http://www.jstor.org/stable/2333709.

Sharpe, William F. (1966). "Mutual Fund Performance". English. In: *The Journal of Business* 39.1, ISSN: 00219398. URL: http://www.jstor.org/stable/2351741.

Swanson, Tim (2014). "Analysis: Around 70 % of Bitcoins Unspent for Six Months or More". In: URL: http://www.coindesk.com/analysis-around-70-bitcoins-dormant-least-six-months/.

Szado, Gitanjali M. Swamy; Irina Zeltser; Hossein Kazemi; Edward (2012). "Setting the Benchmark: Spotlight on Private Equity". In: *Alternative Investment Analyst Review, Volume 1, Issue 1*.

Taylor, James W (2008). "Estimating value at risk and expected shortfall using expectiles". In: *Journal of Financial Econometrics* 6.2, pp. 231–252.

Wuertz, Diethelm et al. (2014). *fBasics: Rmetrics - Markets and Basic Statistics*. R package version 3011.87. URL: `http://CRAN.R-project.org/package=fBasics`.

Zhang, Junni L et al. (2015). *Distillation of News Flow into Analysis of Stock Reactions*. Tech. rep. Sonderforschungsbereich 649, Humboldt University, Berlin, Germany.

Declaration of Authorship

I hereby confirm that I have authored this Master's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, March 30 2015

Simon Trimborn