**Title: Estimating future Bitcoin profitability**

Author: Trenton Potgieter

Date: 10/19/2014

**Introduction:**

Bitcoin is software that tracks and verifies transactions on a public ledger over a peer-to-peer network. Operations and data associated with Bitcoin are decentralized, meaning they are not performed or stored in one single location. Instead, the Bitcoin network consists of computers across the world that automatically store and relay Bitcoin data to each other. The computers' owners voluntarily choose to use and run the Bitcoin software. Anyone can use and run Bitcoin software. Bitcoins are produced through a process called mining. Mining is the competitive use of computational power to calculate a number that falls within a certain range. The first miner to discover a number “target” [1] that meets the criteria is rewarded with a set amount of brand new bitcoins (currently 25 bitcoins), the competition then repeats for the discovery of a new number [2]. In other words, when the number in the block header is equal or lower than the “target”, the block is accepted by the network and the process to create a new block begins.

Understanding just how difficult [3] it is to discover the given number below the given target is the key to determining profitability, or measure of success. More importantly is the ability to estimate future mining profitability by predicting what the network difficulty will be. This paper describes the process of forecasting what the network difficulty will be for a two year period and calculating how the profitability based on the current hashrate of the mining hardware.

**Methods:**

*Data Collection*

The data used for working with the current and previously recorded Network statistics (including Difficulty) is downloaded directly using the R programming language [4]. from the Bitcoin Network by using the Bitcoin Block Explorer website [5] which is an open source web tool that allows a user to view information about the [blocks](https://en.bitcoin.it/wiki/Blocks), [addresses](https://en.bitcoin.it/wiki/Address), and [transactions](https://en.bitcoin.it/wiki/Transactions) created by [Bitcoin](http://bitcoin.org/).

<INSERT R CODE>

*Exploratory Analysis*

Each row within the data is a log containing the following:

* Time when the block was created (UTC)
* Decimal target
* Difficulty
* The average number of hashes it takes to solve a block at this difficulty

For the process of analysis, Time and Difficulty values are extracted. Any duplicated values are removed and the data is converted from UTC to a standard time format for easier sub setting into monthly periods.

<INSERT R CODE>

*Time-Series Modeling*

To model the predicted forecast of the network difficulty, the Arima model is used. Using the auto.arima() function in R uses a variation of the [Hyndman and Khandakar [6] algorithm](http://robjhyndman.com/papers/automatic-forecasting/) which combines unit root tests, minimization of the AICc and MLE to obtain an ARIMA model. The algorithm follows these steps: [7]

1. The number of differences **d** is determined using repeated KPSS tests.
2. The values of **p** and **q** are then chosen by minimizing the AICc after differencing the data **d** times. Rather than considering every possible combination of **p** and **q** , the algorithm uses a stepwise search to traverse the model space.
   1. The best model (with smallest AICc) is selected from the following four: ARIMA(2,d,2), ARIMA(0,d,0), ARIMA(1,d,0), ARIMA(0,d,1). If **d=0** then the constant c is included; if d≥1 then the constant c is set to zero. This is called the "current model".
   2. Variations on the current model are considered:  
        
      vary **p** and/or **q** from the current model by ±1 ;  
      include/exclude c from the current model.  
        
      The best model considered so far (either the current model, or one of these variations) becomes the new current model
   3. Repeat Step 2(b) until no lower AICc can be found.

<INSERT R CODE>

**Results:**

The earthquakes data used in this analysis contains information on the source network that measured the earthquake (Src), the time of the earthquake (TME), the number of sites that measured the earthquake (NST), the location - longitude (Lon) and latitude (Lat), the magnitude (EM, measured on the Richter scale) and the depth (ED, measured in kilometers from the surface) [7]. We identified no missing values in the data set we collected and all measured variables were observed to be inside the standard ranges. Earthquakes in this data set also did not seem to show major patterns over time in magnitude or depth.

Most earthquakes had a small (less than 3 on the Richter scale - 85% of quakes) or medium (3-5 on the Richter scale - 11% of quakes) magnitude. The distribution of earthquake depths was heavily right skewed. Based on the distribution of the earthquake depths we recognized that a transform was necessary to improve the performance of linear regression techniques; we performed a log base 10 transform of the earthquake depths. Subsequent analyses focus on this transformed depth variable.

We first fit a regression model relating earthquake magnitude to earthquake depth. The residuals showed patterns of non-random variation. We attempted to explain those patterns by fitting models including potential confounders. Our final regression model was:

EM = b0 + b1 log10(ED) + f(Lat) + g(Lon) + h(NST) + e

where b0 is an intercept term and b1 represents the change in earthquake magnitude on the Richter scale associated with a change of 1 unit in log base 10 kilometers for earthquakes at the same latitude, longitude, and measured by the same number of stations. The terms f(Lat), g(Lon), and h(NST) represent factor models with 5 different levels each for latitude, longitude, and number of measurement sites. The error term e represents all sources of unmeasured and unmodeled random variation in earthquake magnitude. Our final regression model appeared to remove most of the non-random patterns of variation in the residuals.

We observed a highly statistically significant (P = 8e-15) association between earthquake magnitude and earthquake depth. A change of one unit in log base 10 kilometers of depth corresponded to a change of b1 = 0.41 on the Richter scale (95% Confidence Interval: 0.31, 0.51). So for example, for two earthquakes at the same latitude, longitude, and measured at the same number of stations, we would expect an earthquake measured at a depth of 10 kilometers to be 0.57 units higher on the Richter scale than one measured at a depth of 1 kilometer. We would expect the same difference between earthquakes measured at depths of 10 and 100 kilometers and so forth.

**Conclusions:**

Our analysis suggests that there is a significant, positive association between earthquake magnitude and earthquake depth. Our analysis estimates the relationship using a linear model relating log base 10 depth to magnitude. Therefore, our estimated association is not linear on the depth scale. Even so, there appears to be a strong relationship between the two variables. We also observed that other variables such as latitude, longitude and number of observing stations are associated with both earthquake magnitude and earthquake depth. Including these variables in the regression model relating magnitude to depth improves the model fit, but does not remove the significant positive relationship between the variables.

While our analysis is an interesting first step it is based on a limited sample of earthquakes from only one week in January 2013. A larger collection of representative earthquakes may be more appropriate for understanding the relationship between earthquake magnitude and depth. Our analysis may be of interest to scientists seeking to better understand earthquakes. But for policy makers, an important - and difficult - area of future research would be to develop accurate predictive models of earthquake depth and magnitude that can be used to evaluate the potential human consequence of these natural disasters.

**References**

1. Target page. URL: <https://en.bitcoin.it/wiki/Target>. Accessed 10/19.2014
2. How Bitcoin works page. URL: <https://bitcoinhelp.net/know/how-bitcoin-works>. Accessed 10/19/2014.
3. Difficulty page. URL: <https://en.bitcoin.it/wiki/Difficulty>. Accesses 10/19/2014.
4. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: [http://www.R-project.org](http://www.r-project.org)
5. Bitcoin Block Explorer page. URL: <http://blockexplorer.com>. Accessed 10/14/2014.
6. Rob J. Hyndman and Yeasmin Khandakar. *Journal of Statistical Software* (2008), 27(3).
7. Forcasting: Principles and Practice online textbook. URL: <https://www.otexts.org/fpp/8/7>. Accessed 8/23/2014.

Wikipedia “Earthquake” Page. URL: <http://en.wikipedia.org/wiki/Earthquake>. Accessed 1/31/2013.

2. USGS “Measuring the size of an earthquake” Page. URL: <http://earthquake.usgs.gov/learn/topics/measure.php>. Accessed 1/31/2013

3. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: [http://www.R-project.org](http://www.r-project.org)

4. Seber, George AF, and Alan J. Lee. *Linear regression analysis*. Vol. 936. Wiley, 2012.

5. Ferguson, Thomas S. *A Course in Large Sample Theory: Texts in Statistical Science*. Vol. 38. Chapman & Hall/CRC, 1996.

6. R Markdown Page. URL: <http://www.rstudio.com/ide/docs/authoring/using_markdown>. Accessed 1/31/2013

7. USGS Real Time Data and Feeds Page. URL: <http://earthquake.usgs.gov/earthquakes/feed/>**.** Accessed 1/31/2013.