# NON-LINEAR MODELLING IN BRAZILIAN MARKET: EVALUATING THE FORECASTING PERFORMANCE OF NN (UNIVARIATE NEAREST NEIGHBOR) AND SNN (SIMULTANEOUS NEAREST NEIGHBOR) FORECASTING ALGORITHM

**Marcelo Scherer Perlin**
Escola de Administração - UFRGS
E-mail: marceloperlin@gmail.com

**Paulo Sérgio Ceretta**
UFSM
E-mail: ceretta@smail.ufsm.br

## ABSTRACT

The predictability of stock market's behavior is a topic studied by different academic circles for long time. A popular tool to make predictions about the stock market behavior on short term is the technical analysis. Such tool is based on the analysis of quantitative indicators and also chart patterns in order to identify the time to entry (buy) or exit the market (sell). A quantitative approach that is related to charting is the use of the non-parametric approach of nearest neighbor algorithm in order to produce forecasts of the time series on *t+1*. The main objective of this paper is to study the forecasting performance of the nearest neighbor method for the Brazilian Equity data in two versions, the univariate and also the multivariate case, which is also called simultaneous nearest neighbor. The main conclusion of the paper is that the ability of the algorithm in forecasting the values of the stock prices is mixed. A comparative analysis with the random walk model showed that this naïve approach has more explicative power in numerical accuracy. For the case of directional forecasts, the NN presented better results, resulting in correct directional forecasts moderately higher than 50% for most of the assets and with a maximum of approximately 60% correct market direction forecasts, which indicates that the method may add value in quantitative trading strategies. Comparing the results for both versions of the algorithm, its clear that both presented very similar results, but the univariate case was slightly better.

**Keywords:** Non linear Forecasts, Univariate and Multivariate Nearest Neighbor, Market Efficiency

## 1 INTRODUCTION

The study of the predictability of the financial markets has been, for a long time, the main type of research on the financial academics circles. A popular concept on such type of study is market efficiency hypothesis, which was formalized in the doctoral thesis by Fama.[1] Such concept, on it's weak form, says that the information arrives on the market and is

[1] The paper originated from the doctoral thesis is Fama (1965).

instantly priced, meaning that the price today is always the rational price, where arbitrage opportunities do not exist. Understanding the market as a fair game, then is not possible, according to the theory, for an investor to obtain a significant and positive excessive return in the long run based on public available information (including historical prices).

In opposition to the market efficiency theory, several papers have showed that past information is able, in some extent, to explain future stock market returns. Such predictability can appear in different ways, including time anomalies (day of the weak effect, French (1980)) and correlation between the asset's returns and others variables (Fama and French (1992)). A good review on the market efficiency subject can be found at the papers of Fama (1991) and Dimson e Mussavian (1998).

One type of research on the market's predictability topic is the performance of the popular technical analysis tools. Technical analysis, Murphy (1999), is the use of quantitative tools and graphical tools in order to make decisions about get in (buy) or get out of the market (sell). The quantitative part of such tools includes several indexes (moving averages, stochastics, etc) while the graphical part includes the analysis of patterns in the behavior of stock prices[2].

The quantitative part of technical analysis is easy to implement in trading rules[3], while the graphical part is not so easy to traduce into logical rules. In Lo (2000) was made an attempt to capture such visual patterns based on a non-parametric kernel regression. The main conclusion of this work is that several technical indicators do provide some information and may have some practical value on the analysis of stock's behavior.

An alternative approach to capture the visual part of technical analysis is the use of the forecasting algorithm called nearest neighbor, Clyde e Osler (1997). The basic idea behind nearest neighbor, now referred as NN, is that the series copies its own behavior along the time and such pattern can be explored for forecasting purposes. The algorithm selects the segments of the time series with the most similarity with the last segment available before the observation to be forecasted and, based on those selections, calculates the forecast of the time series on *t+1*. Such method can also be employed using information on multiple time series, which is the multivariate case, simultaneous nearest neighbor.

This method has already been applied to different data sets, such as equities, exchange rates and interest rates. See Bajo-Rubio et al (2002) for an assessment on the past studies

[2] Such patterns includes Head and Shoulders, Triple Top, etc. A complete description of such patterns can be found at Murphy (1999).
[3] As example, with the use of MAs (moving averages). A research that uses the quantitative part of technical analysis in order to produce trading rules is Saffi (2002).

regarding the forecasting ability for the nearest neighbor algorithm, mostly done by the same authors. The conclusion of Bajo-Rubio et al (2002) is that the model's forecasting power for financial time series is, on the majority of the studies, higher than a naïve approach, in this case the random walk model. As an example, Fernández-Rodríguez et al (1997), founds excessive profit from a simple timing strategy based on the forecasts from the nearest neighbor over a buy and hold rule in the Spanish equity market. In this same paper the method is compared against the random walk model and the directional forecasts assessed, both indicators showing the superiority of the mathematical formulation over the naïve predictive approaches (random walk and 50% directional forecasts).

The results for the method are not always positive, some mixed evidence can be found at, as example, Hsieh (1991), where the random walk model yielded lower sum of squared error than the nearest neighbor approach. The purpose of this research is to contribute to the academic discussion by applying the methodology for a different database, more precisely the Brazilian equity data. The NN algorithm has never been used on stocks from the Brazilian stock market, which is another motivation for this research.

The objective of this paper is to verify the forecasting ability for the Brazilian Equity Market of two versions of the nearest neighbor algorithm, the univariate and also the multivariate case. In order to asses the ability of the researched method, it will be analyzed the Utheils statistic and also the direction forecast percentage obtained from the out-of-sample forecasts.

The paper is organized as follows: in the first part will be presented a formal definition of both versions of the nearest neighbor algorithm, in the second part will be explored the method for assessing the forecasting performance of NN algorithm. In the third part will be showed the results obtained from the research and, on the final part, some concluding remarks are going to be presented.

## 2 THE NEAREST NEIGHBOR ALGORITHM

The nearest neighbor method is defined as a non-parametric class of regression. Its main idea is that the series copies its own behavior along the time. In other words, past pieces of information on the series have symmetry with the last information available before the observation on $t+1$. Such way of capturing the pattern on the times series behavior is the main

argument for the similarity between NN algorithm and the graphical part of technical analysis, charting.

The way the NN works is very different than the popular ARIMA model. The ARIMA modeling philosophy is to capture a statistical pattern between the locations of the observations in time. For the NN, such location is not important, since the objective of the algorithm is to locate similar pieces of information, independently of their location in time. Behind all the mathematical formality, the main idea of the NN approach is to capture a non-linear dynamic of self-similarity on the series, which is similar to the fractal dynamic of a chaotic time series.[4]

Next will be described the way the NN works for the univariate and also the multivariate case. For a detailed view of the process, the papers of Farmer e Sidorowich (1987) and Fernández-Rodríguez et al (1997) are indicated.

**2.1 Univariate Nearest Neighbor**

The univariate case for the NN algorithm work with the following steps:

1) The first step is to define a starting training period and divide such period on different vectors (pieces) $y_t^m$ of size *m*, where $t = m,...,T$. The value of *T* is the number of observations on the training period. The term *m* is also defined as the embedding dimension of the time series. For notation purposes, the last vector available before the observation to be forecasted will be called $y_T^m$, and the other pieces will be addressed as $y_i^m$.

2) The second step is to select *k* pieces most similar to $y_T^m$. For the Univariate case, in a formal notation, it is searched the *k* pieces with the highest value of $|\rho|$, which represents the absolute (euclidian) correlation between $y_i^m$ e $y_T^m$. The only difference between the univariate and the multivariate case is on this step: the way that is going to be searched for the *k* pieces with highest symmetry with $y_T^m$.

[4] A classic paper that describes how the chaos theory can be applied to financial market is Hsieh (1991). A practical example of such type of study on Brazilian data can be found at Perlin and Ceretta (2004).

3) With the *k* pieces on hand, each one with *m* observations, is necessary to understand in which way the *k* vectors can be used to construct the forecast on *t+1*. Several approaches can be employed here, including the use of an average or of a tricube function, Fernández-Rodríguez et al (2002). The method chosen for this paper is the one used on Fernández et al (2001), which consists on calculation of the following expression, Equation [1].

$$\hat{y}_{T+1} = \hat{\alpha}_0 + \hat{\alpha}_1 y_T + \hat{\alpha}_2 y_{T-1} + ... + \hat{\alpha}_m y_{T-m+1} \qquad [1]$$

The time varying coefficients in Equation [1], $\hat{\alpha}_0, \hat{\alpha}_1 .. \hat{\alpha}_m$ , are the ones derived from the estimation of a linear model with the dependent variable as $y_{i_r+1}$ and the explanatory variables as $y_{i_r}^m = (y_{i_r}, y_{i_r-1}, ...., y_{i_r-m+1})$, where *r* goes from 1 (one) to *k.* In order to facilitate the understanding of such regression, Equation [1] is presented on a matricial form on next expression, Equation [2].

$$\begin{bmatrix} y_{i_1+1} \\ y_{i_2+1} \\ y_{i_3+1} \\ \vdots \\ y_{i_k+1} \end{bmatrix} = \hat{\alpha}_0 + \hat{\alpha}_1 \begin{bmatrix} y_{i_1} \\ y_{i_2} \\ y_{i_3} \\ \vdots \\ y_{i_k} \end{bmatrix} + \hat{\alpha}_2 \begin{bmatrix} y_{i_1-1} \\ y_{i_2-1} \\ y_{i_3-1} \\ \vdots \\ y_{i_k-1} \end{bmatrix} + ... + \hat{\alpha}_m \begin{bmatrix} y_{i_1-m+1} \\ y_{i_2-m+1} \\ y_{i_3-m+1} \\ \vdots \\ y_{i_k-m+1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_k \end{bmatrix} \qquad [2]$$

For a clarified view of Equation [2], is necessary to comprehend that the NN algorithm is non temporal. The values of $y_{i_k+1}$ are the observations one period ahead of the pieces chosen by the correlation criteria defined earlier. The term $\hat{y}_{i_k-m+1}$ indicates the first values[5] of the *k* chosen pieces, while the term $y_{i_k}$ represents the last terms of each piece chosen. It's easy

[5] When said "first values" of the vector, means, in time notation, the oldest observations. The opposite is true for "last values".

to see that the number of explanatory series on [2] is *m,* and that each one of those will have *k* observations.

The term $\hat{\alpha}_1$, Equation [2], is the coefficient aggregated to the last observation of the chosen series and $\hat{\alpha}_2$ is the coefficient for all the second last observations of all k series. This logic for the coefficients continues until it reaches the first observation of all k chosen series, $\hat{\alpha}_m$. The values of the coefficients on Equation [2] are estimated with the minimization of the sum of the quadratic error ($\sum_{i=1}^{k} \varepsilon_i^2$). The steps 1-3 are executed over time, with time varying parameter for [2], until all out of sample forecasts are created.[6]

**2.2 Multivariate NN Algorithm (Simultaneous Nearest Neighbor).**

The multivariate version of NN works with the same steps presented earlier. The difference is only on the way that the algorithm is going to search for the *k* similar pieces of $y_T^m$. Using the same mathematical notation defined earlier, where $x_t^m$ are the historical pieces of the independent time series $x_t$ and $x_T^m$ is the piece located before the observation to be forecasted on $y_t$, the execution of step 2 is made by maximizing the following formula, Equation [3].

$$\left|\rho\right|(y_i^m, y_T^m) + \left|\rho\right|(x_i^m, x_T^m) \qquad [3]$$

For Equation [3], the term $\left|\rho\right|(y_i^m, y_T^m)$ is the absolute correlation between the pieces ($y_i^m$) and the piece located before the value to be forecasted ($y_T^m$). The basic idea behind the use of [3] is that the independent and the dependent series present regularity on the nearest neighbor location. The purpose on the use of $x_i^m$ is only to help the algorithm to find the location of the *k* nearest neighbor of $y_T^m$ and that's the reason why the multivariate version of NN is also referred as simultaneous NN.

[6] All of the calculations needed for this paper were made based on two functions created by the author on the software MatLab. The functions are public available in http://www.mathworks.com/matlabcentral/fileexchange, keyword: nearest forecasts.

## 3 DATABASE AND METHODOLOGY

The database for this research was prospected from the software Economática. The data to be modeled by NN predictors is based on the daily price series for 5 of the most liquid stocks between 01/01/1996 and 01/01/2006. Such assets are Petrobras, Vale do Rio Doce, Bradesco, Eletrobras and Cemig. The research method needs a large training set in order to produce the forecasts. Given that, the training period for the algorithm is from 01/01/1996 until 01/01/2005, meaning that the year of 2005-2006 will be the out-of-sample period that the forecasts are going to be analyzed. Next, Table 1, is presented the descriptive statistic for the logarithm returns from the assets in the respective researched period.

Table 1 – Descriptive Statistics for the Logarithm Returns from the Assets.

| Statistics | Assets | | | | |
|---|---|---|---|---|---|
| | Petrobras | Vale Rio Doce | Bradesco | Eletrobras | Cemig |
| Average | 0.13% | 0.12% | 0.12% | 0.05% | 0.09% |
| Standard Deviation | 2.72% | 2.65% | 2.64% | 3.45% | 3.21% |
| Excess of Kurtosis | 7.91 | 20.49 | 4.91 | 6.32 | 6.76 |
| Skewness | -0.25 | 1.43 | -0.25 | 0.46 | 0.13 |

From Table 1, it's possible to see that Eletrobras and Cemig presented the largest volatilities, with return's standard deviation of respectively 3.45% and 3.21%. Another information available at Table 1 is the positive excess of kurtosis for all assets, which is a stylized fact for financial time series, meaning that the respective distributions have more frequency of extreme values than a normal distribution.[7]

It's important to remember that the NN algorithm has two parameter, the value of *m* (size of the histories) and *k* (number of nearest neighbors). For this paper the values tested are *m=3, 4, 5, 6, 7,8* and *k*=80,100,120,140.

### 3.1 Assessing the Forecasting Performance of the NN Algorithm

[7] The popular denomination of such fact is that the distribution has fat tails. A class of models capable of simulating such characteristic is the popular Arch family, which was introduced by the seminal paper of Engle (1982).

With all the details about the forecasting method clarified, the next logical step is to formalize the way that the analysis of the forecasts is going to be made. For this paper, the method will be the observation of the Utheils statistic and also the directional forecast percentage. Both statistics are detailed next.

### 3.1.1 U-theils Statistic

The Utheils statistics is a simple comparison between the quadratic error from the method evaluated and the random walk model. [8] Next, Equation [4], is presented the formula for this calculation.

$$U = \frac{\sqrt{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}}{\sqrt{\sum_{t=1}^{T} (y_t - y_{t-1})^2}} \qquad [4]$$

In Equation [4], the denominator of the formula is just the square root of the sum of the quadratic error from a random walk model while the numerator of [4] is the square root of the sum of the quadratic error obtained from the predictions. It's easy to see that if *U* is lower than one (1), then the model that has build the forecasts has more explicative power then the naïve approach, which is clearly a good thing. The U-theils statistic can also be seen as a numerical evaluation of the forecasts accuracy.

### 3.1.2 Directional Forecast Percentage

The directional forecast percentage is a type of analysis that is related to the performance of the model regarding the direction in the behavior of the price series. This index is simply the percentage of correct market direction forecasts. For example, in time *t,* if the model predicts that the stock price will rise on *t*, and that is actually true, then the value of one is assigned to a dummy vector at time *t* and zero otherwise. The formal definition of such concept is presented next, Equation [5].

[8] The random walk represents a naïve prediction. This comparative model is given by $y_t = y_{t-1} + \varepsilon_t$, meaning that the forecast for tomorrow's price is the value of the asset today. It's easy to show that the RW has a variance that grows with the number of observations (not stationary) and that the expected value of such model is the first observation of the series ( $y_0$ ).

$$Df = \frac{\sum_{i=1}^{T} d_t}{T} \qquad [5]$$

For Equation [5], the variable $d_t$ is a logical vector that takes values 1 (one) if the forecasted market direction is correct and 0 (zero) otherwise. [9] The term *T* is the number of forecasts made. If the model tested has good forecasting ability, then the value of *Df* will be higher than 50%, which is the naïve approach here. The value of *Df* is also important here because it partially shows how would the model perform in a timing trading strategy (buy asset if expected return on *t+1* is positive).

## 4 RESULTS

The presentation of results on this paper is structured in two parts: the first is the performance analysis for the univariate version of NN and the second is for the simultaneous NN (multivariate nearest neighbor algorithm). Both versions were executed for parameters m=3, 4, 5, 6, 7, 8 and k=80, 100, 120 and 140. Next, Table 2, is presented the first part of the analysis.

Table 2 – U-theils Statistic and Directional Forecast (DF) for Univariate NN

| **PANEL A – Petrobras** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Value of *m*** | **Values of *k*** | | | | | | | |
| | **80** | | **100** | | **120** | | **140** | |
| | U-Theils | DF | U-Theils | DF | U-Theils | DF | U-Theils | DF |
| 3 | 1.02 | 54.44% | 1.03 | 52.82% | 1.02 | 54.44% | 1.03 | 52.82% |
| 4 | 1.09 | 52.42% | 1.08 | 52.42% | 1.06 | 51.61% | 1.05 | 52.82% |
| 5 | 1.04 | 53.63% | 1.04 | 53.63% | 1.04 | 53.23% | 1.02 | 54.44% |
| 6 | 1.07 | 53.63% | 1.06 | 54.03% | 1.05 | 54.44% | 1.05 | 53.23% |
| 7 | 1.08 | 54.84% | 1.07 | 53.23% | 1.06 | 55.65% | 1.06 | 54.84% |
| 8 | 1.12 | 52.82% | 1.09 | 53.63% | 1.07 | 53.63% | 1.09 | 54.03% |
| **PANEL B - Vale do Rio Doce** | | | | | | | | |
| **Value of *m*** | **Values of *k*** | | | | | | | |
| | **80** | | **100** | | **120** | | **140** | |
| | U-Theils | DF | U-Theils | DF | U-Theils | DF | U-Theils | DF |
| 3 | 1.04 | 57.26% | 1.03 | 59.68% | 1.01 | 59.68% | 1.01 | 59.68% |
| 4 | 1.05 | 56.05% | 1.03 | 58.47% | 1.01 | 57.26% | 1.01 | 56.85% |

[9] An intuitive approach to see if the forecasted market direction is correct is to check if $(\hat{P}_t - P_{t-1}) * (P_t - P_{t-1}) > 0$. This does not work for the cases where $P_t = P_{t-1}$ and $\hat{P}_t = P_{t-1}$.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 5 | 1.05 | 56.05% | 1.04 | 55.65% | 1.03 | 56.05% | 1.03 | 54.44% |
| 6 | 1.06 | 54.03% | 1.05 | 52.02% | 1.05 | 54.03% | 1.06 | 54.03% |
| 7 | 1.11 | 51.61% | 1.08 | 52.82% | 1.09 | 52.42% | 1.07 | 53.63% |
| 8 | 1.12 | 51.61% | 1.10 | 52.42% | 1.10 | 51.21% | 1.08 | 52.42% |

**PANEL C – Bradesco**

| **Value of *m*** | **Values of *k*** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **80** | | **100** | | **120** | | **140** | |
| | U-Theils | DF | U-Theils | DF | U-Theils | DF | U-Theils | DF |
| 3 | 1.05 | 51.21% | 1.04 | 52.42% | 1.03 | 50.00% | 1.01 | 49.60% |
| 4 | 1.04 | 51.21% | 1.02 | 49.60% | 1.02 | 51.21% | 1.01 | 52.42% |
| 5 | 1.01 | 50.00% | 1.01 | 51.21% | 1.01 | 52.82% | 1.00 | 51.61% |
| 6 | 1.03 | 47.58% | 1.03 | 49.19% | 1.03 | 49.60% | 1.04 | 50.00% |
| 7 | 1.10 | 48.79% | 1.09 | 51.61% | 1.08 | 50.00% | 1.08 | 49.60% |
| 8 | 1.08 | 50.00% | 1.08 | 50.00% | 1.06 | 51.21% | 1.02 | 52.42% |

**PANEL D – Eletrobras**

| **Value of *m*** | **Values of *k*** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **80** | | **100** | | **120** | | **140** | |
| | U-Theils | DF | U-Theils | DF | U-Theils | DF | U-Theils | DF |
| 3 | 1.07 | 45.16% | 1.07 | 45.56% | 1.05 | 45.16% | 1.05 | 45.56% |
| 4 | 1.03 | 48.39% | 1.03 | 47.98% | 1.04 | 47.58% | 1.04 | 46.37% |
| 5 | 1.08 | 43.95% | 1.07 | 45.16% | 1.07 | 44.76% | 1.06 | 46.37% |
| 6 | 1.08 | 47.18% | 1.07 | 47.18% | 1.07 | 47.18% | 1.06 | 46.77% |
| 7 | 1.07 | 47.58% | 1.05 | 48.39% | 1.05 | 47.18% | 1.04 | 47.18% |
| 8 | 1.08 | 44.35% | 1.07 | 45.16% | 1.06 | 45.16% | 1.06 | 43.15% |

**PANEL E – Cemig**

| **Value of *m*** | **Values of *k*** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **80** | | **100** | | **120** | | **140** | |
| | U-Theils | DF | U-Theils | DF | U-Theils | DF | U-Theils | DF |
| 3 | 1.06 | 48.39% | 1.07 | 47.98% | 1.06 | 49.60% | 1.05 | 50.81% |
| 4 | 1.07 | 52.42% | 1.06 | 50.40% | 1.05 | 48.39% | 1.04 | 50.40% |
| 5 | 1.11 | 52.02% | 1.06 | 50.00% | 1.04 | 49.19% | 1.04 | 49.60% |
| 6 | 1.09 | 50.40% | 1.07 | 48.79% | 1.07 | 50.40% | 1.06 | 48.39% |
| 7 | 1.09 | 49.60% | 1.06 | 49.19% | 1.05 | 48.39% | 1.04 | 48.39% |
| 8 | 1.05 | 50.00% | 1.05 | 49.60% | 1.03 | 52.42% | 1.03 | 52.42% |

The first thing to be analyzed on Table 2 is the numerical accuracy of the forecasts (U-theils statistic). It's possible to see that the performance of the univariate version of NN, from this point of view, was poor. For all of the combinations between the embedding dimension (*m*) and the number of nearest neighbors (*k*) the values of U-theils are higher than one (1), meaning that the naïve approach of the random walk model had better performance than the researched method for all the assets. The best case was for Bradesco, with *m=5* and *k=140*, where the value of U-theils was approximately, but higher than one (1).

From Table 2, observing the values of DF (directional forecast), it can be seen that the algorithm had a moderate performance. The majority of the values of DF are higher than 50%, meaning that the NN algorithm was able, for most of the cases, to predict correctly the asset

behavior regarding price direction. The best case was for Vale do Rio Doce, with all DFs higher than 50%, and with the maximum of 59.68% of correct direction forecasts.

The second part of the results regards the analysis of the multivariate case of NN. Next, Table 3, the values of U-theils and DF are presented for this version of NN.

Table 3 - U-theils Statistic and Directional Forecast (DF) for Multivariate NN (SNN)

| **PANEL A – Petrobras** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Value of *m*** | **Values of *k*** | | | | | | | |
| | **80** | | **100** | | **120** | | **140** | |
| | U-Theils | DF | U-Theils | DF | U-Theils | DF | U-Theils | DF |
| 3 | 1.06 | 53.23% | 1.04 | 52.82% | 1.03 | 53.63% | 1.03 | 53.23% |
| 4 | 1.07 | 56.85% | 1.06 | 52.82% | 1.06 | 54.03% | 1.06 | 51.61% |
| 5 | 1.09 | 54.44% | 1.07 | 54.84% | 1.05 | 54.84% | 1.04 | 55.65% |
| 6 | 1.09 | 52.02% | 1.08 | 49.19% | 1.07 | 50.00% | 1.07 | 50.00% |
| 7 | 1.09 | 52.02% | 1.09 | 51.61% | 1.08 | 52.42% | 1.05 | 52.42% |
| 8 | 1.12 | 48.79% | 1.09 | 52.02% | 1.08 | 51.61% | 1.05 | 51.61% |
| **PANEL B - Vale do Rio Doce** | | | | | | | | |
| **Value of *m*** | **Values of *k*** | | | | | | | |
| | **80** | | **100** | | **120** | | **140** | |
| | U-Theils | DF | U-Theils | DF | U-Theils | DF | U-Theils | DF |
| 3 | 1.07 | 57.26% | 1.07 | 58.06% | 1.04 | 58.87% | 1.03 | 56.85% |
| 4 | 1.06 | 53.63% | 1.04 | 54.84% | 1.02 | 54.03% | 1.01 | 52.82% |
| 5 | 1.07 | 54.03% | 1.04 | 58.47% | 1.04 | 57.26% | 1.03 | 56.05% |
| 6 | 1.09 | 55.65% | 1.06 | 56.05% | 1.06 | 56.45% | 1.05 | 56.05% |
| 7 | 1.10 | 52.02% | 1.08 | 53.23% | 1.07 | 52.82% | 1.05 | 52.02% |
| 8 | 1.11 | 51.61% | 1.08 | 50.40% | 1.04 | 50.00% | 1.05 | 52.02% |
| **PANEL C – Bradesco** | | | | | | | | |
| **Value of *m*** | **Values of *k*** | | | | | | | |
| | **80** | | **100** | | **120** | | **140** | |
| | U-Theils | DF | U-Theils | DF | U-Theils | DF | U-Theils | DF |
| 3 | 1,07 | 52,82% | 1,05 | 50,81% | 1,03 | 51,21% | 1,02 | 50,81% |
| 4 | 1,06 | 47,98% | 1,05 | 48,39% | 1,03 | 47,58% | 1,02 | 46,77% |
| 5 | 1,07 | 49,60% | 1,05 | 49,60% | 1,04 | 50,40% | 1,05 | 50,81% |
| 6 | 1,04 | 51,61% | 1,03 | 48,79% | 1,04 | 48,79% | 1,03 | 46,77% |
| 7 | 1,07 | 48,79% | 1,05 | 46,37% | 1,04 | 46,77% | 1,04 | 48,39% |
| 8 | 1,06 | 49,19% | 1,05 | 49,19% | 1,05 | 47,98% | 1,05 | 47,58% |
| **PANEL D – Eletrobras** | | | | | | | | |
| **Value of *m*** | **Values of *k*** | | | | | | | |
| | **80** | | **100** | | **120** | | **140** | |
| | U-Theils | DF | U-Theils | DF | U-Theils | DF | U-Theils | DF |
| 3 | 1.06 | 45.56% | 1.06 | 45.97% | 1.06 | 46.37% | 1.05 | 45.56% |
| 4 | 1.06 | 48.39% | 1.04 | 47.98% | 1.02 | 47.58% | 1.02 | 47.98% |
| 5 | 1.06 | 47.58% | 1.05 | 47.98% | 1.06 | 47.58% | 1.05 | 48.79% |
| 6 | 1.05 | 47.18% | 1.05 | 46.37% | 1.04 | 46.77% | 1.03 | 46.77% |
| 7 | 1.04 | 47.98% | 1.04 | 49.19% | 1.03 | 48.79% | 1.02 | 50.00% |
| 8 | 1.06 | 48.79% | 1.06 | 47.18% | 1.05 | 45.97% | 1.05 | 47.18% |

| PANEL E – Cemig | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Value of *m* | Values of *k* | | | | | | | |
| | 80 | | 100 | | 120 | | 140 | |
| | U-Theils | DF | U-Theils | DF | U-Theils | DF | U-Theils | DF |
| 3 | 1.10 | 49.19% | 1.06 | 47.18% | 1.04 | 47.18% | 1.03 | 47.18% |
| 4 | 1.06 | 51.21% | 1.06 | 50.40% | 1.04 | 50.40% | 1.04 | 50.40% |
| 5 | 1.11 | 52.42% | 1.09 | 50.00% | 1.08 | 50.40% | 1.06 | 51.21% |
| 6 | 1.11 | 50.81% | 1.09 | 53.23% | 1.07 | 52.02% | 1.05 | 50.40% |
| 7 | 1.08 | 51.21% | 1.06 | 51.21% | 1.03 | 51.61% | 1.03 | 50.00% |
| 8 | 1.06 | 50.40% | 1.06 | 51.61% | 1.03 | 53.23% | 1.03 | 53.63% |

Analyzing the values on Table 3, it's possible to see that the U-theils statistic are always bellow one (1), meaning, again, that the random walk model presented higher explicative power regarding numerical forecasts of the researched data, which corroborates with the information on Table 2.

For the DF (directional forecast) analysis, the values on Table 3 indicates that the two versions of the NN algorithm presented very similar results, with the values of DF, for most of the cases, oscillating in 50%. For Table 3, the best case is, again, for the asset Vale do Rio Doce, with a percentage of correct directional forecasts in 58.87%.

The previous analysis of Table 2 and Table 3 indicates two things: 1) the NN algorithm, univariate and also the multivariate case, showed bad performance regarding numerical accuracy and 2) for the case of direction forecast, the method presented moderately better results, which indicates that the algorithm may add value in timing strategies, but its important to notes that the evidence wasn't strong.

When looking at the values of Table 2 and 3, it's possible to see that some assets presented better results than others. For example, for the univariate and multivariate case, Vale do Rio Doce presented the best results when looking at the different combination of parameters while Eletrobras presented the worst, with all DF lower than 50% and the majority of U-theils higher than 1.05. This show that the nearest neighbor algorithm may work well for some equity data and not so good to another, and such results is not just bad parameter choice since the different parameter combinations are all going in the same direction as the assets change in the analysis. The contribution of this evidence is that when analyzing the forecasting ability of the nearest neighbor algorithm, the researcher should apply it to different series from the same database and not take conclusions based solely in one time series.

Another type of analysis that is important here is to check the sensitivity of the results regarding the combination of the parameters. For that purpose, Table 4 is presented next.

Table 4 – Mean and Standard Deviation for U-theils and DF values.

| **Panel A - Univariate NN** | | | | |
|---|---|---|---|---|
| Assets | U-Theils | | Direction Forecast | |
| | Mean | Standard Deviation * | Mean | Standard Deviation * |
| Petrobras | 1.06 | 0.03 | 53.61% | 0.93% |
| Vale do Rio Doce | 1.06 | 0.03 | 54.97% | 2.73% |
| Bradesco | 1.04 | 0.03 | 50.55% | 1.31% |
| Eletrobras | 1.06 | 0.01 | 46.19% | 1.44% |
| Cemig | 1.06 | 0.02 | 49.88% | 1.37% |
| **Panel B - Multivariate NN** | | | | |
| Assets | U-Theils | | Direction Forecast | |
| | Mean | Standard Deviation * | Mean | Standard Deviation * |
| Petrobras | 1.07 | 0.02 | 52.57% | 1.97% |
| Vale do Rio Doce | 1.06 | 0.03 | 54.60% | 2.58% |
| Bradesco | 1.05 | 0.01 | 49.04% | 1.71% |
| Eletrobras | 1.05 | 0.01 | 47.48% | 1.18% |
| Cemig | 1.06 | 0.03 | 50.69% | 1.74% |

* The use of the standard deviation is just to have a coefficient that can tell us the dispersion on the distribution of the results. It could be used other kind of dispersion parameter, but the results would be very similar.

The objective of Table 4 is to present a brief review of the results from Table 2 and 3. Such compilation is just the mean and the standard deviation for every combination of *m* and *k,* for each asset. The values of mean for directional forecast and U-theils are much similar for the univariate and the multivariate version of NN of each asset. Observing the values of standard deviation of the results obtained from Table 3 and 4, it can be checked that the results didn't presented much variability regarding the combinations between *m* and *k*. This result suggests that changing the values of the model's parameter (*m* and *k*) do not produced a significant difference in the results, which indicates that the NN algorithm is not very sensible to the parameters combination regarding numerical and directional forecast accuracy.

Another information that can be prospected from Tables 2,3 and 4 is that the both versions of the algorithm presented very similar results, indicating that the output from the two versions aren't so different.

## 5 CONCLUSION

The basic idea of the paper was to evaluate the forecasting performance of the nearest neighbor in two versions, the univariate and also the multivariate case. Such non-linear method has an intuitive appealing that is symmetric to the visual part of technical analysis, the so called charting. The algorithm was applied to the five most liquid assets of Brazilian Market in the researched period. Such assets were Petrobras, Vale do Rio Doce, Bradesco,

Eletrobras and Cemig. In the evaluation of the forecasts, two statistics were computed: the U-theils statistics, which is basically a comparative analysis between the model tested and the random walk approach, and also the directional forecast percentage, which evaluates the model accuracy performance regarding predicted direction of stock price changes.

The results showed that, in the evaluation of the U-theils statistics, the univariate and the multivariate versions of NN had worst performance than the naïve approach (RW). This information leads to the conclusion that the method researched (in both versions) is not suitable regarding numerical forecasts accuracy. In the evaluation of directional forecasts, the NN algorithm presented better results. For most of the assets, the combinations between the values of the model's parameters (*m* and *k*) presented percentage of correct market direction forecasts moderately higher than 50%. The best case is for the asset Vale do Rio Doce, in the univariate case, with approximately 60% of correct market direction forecasts. Another important information obtained from the research, Table 4, is that the NN approach is not very sensible regarding the combination between the model's parameters *m* and *k*.

The main conclusion of the paper is that the use of NN algorithm in two versions, univariate and multivariate, will output similar forecasts, even though the first produced slightly superior prediction power when compared to the second. From the portfolio manager point of view, what this study suggests is that the nearest algorithm may be used in quantitative trading strategies such as timing[10] the market. But, such evidence for profitability (directional forecast) wasn't strong, meaning that this suggestion should be taken carefully, maybe by checking the trading rule returns for the past data before using the algorithm for trading. An example of this type of study can be found at Fernandez Rodriguez et al (2001).

The result from this paper partially corroborates with the conclusion in Fernandez-Rodrigues et al (2001). For the case of numerical accuracy, the performance of the method was poor, which doesn't follow the results obtained in the work of Bajo-Rubio et al (2002) and Fernandez Rodriguez et al (1997). Regarding the contribution of this study to the literature, it can be labeled as a case of mixed results for the nearest neighbor forecasting ability. Analyzing the differences between the papers where the results were positive, can be seen that in those studies the number of observation used for the training data were higher than the one used here (by limitation of reliable data) and also most of those studies were applied to market indexes, and not single stocks as this one. The investigation those

[10] Timing the market is a very simple trading strategy consisting of entering the market when the forecast on t+1 is positive and closing the position when the forecast is negative. This rule is for the long only case, but it can be easily extended for the use of short trades.

differences regarding the forecasting power of the researched method is subject for future studies.

## BIBLIOGRAPHY

BAJO-RUBIO, O., SOSVILLA-RIVERO, S., FERNÁNDEZ-RODRÍGUEZ, F. Non-linear Forecasting Methods: Some applications to the analysis of Financial Series. *Working paper*, n. 1, FEDEA, 2002.

CARVALHAL DA SILVA, A. L., RIBEIRO, T. L. Estimating and Forecasting Latin American Indexes Using Artificial Neural Networks. Proceedings of BALAS, 2002.

CLYDE, W. C., OSLER, C. L. Charting: Chaos Theory in Disguise? J*ournal of Futures Markets*, august, v. 17, n. 5, p. 489-514, 1997.

DIMSON, E., MUSSAVIAN, M. A brief history of market efficiency. *European Financial Management*, v. 4, p. 91-193, 1998.

ENGLE, R. F. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. Econometrica, ed. 50, p. 987-1008, 1982.

FAMA, E. Efficient Capital Markets: II. *Journal of Finance*, v. 46, p. 1575-1617, 1991.

FAMA, E., F. The behavior of Stock Market Prices. *The Journal of Business*. Janeiro, p. 34-105, 1965.

FAMA, E., FRENCH, K. The cross-section of expected stock returns. *Journal of Finance*, v. 47 (2), p. 427-465, 1992.

FARMER, D., SIDOROWICH, J. Predicting chaotic time series. *Physical Review Letters,* v. 59, p. 845-848, 1987.

FERNÁNDEZ-RODRÍGUEZ, F., RIVERO, S. S., FELIX, J. A. Nearest Neighbor Predictions in Foreign Exchange Markets. *Working Paper,* n. 05, FEDEA, 2002.

FERNÁNDEZ-RODRÍGUEZ, F., SOSVILLA-RIVERO, S., GARCÍA-ARTILES, M. Using nearest neighbor predictors to forecast the Spanish Stock Market. *Investigaciones Económicas,* v. 21, p. 75-91, 1997.

FERNÁNDEZ-RODRÍGUEZ, F., SOSVILLA-RIVERO, S., GARCÍA-ARTILES, M. An empirical evaluation of non-linear trading rules. *Working paper,* n. 16, FEDEA, 2001.

FRENCH, K. Stock Returns and the Weekend Effect. *Journal of Financial Economics*, March, p. 55-69, 1980.

HSIEH, David A. Chaos and nonlinear dynamics: applications to financial markets. *The Journal of Finance*, v. 46, n.5, 1839-1877, 1991.

LO. A. Foundations of Technical Analysis: Computational Algorithms Statistical Inference, and Empirical Implementation. *Journal of Finance*, n. 55, p. 1705-1765, 2000.

MURPHY, J. Technical Analysis of the Financial Markets. New York Institute of Finance, 1999.

PERLIN, M. S., CERETTA, P. S. Chaos Theory Applied to the Contracts of Arabic Coffee on the Brazilian Derivative Market. Proceedings of the 4º Brazilian Congress of Finance, 2004.

SAFFI, P. A. C. Technical Analysis: Luck or Reality? Proceedings of the 2º Brazilian Congress of Finance, 2002.