Author: Marcelo Scherer Perlin.

UFRGS – University of Rio Grande do Sul, Brazil.

Graduate Business School, Department of Finance.

1. Description of 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.

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

1.1 The Univariate Nearest Neighbor (method=correlation)

The univariate case for the NN algorithm works 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 observation 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 method of correlation, 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 and 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.

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 ways 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 case of the function is the one used on

Fernández et al (2001), which consists on calculation of the following expression, Equation [1].

$$\hat{y}_{T+1} = \hat{\alpha}_0 + \hat{\alpha}_1 y_{T-1} + \hat{\alpha}_2 y_{T-2} + ... + \hat{\alpha}_m y_{T-m} \qquad [1]$$

The 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].

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

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 y_{i_k-m} indicates the first values of the k chosen pieces, while the term y_{i_k} represents the last terms of each piece chosen. It's easy 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, $\hat{\alpha}_{m-1}$. The values of the coefficients on Equation [2] are estimated with the minimization of the sum of the quadratic error ($\sum_{i=1}^{k} \varepsilon_k^2$). The steps 1-3 are executed in a loop until the point that all forecasts on *t+1* are created.

1.2 The Univariate Nearest Neighbor (method=absolute_distance)

This version of the NN is much simpler than the one with the method of correlations. The steps are:

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 observation 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 method of absolute distance it is searched the *k* pieces with the lowest sum of distances between the vectors y_i^m and y_T^m.

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*. The absolute distance method simply verifies the observations ahead of the *k* chosen neighbors and takes the average of them.

The steps 1-3 are executed in a loop until the point that all forecasts on *t+1* are created. As can be seen from the descriptions, the absolute distance approach is much simpler than the one with correlation, since it does not use any kind of local regression.

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], *k* times.

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

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_t to model y_t 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.

Bibliography.

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.