

Daily sugar price forecasting using the Mixture of Local Expert Models

B. de Melo¹, C. L. Nascimento Júnior² & A. Z. Milioni³

¹*Banco do Brasil, Assets and Business Appraisal Division, Brazil*

²*Instituto Tecnológico de Aeronáutica, Electronic Engineering Division, Brazil*

³*Instituto Tecnológico de Aeronáutica, Division of Mechanical & Aeronautical Engineering, Brazil*

Abstract

This article concerns the application of the Mixture of Local Expert Models (MLEM) technique to predict the daily prices of the Sugar No. 14 contract in the New York Board of Trade (NYBOT). The MLEM technique can be seen as a Data Mining tool that performs data exploratory analysis and mathematical modeling simultaneously. Given a set of input-output data points, the basic idea is as follows: 1) firstly, considering only the input side of the training data, a Kohonen Neural Network is used to divide the data into non-overlapping clusters of points, 2) several modeling techniques are then used to construct competing models for each cluster, considering only the data points in that cluster, 3) the best model for each cluster is then selected and called the Local Expert Model. Finally, the output of all Local Expert Models are combined by a so-called Gating Network (a Radial Basis Function) that considers: a) the distance of the input data point to the center of the cluster of data points used to generate each Local Expert Model, and b) the size of the region of the input space taken by each cluster of training data points. The following modeling techniques were used to develop the Local Expert Models: Artificial Neural Networks, Multiple Regression and Carbon Copy. For comparison purposes, the same modeling techniques are also evaluated when acting as Global Experts, i.e., when the technique uses the entire data set without any clustering.

Keywords: mixture of local experts, clustering, artificial neural networks, multiple regression analysis, time series forecasting.



1 Introduction

The scientific aim of this paper is to advance the research on solving data mining problems through the use and development of modern mathematical modeling and statistical tools capable of handling very large scale data.

It is well known that in general no modeling technique is complete. Some techniques exhibit fast convergence but might fail when submitted to cross validation tests. Other techniques perform well regarding cross validation but have poor convergence properties, as in Nascimento Jr. and Yoneyama [1].

In this paper we propose a modeling technique designed to combine the results of different experts (time series forecasting techniques, in our case) where each expert model (called Local Expert) is developed using only part of the data set. Many expert models are developed for the same part of the data set and only the best expert for each part is then used.

Several of the traditional time series forecasting techniques use *linear* models which, by their nature, are not capable of capturing *non-linear* behavior that is often present in real world situations. Artificial neural networks and other techniques allow the development of non-linear time series forecasting models, which, however, do not necessarily imply better results when compared to traditional *linear* techniques, as in Makridakis et al. [2].

Our purpose is to combine *linear* and *non-linear* techniques in the task of time series forecasting, capturing the best characteristics of each technique as proposed in Melo [3].

2 Mixture of local experts system as a data mining tool

The mixture of local experts system presented in this article follows the idea proposed by Jacobs et al. [4] and has the following procedure: a) divide the data set into regions or clusters, b) for each cluster train all expert models, c) find the best expert for each cluster, and d) implement a composition of the best local experts using a *gating network* which will decide how to weight each local expert output for a given input point (figure 1).

The major hypothesis of the proposed MLEM is that, when the data set can be divided into a set of clusters, one can develop a local model (a local expert) for each data cluster. However, one has to define a procedure to calculate the output when an input point x does not belong exactly to any of the data clusters used to construct the local models.

The steps in order to construct the desired models have the following phases:

- Firstly we cluster the input data set (X) in several regions (X_i).
- For each one of the regions, considering only the data points in the training set of that region, all experts are used to construct the local models.
- The best local expert for each of the regions is found, considering the smallest RMSE (Root Mean Squared Error) measured using only the data points in the training set of the region:



$$\text{RMSE} = \sqrt{\frac{1}{NT} \sum_{t=1}^{NT} (Y_t - \hat{Y}_t)^2} \quad (1)$$

where NT = number of data points in the training set of the region, Y_t is the observed output and $\hat{Y}_t$ is the local model output.

- The best local expert for each of the regions is found, considering the smallest RMSE (Root Mean Squared Error) measured using the data points in the *test* set of the region, as in eqn. (1), but changing NT by NE = number of data points in the *test* set of the region.

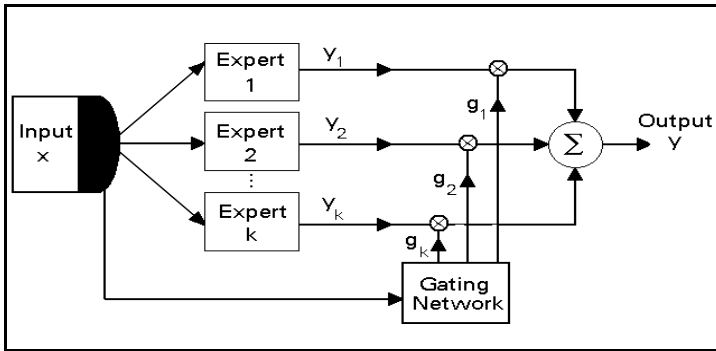


Figure 1: Mixture of local experts models (MLEM).

In our proposal, after the local expert has been developed, the MLEM structure can be used to yield a forecast for a given input point x in the following manner:

- The input point x is delivered to the best expert elected for each cluster i who computes its output y_i .
- Next the gating network is used to compute the weight coefficients g_i which will depend on the distance of the input point x to the center of each cluster as well as the size of the region of the input space taken by each cluster of training data points.
- The final output y will be computed as the weighted average of the outputs y_i using the coefficients g_i as weight factors, i.e.:

$$y = \sum_{i=1}^k g_i y_i \quad (2)$$

where k represents the number of local experts.

3 Data set clustering and weight coefficients computation

3.1 Data set clustering

The data set is clustered using a Kohonen neural network trained by a SOM (Self-Organizing Map) algorithm applied to available data, as in Kohonen [5]. The Kohonen neural network training aims at finding similarities on the input data set. It is possible to show that a Kohonen neural network divides the input data set in such a way that input points that are close to each other (by a given measure) will be assigned to the same cluster, as in Nascimento Jr. and Yoneyama [1] and Haykin [6].

Each cluster defines a region on the input data set space and each input data used during training belongs to one and only one of these clusters. It is possible to show (see Nascimento Jr. and Yoneyama [1]) that, after training, the weights of the Kohonen neural network units indicate the center of each cluster (denoted by ctr_i).

3.2 Weight coefficients computation

According to Bishop [7] and Mitchell [8], the weight coefficients g_i can be computed using a Radial Basis Function (RBF) as follows. Firstly the center of the clusters and their variances are used to compute the coefficients d_i :

$$d_i = \exp \left\{ -\frac{1}{2} \frac{\|x - ctr_i\|^2}{(S_i^2 / S^2)} \right\} \quad (3)$$

where:

x input vector to be forecasted,

ctr_i center of the i^{th} cluster of the training data, $i = 1, 2, \dots, k$,

S_i^2 variance of the distance $(x_j - ctr_i)$ where x_j = training input vector assigned to the i^{th} cluster.

S^2 largest variance S_i^2 , i. e., $S^2 = \max(S_i^2)$ for $i = 1, 2, \dots, k$.

The coefficients g_i can then simply be computed by normalizing the coefficients d_i :

$$g_i = \frac{d_i}{\sum_{j=1}^k d_j} \quad (4)$$

The parameter d_i could also be computed by using the *Mahalanobis* distance (see Kohonen [5] or Bishop [7]):

$$d_i = \exp \left[-\frac{1}{2} (x - ctr_i)^T [\mathbf{M}_i]^{-1} (x - ctr_i) \right] \quad (5)$$

where $\mathbf{M}_i$ is the covariance matrix computed considering only the training input vectors x_j assigned to the i^{th} cluster:



$$\mathbf{M} = E\left[(x_j - ctr_i)^T (x_j - ctr_i)\right] \quad (6)$$

Eqn. (6) adjusts the form of the radial basis function to an elliptical one. However, such procedure is computational intensive and in this article the simpler eqn. (3) was used.

4 Choosing the experts

The expert candidates should be chosen in such a way that the collection of experts represents different types of modeling techniques. In this article the Artificial Neural Networks (ANN) model was chosen for its *non-linear* properties and the Multiple Regression Analysis (MRA) model was chosen for its *linear* properties.

Regarding the ANN model, a multi-layer perceptron (MLP) neural network can be seen as a non-linear regression model when used as a time series forecasting technique. The MLP neural network was chosen because several time series forecasting problems can be addressed by it. There are many training algorithms that can be used to develop MLP models. Among these algorithms we selected the *Back-Propagation*, a gradient-type algorithm which aims at minimizing the root mean squared error (RMSE) between observed values and model outputs (Nascimento Jr. and Yoneyama [1]).

The MRA model uses *ordinary least squares* in order to find the best linear function that fits the given input-output data. When using MRA, several hypotheses are of fundamental importance for the good quality of the results. Pindyck and Rubinfeld [9] and Gujarati [10] recommend checking the normal distribution of the output error, homoscedasticity and the absence of multicollinearity in the data, serial correlation of the output error and the presence of outliers.

Finally, the Carbon Copy (CRB) model, which is considered a very naïve and simple technique, was used as a reference model. The CRB model simply states that the observation at time t becomes the model value at time $t+1$, that is:

$$\hat{Y}_{t+1} = Y_t \quad (7)$$

5 MLEM as a data mining tool to predict the price of a commodity

This section presents a case study where the goal is to predict the deflated daily price of Sugar No. 14 contract in the New York Board of Trade (NYBOT) using a Mixture of Local Experts Models (MLEM). The use of mixture of local experts to forecast the deflated monthly price of Sugar No. 14 contracts in the NYBOT can be seen in Melo, Nascimento Jr. and Milioni[11]. The time series used in this case has 3521 consecutive daily prices (around 14 years, from 03/Jan/1989 to 03/Feb/2003). The first 3231 points of the time series were used as data to construct the models (training set: 2996 first points, test set: 235 last points). The last 260 points were used only to validate the models (validation data set) and



were not used during the construction of the MLEM phase. The real price commodity was obtained using the CPI (Consumer Price Index) as a deflator. Since CPI is presented in a monthly basis, we computed a linear adjustment among sequence values in order to have it too in a daily basis, as the commodity price. The reference date to obtain daily real price was 1-january-2003. The behavior of the time series (training and test data sets) can be seen in figure 2 that shows both nominal price as the real price (in bold).

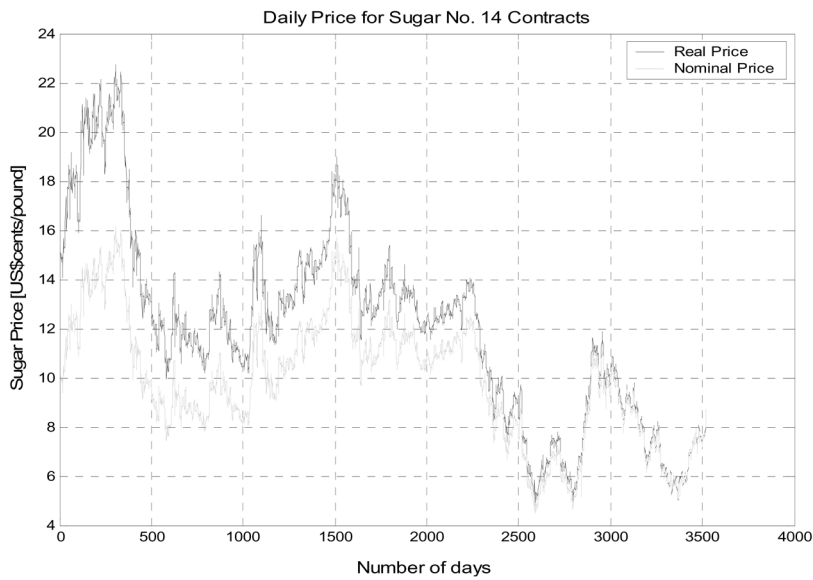


Figure 2: Time Series with the commodity daily price.

A window size of 10 points was used for all experts, that is, the points Y_{t-9} , Y_{t-8} , ..., Y_t where used as model inputs in order to predict the model output Y_{t+1} . This window size was chosen in order to capture at least the past 2 weeks of this commodity trading.

The first step was to use a Kohonen neural network over the 3231 data points (training and test sets). The data set was divided into 6 regions. Table 1 summarizes how the data points used to construct the models were distributed:

Table 1: Distribution of the points in the data set used to construct the MLEM.

Region Data Set	I	II	III	IV	V	VI	Total
Training	233	233	1017	608	369	536	2996
Test	--	--	--	--	95	140	235
Total	233	233	1017	608	464	676	3231



Four modeling techniques were used to develop the 4 (four) local experts for each one of the 6 regions, considering only the data points in the training set of the region:

- ANN I (Artificial Neural Network I): *Multi Layer Perceptrons* without *Input - Output* direct connection and a $[10, N_h, 1]$ topology, which means 10 input units, N_h hidden units and 1 output unit.
- ANN II (Artificial Neural Network II): *Multi Layer Perceptrons* with linear *Input - Output* direct connection and a $[10, N_h, 1]$ topology II.
- MRA (Multiple Regression Analysis): the model used to predict Y_{t+1} was a linear combination of Y_t , Y_{t-1} , ..., Y_{t-9} and their inverse and logarithmic transformations.
- CRB (Carbon Copy): $\hat{Y}_{t+1} = Y_t$.

The main difference between the two types of ANN is the direct input-output connections. According to Weigend & Gershenfeld [12], the direct linear connections between each input and the output units can help the training algorithm to quickly find the linear input-output relationship (if it exists) and then the hidden nonlinear units can be used to find the nonlinear part of the desired input-output mapping.

The second step in order to construct the desired models is to find the best local expert for each of the 6 regions. The best local expert for a region is the expert that has the smallest RMSE (Root Mean Squared Error) measured using only the data points in the test set of that region. However, if there is no test data points in a region, the RMSE is measured considering the training data points assigned to that region.

Finally, in the validation phase, a one-step ahead forecast is performed for the 260 points in the validation set using the mixture of the best local experts, as described by eqns. (3) and (4).

For comparison purposes, the same 4 modeling techniques (ANN I, ANN II, MRA and CRB) were used to develop the so-called “global experts”, that is, considering that the training data set (2996 points) formed just one region.

6 Analysis of the experimental results

For regions I, II and IV the best local expert was the ANN I model, while for region III the best local expert was the ANN II model, considering the RMSE measured on the *training* data set for those regions (since these regions have no *test* data set). For regions V and VI the best local expert was the MRA considering the RMSE measured on the *test* data set (if the RMSE was measured on the *training* data set the best local expert for region V would be the ANN I model and for region VI would be the ANN II).

So, the one-step ahead forecast for the 260 points in the validation data set was calculated by eqn. (8) as follows:

$$\hat{Y}_{t+1} = g_I \hat{Y}_{t+1}^{ANN I} + g_{II} \hat{Y}_{t+1}^{ANN II} + g_{III} \hat{Y}_{t+1}^{ANN II} + g_{IV} \hat{Y}_{t+1}^{ANN I} + g_V \hat{Y}_{t+1}^{MRA} + g_{VI} \hat{Y}_{t+1}^{MRA} \quad (8)$$



Figure 3 shows the forecast for the validation data set made by the MLEM technique.

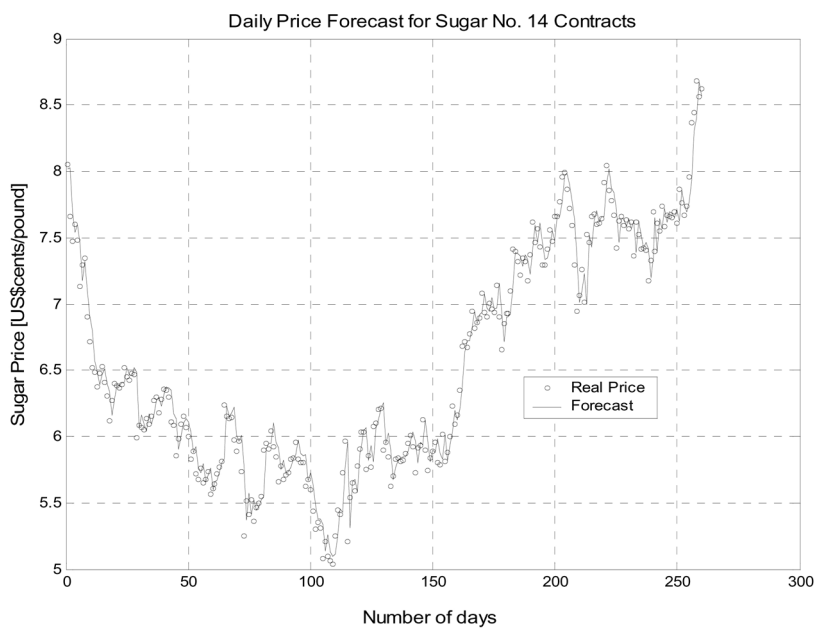


Figure 3: Forecasts for the validation data set points.

Table 2: RMSE measures for different expert models.

Expert Model	RMSE	Relative RMSE
Global ANN I	0.10943	1.2048
Global ANN II	0.15129	1.6656
Global MRA	0.09089	1.0007
Global CRB	0.09083	1.0000
MLEM	0.08982	0.9889

From this figure one can see that, although the validation data set points have a little variation ($\approx 3,5$ US\$ cents/pound), the proposed mixture of local experts technique follows the time series volatility. Table 2 shows the RMSE measures (considering the validation data set) for different expert models. Given the high volatility of this time series, it is no surprise that the best global model is the Carbon-Copy model. However, the MLEM technique produced a model that is better (even though only marginally better) than the Carbon-Copy model, which is a considerable achievement.



7 Conclusion

The proposed system for the implementation of the mixture of local experts technique can be applied to a large number of modeling problems, including the forecast of time series, a fundamental part of most decision-making processes that continues to be a challenge for researchers from all over the world.

It is important to note that, when using any modeling technique, the benefits must be large enough to outweigh the costs. Therefore any improvement obtained by the application of the technique proposed in this article should be weighted against the increased complexity and extra burden needed to implement the technique. The costs and results when one uses global experts based on single techniques should also be considered in this comparison, as mentioned by Khotanzad et al [13].

Future research will investigate other clustering procedures, more modelling techniques to develop the local expert models, and different strategies to combine the local expert models.

References

- [1] Nascimento Jr, C. L. & Yoneyama, T. , *Inteligência Artificial em Controle e Automação*, Editora Edgard Blücher, São Paulo, 2000 (in Portuguese).
- [2] Makridakis, S.; Wheelwright, S. & Hyndman, R. J., *Forecasting Methods and Applications*, 3rd ed., John Wiley & Sons, New York, 1998.
- [3] Melo, Brício; *Previsão de Séries Temporais Usando Modelos da Composição de Especialistas Locais*. MSc Thesis, Instituto Tecnológico de Aeronáutica – ITA, São José dos Campos (SP), 2003 (in Portuguese).
- [4] Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J. & Hinton, G. E., Adaptive Mixture of Local Experts, *Neural Computation*. Vol. 3, No. 1, pp.79-87, MIT Press, 1991.
- [5] Kohonen, T., *Self-Organization and Associative Memory*. 3rd ed., Springer-Verlag, Berlin, 1989.
- [6] Haykin, S., *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice Hall, New York, 1994.
- [7] Bishop, Christopher M., *Neural Networks for Pattern Recognition*, New York, USA: Oxford University Press Inc, 1995.
- [8] Mitchell, Tom Michael, *Machine Learning*, Singapore: McGraw-Hill, 1997.
- [9] Pindyck, R. S. & Rubinfeld, D. L., *Econometric models and economic forecasts*, 4th ed., McGraw-Hill, New York, 1998.
- [10] Gujarati, D. N., *Econometria Básica*. Makron Books, São Paulo, 2000 (in Portuguese).
- [11] Melo, Brício; Nascimento Jr, Cairo Lúcio & Milioni, Armando Zeferino, *Time Series Forecasting Using Mixture of Local Expert Models*. Proc. of the IX International Conference on Industrial Engineering and Operations Management, Ouro Preto, Brazil, October, 21th to 24th, pp 165-171, 2003.



- [12] Weigend, A. S. & Gershenfeld, N. A.; *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison Wesley, Reading, 1994.
- [13] Khotanzad, A.; Elragal, H.; Lu, T.L., Combination of Artificial Neural Network Forecasters for Prediction of Natural Gas Consumption. *IEEE Transactions on Neural Networks*, Vol. 11, No. 2, p.464-473, 2000.

