Time Delayed Independent Component Analysis for Data Quality Monitoring

José Márcio Faier
Signal Processing Laboratory, COPPE/Poli
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, BRAZIL
faier@lps.ufrj.br

José Manoel de Seixas
Signal Processing Laboratory, COPPE/Poli
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, BRAZIL
seixas@lps.ufrj.br

Abstract— In the information era, databases in companies and research centers are getting larger, which makes the quality of data a key issue. In this paper, time delayed independent component analysis is used for data quality monitoring of electric load time series. The independent component analysis (ICA) was applied in the preprocessing phase, which increased the data quality system performance. The extraction of signal sources reduced the forecast error, revealed relevant information and narrowed the validation corridor width.

Keywords- Data Quality; Time Series; Independent Component Analysis; Neural Networks.

I. INTRODUCTION

In present days, the global development is due, in large part, to wide data dissemination, especially due to Internet. In fact, with the enormous data volume increase, the attention has turned to the ability to absorb information and respond appropriately [1]. Thus, data quality issues have become a key factor to the transformation from data to relevant information.

Data quality is the level of correctness, completeness, consistency, interpretability, security, aggregated information and other data characteristics [2]. These data quality dimensions must be specified and monitored in accordance to user specifications.

In the electric sector, data quality study is important due the recent increase on electric load demand, especially in emerging countries, such as Brazil. The demand increases has resulted companies fusion (data integration from different systems), decisions to avoid blackout and other decisions to manager the electric system.

In this work, a data quality monitoring system is developed to analyze electric load time series with respect to the peak energy. The methodology uses adjacent series with respect to the peak hour, the daily peak series and temperature series. These data contain fundamental patterns that impact significantly a number of decision taking processes and they should not be corrupted. Thus, a data quality monitoring may identify problems and, eventually, correct mistakes and enrich the information, in accordance to user specifications.

To monitor key data quality dimensions in this time series, a validation corridor is proposed for evaluating an incoming sample included in the database and correct for it, if necessary/requested. Here, the corridor is built dynamically using Independent Component Analysis [3], aiming at identifying more structured data in the incoming time series. This more structured information may make the data quality monitoring system more efficient. Over the estimated independent sources, signal preprocessing is applied for removing seasonality, cycles and tendency [4]. Neural network modeling [5] estimates the target application from the resulting residual signal. The validation corridor center for data quality evaluation is the forecasted value for a given sample and its limit is proportional to the estimation error. This method allows the correction for outliers and missing data [6].

The Independent Component Analysis (ICA) is a statistical technique to find hidden factors in observed signals. ICA defines a model generator from observed data, which are assumed to be mixtures of unknown independent variables (sources). ICA has been used as an auxiliary tool in autoregressive processes for time series forecast [7].

The paper is organized as it follows. In the next section, a more detailed explanation of the data quality monitoring system is made. Section III presents the methodology used in the case study in data quality monitoring for electric load time series, which is conducted in Section IV. Conclusions are derived in Section V.

II. TIME SERIES DATA QUALITY MONITORING

The aim of the data quality monitoring system is to evaluate the quality of a new sample, which is to be incorporated into the database, and correct for the incoming sample, if necessary. The system is built as a control system [6], where past samples are used to build the time series model and produce a validation corridor, within which the incoming sample should stay (see Fig. 1).

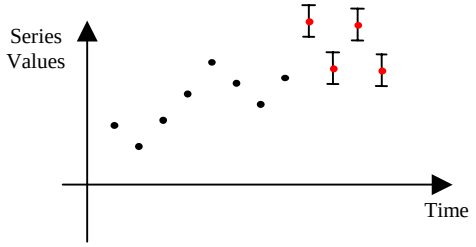

Fig. 1. The validation corridor concept

The validation corridor is defined dynamically, at moment n, by the mean absolute error ($\mu_{|error|}$) between estimated (x_{exti}) and real (x_i) sample values, adjusted by a constant to define the missing/fail probability:

$$Corridor(n) = 2.k.\frac{\sum_{i=1}^{n-1}\left|x_{est_i} - x_i\right|}{n-1} = 2.k.\mu_{|error|} \quad (1)$$

The k parameter allows include the user role and determine a compromise between the context and the user specifications. Typically, k is adjusted to detect theoretical presence of soft outliers in training set (one outlier for 150 samples).

The time series model (corridor centers) is derived from preprocessed data estimations (see Fig. 2). The preprocessing stage extracts any modeled component [4].

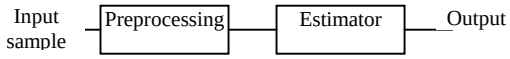

Figure 2. Basic block diagram of the monitoring system

The presence of seasonality and cycles are analyzed in frequency-domain by the Fast Fourier Transform [8]. To remove the identified components from data, it is verified their significance level. Next, the presence of heteroscedasticity is analyzed with Goldfeld-Quandt test [9]. In case of heteroscedasticity, an appropriate action, such as the application of the logarithmic function, should be considered. With homoscedastic series, the tendency is analyzed. For this, a combination of the Dickey-Fuller (ADF) [10] and Phillips-Perron [11] tests is used. Such test combination checks for unit roots in time series. In case of finding unit roots, the trend is stochastic and the first difference is applied m times (where m is the integration order of the process). If the test does not detect unit roots, the trend is deterministic, and it is removed from a polynomial fitting.

The estimator block is performed by a linear model or a neural network. Neural estimators for time series forecasting have been widely used [12]. It has been shown that neural systems are most effective when input data are preprocessed. In recent works, neural estimators have been fed from a residue series, which is obtained at the output of the preprocessing phase [4][6]. This residual information is the result of subtracting from the incoming raw data the modeled time series components (tendency, seasonality, cycles, etc), obtained from the preprocessing block. Therefore, the estimator aims at forecasting what is unknown from data.

In this work, we proposed to include an ICA block to the data preprocessing chain (see Fig. 3). The ICA finds the independent sources (**y**) derived from the observed signals (**x**). In time series, ICA finds independent signals analyzing delayed correlations to estimate the demixing matrix **B** [3]:

$$\mathbf{y} = \mathbf{B}\mathbf{x} \quad (2)$$

If an independent component is assigned to noise, through correlating components to the sources obtained for modeling, deflation may be applied. Next, after the preprocessing step (PP block - Fig. 3), according to what was described before, the system evaluates the correlation (CORR block - Fig. 3) of delayed versions of the estimated (independent preprocessed sources) sources, in order to determine the relevant samples which will be fed into estimator input nodes. If correlations are above a threshold, the qualified samples feed the estimator (EST block – Fig. 3).

The estimator design is based on parsimonious criterion [13]. From simple models, the complexity is gradually increased and evaluated. In the non-linear case, non-recurrent (mult-layer perceptron - MLP [5]) and recurrent (Elman [14]) networks are used. Thus, from a single hidden neuron, the number of hidden neurons is increased until hypothesis test rejection. Besides, early stop of network training is applied to avoid over training [5]. In the linear case, the estimator is an auto-regressive moving average (ARMA [12]) model.

The forecasted time series is reconstructed from the modeling part of the sources (block PP^{-1} - Fig. 3), resulting in the estimated sources ($\mathbf{y}_{est}$). The ICA process is reversed (ICA^{-1} block) and the forecasted values ($\mathbf{x}_{est}$) are obtained. From $\mathbf{x}_{est}$, the corridor is finally constructed for the original data space.

For data quality assessment, the data samples should remain within the corridor limits. Thus, the aim is to obtain a corridor as narrow as possible, for detecting errors and allowing their correction with good accuracy, if necessary / requested.

III. METHODOLOGY

The data quality monitoring system was analyzed in the

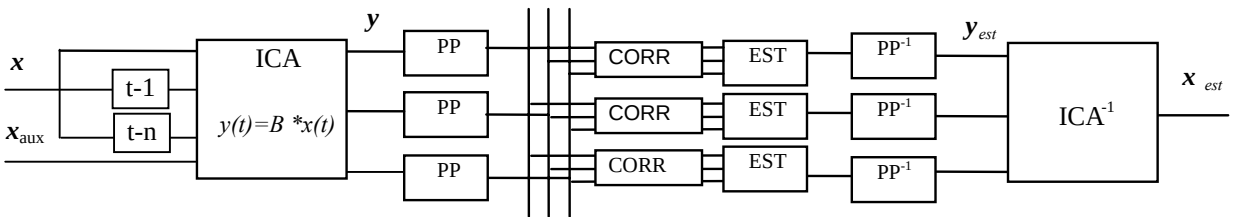

Figure 3. Data Quality Monitoring System structure.

framework of the electric load time series from a European energy supplier (East-Slovakia Power Distribution Company), which was used in a competition in 2001 by the European Network on Intelligent Technologies for Smart Adaptive Systems [15]. This database comprises electric load series, in MW, collected every thirty minutes from 01 January 1997 to 31 January 1999 and the daily temperature averaged in °C, covering the same time period. In the competition held in 2001, the competition task was to develop models to forecast daily peak load for January 1999. Here, the 1997 period was used for time series training, the year 1998 for training validation and January/1999 for testing (generalization).

Besides daily load peak, groups of series near the mean peak time (20:00) were also considered. Thus, seven adjacent series between 18:30 and 21:30 were used for series modeling. The temperature was used as an auxiliary series.

The validation corridor was estimated with the *k* constant defined by fail/missing probability on training / validation set. The constant *k* was determined assuming one outlier for 150 samples.

For finding independent sources (**y**) – see (1) -, a specific method for time series was employed. The method finds a demixing matrix (**B**) by diagonalization of the Delayed-Auto-Cross-Covariance Matrix:

$$C_{\tau}^{x} = E\left\{x(n)x(n-\tau)^{T}\right\} \qquad (3)$$

Where, n is the time sequence, τ is the time delayed and ***x*** the mixing series.

We use the Second Order Blind Identification algorithm with Robust Orthogonalization (SOBI-RO [16][17]) for find independent components. This method diagonalizes a group of Delayed-Auto-Cross-Covariance Matrixes.

The series are presented to the ICA Block in parallel. Target series could also be presented delayed to ICA Block - named here time-delayed ICA (TDICA [1]) -, increasing the number of explicative series. Analysis and forecasts are performed in the ICA space and transformed back to the original space, with reversed ICA. In the original space, just the target series no delayed is considered.

The neural network input layer was constructed from sources considering samples delayed – time-delayed neural networks (TDNN). We used a correlation test, typically 0.05 for threshold, to find relevant input delays.

For hidden layer, the hypothesis testing of a model with n neurons against the hypothesis of n +1 neurons was computed for 5% significance level.

The results were analyzed using performance indexes evaluated over the test set: the normalized mean square errors (NMSE). The index $NMSE_1$ normalizes the MSE with respect to the variance of the estimated series - see (4) -, and $NMSE_2$ uses the best random walk estimator as the normalization factor - see (5).

[1] Delays for TDICA not confuse with delays for Delayed-Auto-Cross-Covariance Matrix.

$$NMSE_1 = \frac{MSE}{\sigma_x^2} = \frac{E\left[(x_{est} - x)^2\right]}{E\left[(\mu_x - x)^2\right]} \qquad (4)$$

$$NMSE_2 = \frac{E\left[(x_{est} - x)^2\right]}{E\left[(x_{t-1} - x)^2\right]} \qquad (5)$$

The corridor center with $NMSE_1$ smaller than 1 is better than a corridor constructed with the mean of the process (μ_x). The same occurs when the NMSE2 is smaller than 1 and the corridor is constructed using the sample immediately before (x_{t-1}).

The results with and without ICA were also compared using three others performance indexes. The R indicator is the rate between correlations from original series and forecasted series delayed one sample (Lag_1) and no delayed (Lag_0) – see (6). The MAC indicator is the mean absolute corridor width, given in MW, and the MAPE is the mean absolute percentage error between forecasted and real sample - see (7) and (8).

$$R = \frac{Lag_{t=0}}{Lag_{t=1}} \qquad (6)$$

$$MAC = 2.k.\frac{\sum_{i=1}^{N}\left|x_{est_i} - x_i\right|}{N} \qquad (7)$$

$$MAPE = \frac{\sum_{i=1}^{N}\left|\frac{x_{est} - x_i}{x_i}\right|}{N} x100 \qquad (8)$$

Here, x_i and x_{est_i} are observed and forecasted samples, at time instant i, respectively, N is the number of samples, and *k* is the corridor adjustment constant obtained during the train phase.

IV. Analisys and Results

The best results with ICA are obtained trough second order statistics with robust orthogonalization (SOBI-RO), diagonalizing the first 225 delayed-cross-covariance matrixes. Better performance is achieved using TDICA with 02 delays for target series and without delays for temperature or other explicative series. The preprocessing extracted frequency components above 06 standard deviations with respect to the mean amplitude value. After the removal of cycles and seasonality, the hypothesis tests did not detect stochastic tendency, to 5% significance level. Thus, a linear trend was removed. Three standard deviations were used to define the network input samples from the correlation function. The data quality monitoring systems used both linear (ARMA) and non-linear estimators (MLP) to modeling the sources. For non-linear case, the hypothesis test defined maximum 02 hidden neurons, at a 5% significance level. The sources #1 to #3 are modeling with non-linear estimators and the others used linear models.

Table I shows $NMSE_1$ and $NMSE_2$ indexes computed from the testing series with ICA. It is observed that only $NMSE_2$ for Series #7 is around 1 and the others are well below. Then, the data quality monitoring system performance increases when compared to mean and to the best random walk estimator.

TABLE I. $NMSE_1$ AND $NMSE_2$

Series	$NMSE_1$	$NMSE_2$
Series #1 (18:30)	0,44	0,51
Series #2 (19h)	0,44	0,62
Series #3 (19:30)	‘0,37	0,39
Series #4 (20h)	0,39	0,48
Series #5 (20:30)	0,24	0,29
Series #6 (21h)	0,27	0,51
Series #7 (21:30)	0,52	1,01
Series #8 (Peak)	0,26	0,26

Without ICA, for all series, the best model was an ARMA with maximum 20 delays and no feedback, becoming a Moving Average (MA) model. The preprocessing was equal for both with and without ICA.

Table II shows the performance for both using or not ICA. The best results for each case are expressed as boldfaced values. In general, ICA performed better. For all series modeled with ICA, R is above 1. Without ICA, R indicator is below 1 for series #1, #6 and #7, indicating worse performance. Also, in general, the corridor is narrower and the MAPE is smaller, when ICA is used in the preprocessing chain.

Figure 4 shows five sources extracted from ICA, indicating that temperature (first series and first source) is one of the estimated sources. The second source suggests a semester dependency. The third source is from annual variation and the others suggest a trimester dependency. The others sources did not show an easy interpretation in the context of the application. This ability to identifying original and better structured information proved here to facilitate the work of the estimation block.

TABLE II. RESULTS FOR MAPE, MAC AND R

Series	With ICA (SOBI-RO)			Without ICA		
	MAPE (%)	MAC (MW)	R	MAPE (%)	MAC (MW)	R
Series #1 (18:30)	**3.1**	**158**	**1,87**	4.6	421	0,92
Series #2 (19h)	**2.5**	**150**	**1,55**	4.6	406	1,1
Series #3 (19:30)	3.1	156	**1,64**	**2.9**	**148**	1,04
Series #4 (20h)	2.2	**137**	**1,64**	2.2	151	1,28
Series #5 (20:30)	2.2	**124**	**1,3**	2.2	136	1,06
Series #6 (21h)	**1.9**	136	**1,18**	1.9	**125**	0,97
Series #7 (21:30)	**2.4**	**127**	**1,2**	3.1	158	0,94
Series #8 (Peak)	**2.0**	153	**1,85**	4.0	**149**	1,1

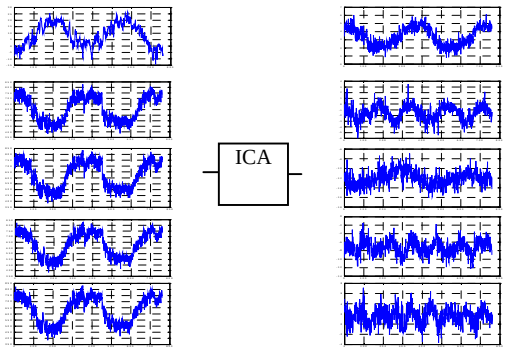

Figure 4. On the left, the series for temperature (top) and series from #1to #4. On the right, independent sources obtained through ICA block.

Figure 6 shows the real values, the validation corridor centers and the corridor width for the peak series, when ICA is used in the series preprocessing chain and MLP and ARMA are used to modeling the sources.

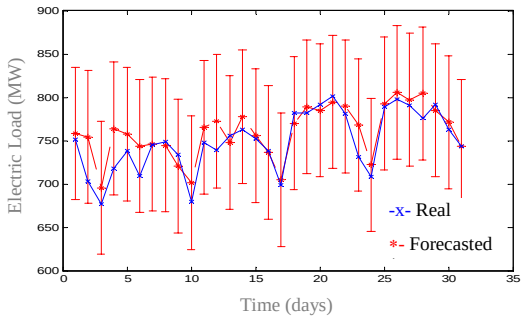

Figure 6. Real and forecasted peak series and the validation corridor for Jan/1999.

V. CONCLUSIONS

The data quality monitoring system proposed here uses a validation corridor to evaluate incoming samples of a target time series. The corridor is built around a forecasted value that is obtained from preprocessed data. The dynamic corridor adapts to the series statistical variations and the system alerts the user when the incoming sample is out of the corridor limits. In case a correction is required from the expert user or a missing value is detected, forecasted sample may be used.

In the proposed system, neural networks and linear models were applied in combination with Time Delayed Independent Component Analysis and a pre-processing stage. The parsimonious models presented better results (linear or non-linear models with few neurons). The impact of ICA was analyzed for a particular case of electric load time series. The ICA algorithm used second-order statistics with time sequence analyses (SOBI-RO) to extract the independent sources. The neural (recurrent and non-recurrent) or linear model operated over preprocessed independent sources (analyzing and removing heteroscedasticity, trends, cycles and seasonality). The monitoring system proposed, using ICA, reduced the validation corridors and improved the forecast performance, which have positive impact on data quality dimensions such as correctness, completeness, outdating, and interpretability.

ACKNOWLEDGMENT

We are thankful to CNPq and FAPERJ (Brazil) for their support to this work.

REFERENCES

[1] Eckerson, W. W. (2002). Data Quality and the Botton Line, Report, The Data Warehousing Institute.

[2] Chrisman, N. R. (1983). The Role of Quality Information in the Long-Term Functioning of a GIS. In: Proceedings of the AUTOCART06, v. 2, pp. 303-321.

[3] Hyvarinen, A., Karhunen, J. e Oja, E., (2001). Independent Component Analysis, ISBN 0-471-40540-X John Wiley & Sons, Inc.

[4] DANTAS, A. C. H. ; DINIZ, F. C. da C. B. ; FERREIRA, T. N. ; SEIXAS, J. M. de . Statistical and Signal Processing Based

System for Data Quality Management. In: IV International Conference on Data Mining Including Building Applications for CRM & Competitive Intelligence, 2003, Rio de Janeiro. Proceedings of the IV International Conference on Data Mining Including Building Applications for CRM & Competitive Intelligence, 2003. p. 01-10.

[5] Haykin, Simon. (2008) Neural Networks and Learning Machines, 2da. Edition ISBN 0131471392, Prentice Hall.

[6] DANTAS, A. C. H. ; SEIXAS, J. M. de . Neural Networks for Data Quality Monitoring of Time Series. In: 9th International Conference on Enterprise Information Systems, 2007, Funchal, Madeira. Proceedings of the 9th International Conference on Enterprise Information Systems, 2007. p. 411-415.

[7] Kiviluoto, K., Oja, E. (1998). Independent Component Analysis for Parallel Financial Time Series. In Proc. Int. Conf. on Neural Information Processing (ICONIP'98), v. 2, pp. 895-898, Tokyo, Japan.

[8] Brigham, E.O. (2002), The Fast Fourier Transform, New York: Prentice-Hall .

[9] S.M. Goldfeld and R.E. Quandt (1965), "Some Tests for Homoscedasticity". Journal of the American Statistical Association 60, 539–547.

[10] Dickey, D. A. and Fuller, W. A. (1979). Distributions of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, v. 75, pp. 427-431.

[11] Phillips, P. C. B. (1987). Time series regression with a unit root. Econometrica, v. 55, n. 2, pp. 277-301.

[12] George Box, Gwilym M. Jenkins, and Gregory C. Reinsel. Time Series Analysis: Forecasting and Control, third edition. Prentice-Hall, 1994.

[13] Medeiros, M. C., Teraasvirta, T., Rech, G. (2006). Building Neural Network Time Series Models: A Statistical Approach, Journal of Forecasting, v. 25, n. 1, pp. 49-75.

[14] Elman, J. L. (1990). Finding structure in time. Cognitive Science, v. 14, pp. 179-211.

[15] EUNITE, European Network on Intelligent Tech-nologies for Smart Adaptive Systems (2001), http://neuron.tuke.sk/competition.

[16] Belouchrani, A., Abedi-Meraim, K., Cardoso, J., Moulines, E. (1997). A Blind Source Separation Technique Using Second Order Statistics. IEEE Transactions on Signal Processing, 45 (2):434-444.

[17] Belouchrani, A., Cichocki, A. (2001). Robust whitening procedure in blind source separation context, Electronics Letters, Vol. 36, No. 24, pp. 2050-2053.