DATA QUALITY MONITORING: INDEPENDENT COMPONENT ANALYSIS FOR TIME SERIES

José Márcio Faier, and José Manoel de Seixas

Signal Processing Laboratory, COPPE/Poli, Federal University of Rio de Janeiro,
Rio de Janeiro, RJ, BRAZIL
(faier@lps.ufrj.br, and seixas@lps.ufrj.br)
http://www.lps.ufrj.br

ABSTRACT

In the information era, databases in companies and research centres are getting larger, which makes the quality of data a key issue. In this paper, independent component analysis is used for data quality monitoring of electric load time series. The independent component analysis was applied in the pre-processing phase, which increased the data quality system performance. The extraction of signal sources revealed relevant information, and narrowed the corridor width used for data validation.

1. 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, aggregated information and other data context dependent characteristics [2]. These data quality dimensions must be specified and monitored in accordance to user specifications. The users define what is high or low quality.

In the electric sector, data quality studies are even more important due to the recent increase on electric load demands, especially in emerging countries, such as Brazil. The demand increases have resulted in companies fusion (data integration from different systems), and decisions must be taken to avoid blackout and to manager the electric system.

In this work, a data quality monitoring system is developed to analyse 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 system may identify problems and, eventually, correct for 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 pre-processing is applied for removing seasonality, cycles and tendency [4]. Neural network [5] or linear modelling [6] estimates the target application from the resulting residual signal. The validation corridor centre 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 [7].

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 [8]. It has been shown that the estimated sources concern structured data, which can be used for data prospection or signal processing

The paper is organized as it follows. In the next section, a more detailed explanation of the data quality monitoring system is given. Section 3 presents the methodology used in the case study of data quality monitoring for electric load time series, which is conducted in Section 4. Conclusions are derived in Section 5.

2. 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/requested. The system is built as a control system [7], 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).

The validation corridor is defined dynamically, at sampling time instant n, by the mean absolute error ($\mu_{|error|}$) between estimated (x_{est_i}) and real (x_i) sample values, and it is adjusted by a constant to define the missing/fail probability (1).

 ISSN 2076-1465

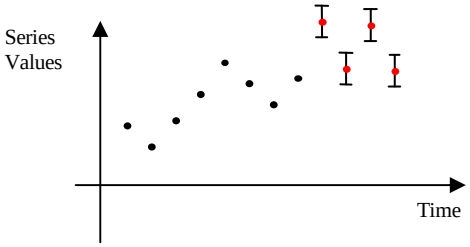

Fig. 1. The validation corridor concept.

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

The k parameter allows to include the user role and determines 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 each 150 samples).

The time series model (corridor centres) is derived from pre-processed data estimations (see Fig. 2). The pre-processing stage extracts typical time series components such as seasonality, cycles, and tendency [4].

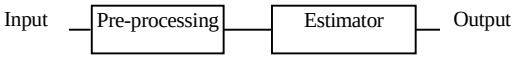

Figure 2. Basic block diagram for determining the corridor centres of the monitoring system

The presence of seasonality and cycles are analysed in frequency-domain by the Fast Fourier Transform [9]. From spectral information, we remove the identified components verifying their significance level. Significant frequency component, above a threshold, are subtracted from the original time series.

Next, the presence of heteroscedasticity is analysed with Goldfeld-Quandt test [10]. In case of heteroscedasticity, an appropriate action, such as the application of the logarithmic function, should be considered. With homoscedastic series, the tendency is analysed. For this, a combination of the Dickey-Fuller (ADF) [11] and Phillips-Perron [12] 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 either by a linear model or a neural network. Neural estimators for time series forecasting have been widely used [6]. It has been shown that neural systems are most effective when input data are pre-processed. In recent works [4][7], neural estimators have been fed from a residue series, which is obtained at the output of the pre-processing phase, as from Figure 2. This residual information is the result of subtracting from the incoming raw data the modelled time series components (tendency, seasonality, cycles), obtained from the pre-processing block. Therefore, the estimator aims at forecasting what is unknown from data.

In this work, we proposed to include an ICA block to the pre-processing chain (see Fig. 3). The aim is to access more structured signals with respect to the original data and facilitate the data pre-processing step. In the sequence, the estimator block (EST) models the pre-processed (PP) residue in the ICA space.

The ICA finds the independent sources ($\boldsymbol{y}$) derived from the observed signals ($\boldsymbol{x}$), estimating the de-mixing matrix $\boldsymbol{B}$ [3] – see (2). If an independent component is assigned to noise, deflation may be applied (ICA block).

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

The estimator design is based on parsimonious criterion [13]. From simple models (linear models), the complexity is gradually increased by introducing non-linear neurons, and evaluated. Thus, from a single hidden neuron, the number of hidden neurons is increased until the error decrease hypothesis can be rejected. Early stop of the neural network training is applied to avoid over training [5]. In the non-linear case, we use feed forward (mult-layer perceptron - MLP [5]) neural networks. In the linear case, the estimator is an auto-regressive moving average (ARMA [6]) model.

The estimated time series is reconstructed over the modelled sources (block PP^{-1} - Fig. 3), resulting in the estimated sources ($\boldsymbol{y}_{est}$). The ICA process is then reversed (ICA^{-1} block – Fig. 3) and the estimated values ($\boldsymbol{x}_{est}$) are obtained. From $\boldsymbol{x}_{est}$, the corridor is finally constructed for the original data space – see (1).

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 but emitting only correct alerts, for detecting errors and allowing their correction with good accuracy, if necessary / requested.

3. METHODOLOGY

The data quality monitoring system was analysed in the 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 [14]. This database comprises electric load series, in MW, collected every thirty minutes from 01 January 1997 to 31 January 1999 and the daily temperature

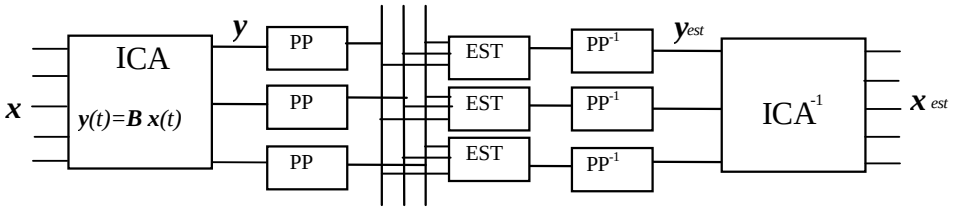

Figure 3. The structure of the Data Quality Monitoring System. The estimator block (EST) models the preprocessed (PP) residue in the ICA space. Finally, both pre-processing and ICA are reversed, and the model is mapped onto the original space.

averaged in °C, covering the same time period. In the competition held in 2001, the competition task was to develop models to forecast the daily peak load along 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 the daily load peak series, 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 modelling. 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 equation (1) -, a specific time series algorithm was employed. The method finds a de-mixing matrix (***B***) by diagonalizing the Delayed-Auto-Cross-Covariance Matrix:

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

where, *n* is the time sequence, τ is a time delay and ***x*** the observed series.

We use the Second Order Blind Identification algorithm with Robust Orthogonalization (SOBI-RO [15]) for determining the independent components. This method first whitens data and then diagonalizes a group of Delayed-Auto-Cross-Covariance Matrixes. The time series are presented to the ICA Block in parallel. Analysis and forecasts are performed in the ICA space and transformed back to the original space using reversed ICA.

The neural network input layer was constructed from sources considering delayed samples. We used a correlation test, typically 95% 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. If the output error increases when a neuron is included, at 95% confidence level, the *n+1* hypothesis is rejected.

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

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

$$NMSE_2 = \frac{E\left|(x_{est} - x)^2\right|}{E\left|(x_{n-1} - x)^2\right|} \tag{5}$$

A corridor centre 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 $NMSE_2$ is smaller than 1 and the corridor is constructed using the sample from the previous time instant (x_{n-1}).

The results with and without the application of ICA were also compared using three others performance indexes. The R indicator is the rate between correlations from original series and forecasted series delayed from one sample ($Lag_{n=1}$) and without any delay ($Lag_{n=0}$) – 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).

$$R = \frac{Lag_{n=0}}{Lag_{n=1}} \tag{6}$$

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

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

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.

4. ANALISYS AND RESULTS

The better performance could be explained, in part, by robust orthogonalization, which minimizes the noise effects. Besides, the second order algorithm diagonalizes the first 225 delayed-cross-covariance matrixes – see τ (3) -, performing more than orthogonalization. In fact, performing the independence [3].

The pre-processing extracted frequency components above 06 standard deviations with respect to the mean amplitude value. The spectral information from independent sources is clearer than without ICA block, which facilitates the pre-processing.

The input network is defined using the spatial-temporal correlation function. When ICA block was included, there were not significant correlations between the independent sources, only temporal correlations, which simplified the estimation. The data quality monitoring system used both linear (ARMA) and non-linear estimators (MLP) for residue source modelling. For non-linear case, the hypothesis test defined maximum 02 hidden neurons. The first three sources were modelled with non-linear estimators (MLP) and, to the others sources, linear models (ARMA) proved to be enough.

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 the usage of either the mean or the best random walk estimator.

TABLE I. $NMSE_1$ and $NMSE_2$ performance indexes.

Series	$NMSE_1$	$NMSE_2$
Series #1 (18:30)	0,47	0,54
Series #2 (19h)	0,47	0,66
Series #3 (19:30)	0,39	0,42
Series #4 (20h)	0,41	0,49
Series #5 (20:30)	0,29	0,31
Series #6 (21h)	0,31	0,53
Series #7 (21:30)	0,54	1,02
Series #8 (Peak)	0,29	0,29

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

Table II shows the performance indexes for both using or not ICA. The pre-processing was automated using the same parameters for both with and without ICA. The best results for each case are expressed as boldfaced values. In general, ICA performed better. For all series modelled 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 validation corridor is narrower and the MAPE is smaller, when ICA is used in the pre-processing chain.

In Figure 4, due to the similar series shapes, we show only the first five series (temperature and series #1 to #4), and the more structured sources (sources #1 to #5). We observe that the 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 sources not showed did not allow 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. MAPE, MAC and R performance indexes with and without ICA block

Series	With ICA (SOBI-RO)			Without ICA		
	MAPE (%)	MAC (MW)	R	MAPE (%)	MAC (MW)	R
Series #1 (18:30)	**3.3**	**162**	**1,81**	4.6	421	0,92
Series #2 (19h)	**2.7**	**156**	**1,51**	4.6	406	1,1
Series #3 (19:30)	3.3	159	**1,60**	**2.9**	**148**	1,04
Series #4 (20h)	2.2	**140**	**1,61**	2.2	151	1,28
Series #5 (20:30)	2.2	**129**	**1,28**	2.2	136	1,06
Series #6 (21h)	**1.9**	139	**1,15**	1.9	**125**	0,97
Series #7 (21:30)	**2.6**	**130**	**1,16**	3.1	158	0,94
Series #8 (Peak)	**2.1**	157	**1,80**	4.0	**149**	1,1

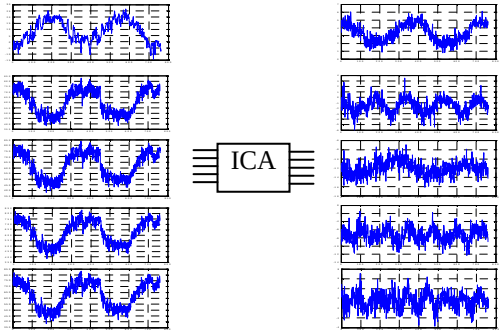

Figure 4. On the left, the first five series for temperature (top) and series from #1to #4. On the right, the more structured independent sources obtained through ICA block.

Figure 5 shows the real values, the validation corridor centers and the corridor width for the peak series, when ICA is used in the series pre-processing chain and linear model are used to modeling the sources.

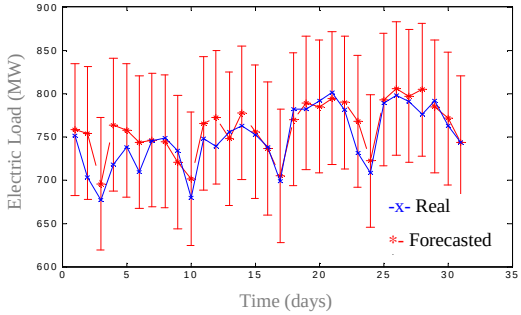

Figure 5. Actual and forecasted peak series and the validation corridor for Jan/1999.

5. 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 pre-processed 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, the forecasted sample may be used.

In the proposed system, neural networks and linear models were applied in combination with Independent Component Analysis, which was included in the pre-processing stage. The parsimonious models presented better results (linear or non-linear models with few neurons). The impact of ICA was analysed 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. Both neural and linear models operated over pre-processed independent sources (analysing and removing heteroscedasticity, trends, cycles and seasonality). Using ICA, the validation corridors were reduced and the forecast performance was improved, which produced a positive impact on data quality dimensions such as correctness, completeness, and interpretability.

ACKNOWLEDGMENT

We are thankful to ICSystems, 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; John Wiley & Sons, Inc.

[4] DANTAS, A. C. H. ; DINIZ, F. C. da C. B. ; FERREIRA, T. N. ; SEIXAS, J. M. de (2003). Statistical and Signal Processing Based System for Data Quality Management. In: IV International Conference on Data Mining Including Building Applications for CRM & Competitive Intelligence, Rio de Janeiro. pp. 01-10.

[5] Haykin, Simon. (2008) Neural Networks and Learning Machines, Second Edition, Prentice Hall.

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

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

[8] 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.

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

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

[11] 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.

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

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

[15] 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): pp. 434-444.