Barceló Cerdá, S.; Vidal Puig, S.; Ferrer, A. (2011). Comparison of multivariate statistical methods for dynamic systems modeling. Quality and Reliability Engineering International. 27(1):107-124. doi:10.1002/qre.1102.

Comparison of Multivariate Statistical Methods for Dynamic Systems Modeling

Susana Barceló*, Santiago Vidal-Puig, Alberto Ferrer
Department of Applied Statistics and Operational Research, and Quality. Universidad Politécnica de Valencia. Camino de Vera s/n 46022 Valencia, Spain

Abstract

In this paper two multivariate statistical methodologies are compared in order to estimate a MIMO (multi-input multi-output) transfer function model in an industrial polymerization process. In these contexts, process variables are usually autocorrelated (i.e., there is time-dependence between observations), posing some problems to classical linear regression models. The two methodologies to be compared are both related to the analyses of multivariate time series: Box-Jenkins methodology, and partial least squares (PLS) time series. Both methodologies are compared keeping in mind different issues, such as the simplicity of the process modeling (i.e. the steps of the identification, estimation and validation of the model), the usefulness of the graphical tools, the goodness of fit, and the parsimony of the estimated models. Real data from a polymerization process is used to illustrate the performance of the methodologies under study.

Key words: MIMO transfer function model; Box-Jenkins, PLS; Time series; Process dynamics.

1. Introduction

When the pursued objective is to implement a model-based process control system as a Model Predictive Control, the model estimation stage is critical in the development of the system. The estimated model should faithfully represent the process dynamics, for the process control system to be efficient. It must describe the dynamic relationships between the variables intervening in the process with the purpose of forecasting the process evolution that would allow an efficient control.

In many industrial processes, consecutive measurements are usually auto and cross-correlated. This is a consequence of the presence of inertial elements such as the raw materials, storage tanks, reactors, refluxes, environmental conditions, etc. with dynamics larger than the sampling frequency [1]. In this context, the classical linear

*Correspondence to: Susana Barceló. Department of Applied Statistics and Operational Research, and Quality. Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain.
†*E-mail address*: sbarcelo@eio.upv.es ; Tel.: +34 963877490; fax: +34 963877499
Contract/grant sponsor: Spanish Government (MICINN) and European Union (RDE funds)
Contract/grant number: DPI2008-06880-C03-03/DPI

regression models present remarkable drawbacks. For example, to assume that the relationships between the inputs and the outputs in the system are instantaneous, when, actually, these dynamic processes show inertias and delayed responses. Another is to model the residuals (the part of the responses not explained by the predictor variables) as white noise (i.e. independent and identically normally distributed, $N(0,\sigma)$). These two restrictions that can be reasonable in static situations, in which by hypothesis we can disregard the temporary dimension of the data, are often erroneous when the variables are observed through time and a dynamic structure is expected.

The dynamic regression model associated with this type of systems can be characterized by the *discrete transfer function models*. The dynamic transfer functions come up as a generalization of the regression models to study the relationship between M predictors $x_1, x_2, ..., x_M$ and L responses $y_1, y_2, ..., y_L$ when it is suspected that the involved variables are autocorrelated and, furthermore, the relationships between them are not exclusively instantaneous and can show time delays or inertias. These models allow us to assess how the effects between variables are transmitted. They are a useful tool in process engineering, since only when the dynamic system is known, it is possible to build a reliable inferential model (*soft sensor*) and implement an efficient process control.

There are different multivariate statistical methodologies to estimate a *multivariate transfer function model* from process data. In this paper, two methodologies will be compared, the classical multivariate time series Box-Jenkins methodology [2, 3, 4], and the projective PLS time series methodology [5].

The main objective of this research is the comparison of two multivariate statistical methodologies to estimate the transfer function model (TFM) of a continuous industrial polymerization process. The quality of these models is critical in order to obtain good predictions and also in the successful implementation of a process control system such as a Model Predictive Control. These models can be estimated from the data, either as a low order transfer functions using prediction error-based methods or identifying non-parsimonious finite impulse response (FIR) functions, by using regression methods, such as ordinary least squares or biased methods such as ridge or partial least square regression. Although, directly identifying non-parsimonious FIR models have certain advantages -e.g., any complex dynamic linear system can be fitted no matter the structure of the model- [6, 7], the trade-offs involved in identifying the FIR models by directly fitting various regression methods versus first identifying low order parsimonious models using prediction error methods, have not been well documented as commented by Dayal and MacGregor [8].

The remainder of the paper is organized as follows. Section 2 describes the petrochemical process studied. Box-Jenkins methodology is described in Section 3. Section 4 introduces the PLS time series methodology. Models estimated from both methodologies are discussed in Section 5. Finally, Section 6 deals with the conclusions.

2. Process Description

The case study that serves as the basis of this comparison, involves a commercial-scale polymerization process that produces large volumes of a polymer (high-density polyethylene) of a certain grade used in many familiar consumer products.

Processing is performed continuously. This industrial process was briefly introduced by Ferrer *et al*. [9].

This is a MIMO process with two outputs (*MI* and *APRE*) and two inputs (*T* and *E*). Observations of the output polymer properties and opportunities for adjustments occur at discrete equidistant intervals of time *t*: samples of reactor effluent are taken every two hours and analyzed off-line. The key quality characteristic is polymer viscosity, which is measured by melt index (MI_t). The sampling and measurement process introduce a modest analytical error. Added to MI_t, an index of process productivity ($APRE_t$) is worked out by energy balance every two hours, in a contemporary manner to the viscosity data (MI_t). The input registered variables are both the averages of reactor temperature $T_{t\text{-}1}$ and ethylene flow $E_{t\text{-}1}$ during the two hours before *t*. Four manufacturing periods (campaigns) for the production of the same polymer grade have been investigated.

The objective of the control system is to minimize *MI* variation around a target level of 0.8 viscosity units and to keep the productivity (*APRE*) as high as possible, guaranteeing that *MI* matches the specification limits. Adjustments to these variables can be made by varying the temperature of the reactor (*T*) and ethylene flow (*E*). These are ready compensatory process variables, whose changes represent negligible cost when compared to off-target viscosity cost or low productivity level. The opportunity to adjust the process occurs immediately after values of the output variables are obtained every two hours, so that the input variables (temperature and ethylene flow) are allowed to remain at the same level between observations. Figure 1 displays some typical input/output data for a particular manufacturing period. Every output is plotted against both inputs to illustrate the characteristics of these process data: autocorrelations, cross-correlations, and non-stationary behavior.

Figure 1 (here)

3. Box-Jenkins Methodology

3.1. MIMO transfer function model

The MIMO *transfer function model* is an alternative model formulation for vector time series [3, 4] used in multivariate time series analysis. In the econometric setting these models are known as simultaneous transfer function models. They allow representing the dynamic relationships between *M* input variables $x_1, x_2, ..., x_M$ and *L* output variables $y_1, y_2, ..., y_L$ of a dynamic system. The general mathematical multivariate transfer function expression in matrix form is:

$$\boldsymbol{y}_t = \boldsymbol{c} + \sum_{b=0}^{\infty} \boldsymbol{\Psi}_b^* B^b \boldsymbol{x}_t + \boldsymbol{n}_t \qquad (1)$$

where $\boldsymbol{x}_t = \left(x_{1t}, x_{2t}, ..., x_{Mt}\right)^T$ is a vector with *M* observable input variables; $\boldsymbol{y}_t = \left(y_{1t}, y_{2t}, ..., y_{Lt}\right)^T$ is a vector with *L* output variables; *B* is the backshift operator; $\boldsymbol{\Psi}_b^*$ are coefficient matrices (*L*×*M*) representing the effects that changes in the

input variables $\boldsymbol{x}_t$ have on the output variables at b time periods ahead, $\boldsymbol{y}_{t+b}$, and are called the impulse response matrices; * is a superscript indicating a matrix associated to the input variables; $\boldsymbol{c}$ is a constant vector ($L\times1$); $\boldsymbol{n}_t$ is a noise vector ($L\times1$) following a stationary vector ARMA process $\boldsymbol{\Phi}(B)\boldsymbol{n}_t = \boldsymbol{\Theta}(B)\boldsymbol{a}_t$, where $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ are $L\times L$ finite order polynomial matrix in B; and $\boldsymbol{a}_t = \left(a_{1t}, a_{2t}, \ldots, a_{Lt}\right)^T$ an L-dimensional white noise time series vector independently and identically multivariate normally distributed, $N(\mathbf{0}, \boldsymbol{\Sigma})$.

Another form of a time series model with exogenous input variables is the so-called vector ARMAX model. This form of model can be motivated by (1) assuming that the transfer function operator $\boldsymbol{\Psi}^*(B) = \sum_{b=0}^{\infty} \boldsymbol{\Psi}_b^* B^b$ can be represented in a rational factorization as $\boldsymbol{\Psi}^*(B) = \boldsymbol{\Phi}(B)^{-1}\boldsymbol{\Theta}^*(B)$, where $\boldsymbol{\Theta}^*(B) = \sum_{b=0}^{s} \boldsymbol{\Theta}_b^* B^b$ is of order s, and the $\boldsymbol{\Theta}_b^*$ are coefficient matrices ($L\times M$). For convenience, we also assume that the factor $\boldsymbol{\Phi}(B)$ in $\boldsymbol{\Psi}^*(B)$ is the same as the AR factor in the model for the noise $\boldsymbol{n}_t$. Then, the following equation:

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\Psi}^*(B)\boldsymbol{x}_t + \boldsymbol{n}_t = \boldsymbol{c} + \boldsymbol{\Phi}(B)^{-1}\boldsymbol{\Theta}^*(B)\boldsymbol{x}_t + \boldsymbol{n}_t \tag{2}$$

can be expressed as:

$$\boldsymbol{\Phi}(B)\boldsymbol{y}_t = \boldsymbol{c}' + \boldsymbol{\Theta}^*(B)\boldsymbol{x}_t + \boldsymbol{\Phi}(B)\boldsymbol{n}_t \equiv \boldsymbol{c}' + \boldsymbol{\Theta}^*(B)\boldsymbol{x}_t + \boldsymbol{\Theta}(B)\boldsymbol{a}_t \tag{3}$$

or

$$\boldsymbol{y}_t - \sum_{b=1}^{p} \boldsymbol{\Phi}_b \boldsymbol{y}_{t-b} = \boldsymbol{c}' + \sum_{b=0}^{s} \boldsymbol{\Theta}_b^* \boldsymbol{x}_{t-b} + \boldsymbol{a}_t - \sum_{b=1}^{q} \boldsymbol{\Theta}_b \boldsymbol{a}_{t-b} \tag{4}$$

which is referred to as a vector ARMAX model (the X stands for exogenous) with the exogenous input variables $\boldsymbol{x}_t$.

The main hypotheses of the model are the following:

(1) The (exogenous) input process $\{\boldsymbol{x}_t\}$ is generated independently of the noise process $\{\boldsymbol{n}_t\}$.
(2) The input process $\{\boldsymbol{x}_t\}$ can affect the output process $\{\boldsymbol{y}_t\}$, but not on the contrary, since the relationship is unidirectional.
(3) Another tacit assumption of the model is that the system to be modeled is stable, that is to say, all roots of $det\{\boldsymbol{\Phi}(B)\} = 0$ are greater than one in absolute value. In other words, it is assumed that the $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ and $\boldsymbol{n}_t$ are stationary multivariate stochastic processes that allow a convergent representation similar to (1):

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\Psi}^*(B)\boldsymbol{x}_t + \boldsymbol{\Psi}(B)\boldsymbol{a}_t = \boldsymbol{c} + \sum_{b=0}^{\infty} \boldsymbol{\Psi}_b^* \boldsymbol{x}_{t-b} + \sum_{b=0}^{\infty} \boldsymbol{\Psi}_b \boldsymbol{a}_{t-b} \qquad (5)$$

where the sums converge.

For the practical use of model (1) the transfer function individual operators $\boldsymbol{\Psi}^*(B) = \sum_{b=0}^{\infty} \boldsymbol{\Psi}_b^* B^b$ can be represented as a quotient of polynomials of finite order in the backshift operator B, $\omega_{lm}(B)/\delta_{lm}(B)$ leading to an alternative formulation of the MIMO *transfer function model*:

$$y_{lt} = c_l + \sum_{m=1}^{M} \frac{\omega_{lm}(B)}{\delta_{lm}(B)} B^b x_{mt} + \sum_{l'=1}^{L} \frac{\theta_{ll'}(B)}{\phi_{ll'}(B)} a_{lt} \; ; \quad l \in (1, \ldots, L) \quad (6)$$

where $\omega_{lm}(B), \delta_{lm}(B), \theta_{ll'}(B)$ and $\phi_{ll'}(B)$ are finite order polynomials in the backshift operator B. The described model in (6) is a multivariate generalization of the *single-input-single-output* (SISO) *transfer function model* [2].

The equation (6) expressed in vector-matrix form for the particular case of two inputs M=2 and two outputs L=2 has the following form:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \dfrac{\omega_{11}(B)B^{b_{11}}}{\delta_{11}(B)} & \dfrac{\omega_{12}(B)B^{b_{12}}}{\delta_{12}(B)} \\ \dfrac{\omega_{21}(B)B^{b_{21}}}{\delta_{21}(B)} & \dfrac{\omega_{22}(B)B^{b_{22}}}{\delta_{22}(B)} \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + \begin{bmatrix} \dfrac{\theta_{11}(B)}{\phi_{11}(B)} & \dfrac{\theta_{12}(B)}{\phi_{12}(B)} \\ \dfrac{\theta_{21}(B)}{\phi_{21}(B)} & \dfrac{\theta_{22}(B)}{\phi_{22}(B)} \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} \quad (7)$$

In order to estimate the model (6), it is assumed that this model is identifiable [10, 11]. For matching this requirement, the model has to be stable, i.e. the roots of the polynomials $\phi_{ll'}(B)$ and $\delta_{lm}(B)$ are greater than one in absolute value; and invertible, i.e. the roots of the polynomial $\theta_{ll'}(B)$ lie outside the unit circle.

3.2. Multivariate Transfer Function Identification

3.2.1. Preliminary Time Series Analyses

Before proceeding to identify the model, it is necessary to carry out some preliminary analyses of the series for the following purposes:

a) To detect outliers.
b) To check for heteroscedasticity. In this case, it can be necessary to apply some transformation to the original series to stabilize the variance like the Box-Cox transformations [12].
c) To verify whether the original series are stationary or if it is necessary to difference them.

At this step, there are some useful statistical tools, such as the time series plots and the simple autocorrelation function. These descriptive tools could even inform us about the nature of the transfer function, either in the Box -Jenkins or in the PLS time series methodologies.

As commented in Section 2, we have investigated four manufacturing periods (campaigns) for the production of the same polymer grade. The registered variables are MI_t at the outlet of the reactor and $APRE_t$ every two hours and the averages of reactor temperature $T_{t\text{-}1}$ and ethylene flow $E_{t\text{-}1}$ during the two hours before t.

An examination of the time series plots allows the detection of some potential outliers. As an example, Figure 2 shows the input series plots of campaign 4, reflecting a breakdown that takes place during the process at around t=70. An iterative procedure for joint estimation of model parameters and outliers effects developed by Chen and Liu [13] has been used to deal with this problem.

Figure 2 (here)

If the data series are nonstationary, the auto and cross-correlation functions cannot be estimated. However in the Box-Jenkins methodology, they are necessary for model identification. In our case the input series are nonstationary, as it can be seen in Figures 3 (top) and 4 (top), where input data from campaign 0 are represented as an example. In both figures the autocorrelation function fails to damp out quickly, indicating that time series needs a degree of differencing d to induce stationarity. Figures 3 (bottom) and 4 (bottom) show that after differencing both input series once (d=1), stationary is achieved. On the other hand, heteroscedasticity is not detected.

Figure 3 (here)

Figure 4 (here)

3.2.2. Prewhitening and cross-correlations

The identification process is simplified when the input series of the system are white noise. When an original input series follows any other stochastic process this simplification is possible through the prewhitening methodology. We suppose that an input series x_t^*, after being differenced the necessary number of times d, becomes a stationary series x_t that can be represented by an ARMA model. Therefore, after prewhitening the input series x_t^* with this ARIMA model, it becomes white noise. Thus, the cross-correlation function between the prewhitened input and the filtered output with the same ARIMA model is directly proportional to the impulse response function [2].

3.2.3. Study of cross-correlation between the output variables

From the study of the estimated cross-correlation functions (CCF) between the outputs MI_t and $APRE_t$ for the different campaigns, it can be concluded that the outputs are not cross-correlated. In Figure 5 a plot of the estimated CCF between the

prewhitened and filtered output series of campaign 5 is shown as example. In this case the filter is the first difference, $\nabla y_t = y_t - y_{t-1}$.

Figure 5 (here)

The non-existence of correlation between the two output variables MI_t and $APRE_t$ does not improve the efficiency of the joint estimation of the MIMO transfer function model, with respect to the estimation of both MISO (multiple-input-single-output) model (one for each output). Therefore, both MISO models were estimated independently.

3.2.4. Close-loop Transfer Function Model Identification

The general close-loop process scheme is showed in Figure 6, in which $H_1(B)$ and $H_2(B)$ are the transfer function equations representing the true model of the process; N_{1t} and N_{2t} are the noise components; MI_t and $APRE_t$ are the actual values of the process outputs; TV_1 and TV_2 are the output targets values; $C_1(B)$ and $C_2(B)$ are the transfer function equations of the controller; and T_t and E_t are the adjustment values of the process input. For economy and safety reasons, data were obtained under a modified closed-loop operation, using only temperature adjustments to control *MI*.

For identification and estimation purposes of the transfer function model of the process, Box and MacGregor [14] suggested that a persistently exciting dither signal (pseudorandom binary signal) has to be added to the intended adjustments. Nevertheless, in this case dither signal was not explicitly added to the *T* adjustments, because the magnitude of feedback adjustments were based on the particular experience of the various operators, who frequently have different control philosophies, so the dither signals can be considered as included in the manual input *T* adjustments.

Figure 6 (here)

As commented before, the estimation of the cross-correlation functions (CCF) between the different input and output is needed for transfer function model identification. In Figure 7, the estimated CCF of the prewhitened and filtered series (the filter is the first difference of both series) shows the dynamic relationships between *MI* and *T*, which are consistent in the four manufacturing periods. The effect of changing *T* at time *t* begins to show in *MI* two hours later, MI_{t+1} (cross-correlation at lag 1, r(1) statistically significant and positive) and lasts for two more hours, MI_{t+2} (cross-correlation at lag 2, r(2) statistically significant and positive). This will be taken into account in the model formulation. Note that due to the closed-loop nature of the process, in Figure 7 the CCF also shows the reaction of the operator (cross-correlation at lag 0, r(0) statistically significant and negative), who decreased or increased T_t depending on MI_t laboratory measurement.

Figure 7 (here)

Figure 8 shows the CCF representing the dynamic relationships between ∇MI and ∇E (again the filter is the first difference) which are consistent enough in the four campaigns. The effect of changing *E* at time *t*, begins to show in *MI* two hours later,

MI_{t+1} (r(1) statistically significant in campaign 2 and 4) with positive effect, and lasts for two more hours MI_{t+2} (r(2) statistically significant in campaign 2 and 5), with a negative effect. Note that there is no instantaneous relationship between both variables (r(0) statistically non-significant) because in this process *MI* was not controlled based on ethylene flow measurements.

Figure 8 (here)

In figure 9 the cross-correlation functions show the dynamic relationships between $\nabla APRE$ and ∇T (again the filter is the first difference). It is shown how a change in the variable *T* at time *t*, has a positive effect on *APRE* which is consistent in the four campaigns. This effect begins to show in *APRE* two hours later, $APRE_{t+1}$ [r(1) statistically significant in the four campaigns] and finishes four hours later, $APRE_{t+2}$ [r(2) statistically significant in campaign 5], being positive in campaigns 0 and 5. The statistically significant and negative coefficient r(0) in campaigns 0, 2 and 5 is reflecting the side-effect of manual temperature adjustments performed by operators trying to control the melt index, and the slight correlation between temperature and ethylene flow.

Figure 9 (here)

Figure 10 displays the CCF between $\nabla APRE$ and ∇E showing the dynamic relationships between *APRE* and *E*. The effect of changing *E* at time *t* has the biggest impact in *APRE* two hours later, $APRE_{t+1}$ (r(1) statistically significant in the four campaigns). In this case the instantaneous relationship estimated by r(0) is positive and statistically significant in campaigns 4 and 5, and with the same sign as r(1). This is a consequence of the way *APRE* is worked out: productivity is directly related to ethylene flow. Note that in this process ethylene flow measurements were not used for process control.

Figure 10 (here)

The overview of the polymerization process and the consistent CCF´s suggests the following tentative transfer function model for the viscosity ∇MI and productivity $\nabla APRE$ variation at time *t*:

$$\begin{bmatrix} \nabla MI_t \\ \nabla APRE_t \end{bmatrix} = \begin{bmatrix} (\omega_{11,1} + \omega_{11,2}B)B & (\omega_{12,1} + \omega_{12,2}B)B \\ (\omega_{21,1} + \omega_{21,2}B)B & \omega_{22,1}B \end{bmatrix} \begin{bmatrix} \nabla T_t \\ \nabla E_t \end{bmatrix} + \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad (8)$$

where ∇T_t and ∇E_t are the adjustments of temperature and ethylene flow at the time *t* respectively, and a_l are white noise processes $N(0, \sigma_l^2)$. The model for ∇MI_t is a discrete transfer function model of second order for the two inputs. The model for $\nabla APRE_t$ is a discrete transfer function model of second order for the T_t and first order for E_t.

3.2.5. Model Estimation

The parameters of model (8) have been estimated with the statistical package SCA [15] in an iterative fashion. First, the model has been preliminary estimated assuming a white noise model for the noise terms (i.e. n_1=n_2=1). Then, by inspecting the

autocorrelation and partial autocorrelation functions of the residuals, new ARMA models for n_1 and n_2 are proposed, and the model is fitted again. This process is iterated until residuals do not show any autocorrelation structure. As Figure 11 illustrates, in this case the study suggests a white noise structure for ∇MI_t (equation (9)) and an $AR(1)$ structure for $\nabla APRE_t$.(equation (10)). Tables 1 and 2 give the results of the estimation procedure for the four campaigns studied for the ∇MI_t and $\nabla APRE_t$ models, respectively (standard errors of the estimations in parentheses).

∇MI_t Model:

$$\nabla MI_t = (\omega_{11,1} + \omega_{11,2}B)\nabla T_{t-1} + (\omega_{12,1} + \omega_{12,2}B)\nabla E_{t-1} + a_{1t} \quad (9)$$

Table 1 (here)

$\nabla APRE_t$ Model

$$\nabla APRE_t = (\omega_{21,1} + \omega_{21,2}B)\nabla T_{t-1} + \omega_{22,1}\nabla E_{t-1} + \frac{1}{(1-\phi B)}a_{2t} \quad (10)$$

Table 2 (here)

Estimates vary somewhat over the different campaigns. Models (9) and (10), however, give a reasonable description of the process over a long period of time, and residual analyses from them, gave no reason to consider another models to produce a substantial increase in explanatory power.

4. PLS Time Series Methodology: FIR Model

The use of Box-Jenkins methodology in the estimation of parsimonious transfer functions becomes a very complicated task when many highly correlated input variables are involved. An interesting approach to avoid those difficulties could be the use of methodologies based on the estimation of the finite impulse response (FIR) models. Among all the available methods for estimating the FIR model, we have used the Partial Least Square Regression (PLS) model [16, 17, 18]. PLS is a regression-like method highly recommended in situations with high co-linearity among the input variables or even with rank-deficient data. This technique is very efficient in handling missing data, providing a wide package of easy-to-use graphical tools which facilitates the outlier detection and the identification of the transfer function model. Ferrer *et al.* [9] illustrate the versatility of PLS through several real industrial cases including a brief description of the role of PLS methods to assist in the empirical model building of the continuous polymerization process thoroughly described in the present paper.

PLS is a projection method that models the relationship between a response matrix ***Y*** and a predictor matrix ***X***. PLS projects the data from the original space (***X, Y***) of high dimensionality, into a new subspace of lower dimension A where a new set of variables known as *latent variables* ($t_1, t_2,..., t_A$) and ($u_1, u_2,..., u_A$) are defined. These

new variables comprise all the relevant and meaningful aspects of the predictive variability contained in the original set. The dimensionality is reduced in both matrices (***X*** and ***Y***) according to the objective of searching the directions in the ***X*** and ***Y*** spaces which simultaneously explain more variability and are more useful in the prediction of the quality variables (***Y***). This is done by maximizing the covariance between each pair of latent variables, $cov(t_a; u_a)$. Both matrices are decomposed into smaller ones as follows:

$$\boldsymbol{X} = \sum_{a=1}^{A} \boldsymbol{t}_a \boldsymbol{p}_a^T + \boldsymbol{E} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}$$

$$\boldsymbol{Y} = \sum_{a=1}^{A} \boldsymbol{u}_a \boldsymbol{c}_a^T + \boldsymbol{F} = \boldsymbol{U}\boldsymbol{C}^T + \boldsymbol{F}$$

where ***T*** and ***U*** are the score matrices, ***P*** and ***C*** are the loading matrices, and ***E*** and ***F*** are the residual matrices for ***X*** and ***Y***, respectively, for a model with A latent variables. The x-scores $\boldsymbol{t}_a$ are linear combinations of the **X** matrix (in the first PLS latent variable) or **X**-residual matrix ($\boldsymbol{X}_a$) (in the a-th latent variable, a>1):

$$\boldsymbol{t}_a = \boldsymbol{X}_{a\text{-}1}\boldsymbol{w}_a \quad ; \quad \boldsymbol{X}_a = \boldsymbol{X}_{a\text{-}1} - \boldsymbol{t}_a \boldsymbol{p}_a^T$$

being $\boldsymbol{w}_a$ the weight vector for the a-th latent variable.

This is done in a way to maximize the covariance between ***T*** and ***U***, both related by the inner relation $\boldsymbol{U} = \boldsymbol{T}\boldsymbol{B} + \boldsymbol{H}$, where ***B*** is a diagonal matrix and ***H*** is a residual matrix. This allows PLS to be expressed as a regression-like model (i.e. a function of the original x-variables):

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{B}\boldsymbol{C}^T + \boldsymbol{F}^* = \boldsymbol{X}\boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}\boldsymbol{B}\boldsymbol{C}^T + \boldsymbol{F}^* = \boldsymbol{X}\boldsymbol{B}_{\mathbf{coef}} + \boldsymbol{F}^* \qquad (11)$$

where $\boldsymbol{B}_{\mathbf{coef}}$ is a regression-like coefficients matrix and ***F**** is a residual matrix.

For a more detailed explanation of the different algorithms used and the mathematical and statistical structure of PLS see e.g. Helland [17] and Höskuldsson [18].

In this paper the PLS models will be frequently expressed as a predictive regression model (11). This will facilitate the comparison between these models and the ones obtained with the Box-Jenkins methodology. But it should be noted that the use of the PLS model expressed as weights and scores in the different components in addition to all the graphical tools such as contribution, loading and score plots, helps us to understand the covariance structure existing among the different variables under study. That covariance structure remains hidden when we use the model expressed in the classical regression form.

PLS is able to model the dynamic behavior in the relationship between the input and the output variables, by using a dynamic version known as PLS Time Series (PLS-TS). The inclusion of the dynamics is accomplished by including lagged variables in the model. The use of lagged variables was first suggested by Wold *et al*. [19]. This

methodology has been employed in this paper according to the proposals of Eriksson *et al*. [20], and Dayal and MacGregor [8].

We have selected a finite impulse response (FIR) model, where the output variable y_t is expressed as a linear combination of the original input variables x_t and a new set of lagged variables (sequential lags of the original input variables x_t). In the case of two inputs (as in the present case study) FIR model can be expressed as:

$$y_t = \sum_{i=1}^{2} \left(\beta_{i0} x_{i,t} + \beta_{i1} x_{i,t-1} + \beta_{i2} x_{i,t-2} + ... \right) + \varepsilon_t \qquad (12)$$

As commented in Section 3.2.3 no meaningful cross-correlation between both output variables $\boldsymbol{y}_t = (y_{t1;}\ y_{t2})'$ was detected (see Figure 5). Therefore, two separated transfer function models were estimated (one for each output variable). The number of lags to consider in the models was determined by the results of an initial study of the dynamic behavior in the data (autocorrelation functions, and the cross-correlation function among input and output variables shown in Figures 7, 8, 9, 10). The number of lags to include in the model was selected following conservative criteria, to capture the entire dynamics in the data.

4.1 FIR Model Formulation (PLS-TS Model)

4.1.1 Initial study of the data

In order to visualize the process evolution we have fitted two PLS models, one for each output variable, including all the process variables lagged from *t*-1 to *t*-6

$$MI_t = f(T_{t-1}T_{t-2}T_{t-3}T_{t-4}T_{t-5}T_{t-6}E_{t-1}E_{t-2}E_{t-3}E_{t-4}E_{t-5}E_{t-6})$$
$$APRE_t = f(T_{t-1}T_{t-2}T_{t-3}T_{t-4}T_{t-5}T_{t-6}E_{t-1}E_{t-2}E_{t-3}E_{t-4}E_{t-5}E_{t-6})$$

Figure 12 (here)

The *score* plots can serve to detect anomalies in the process evolution**.** Figure 12(top) shows a *score* plot for the first and third components of the *APRE* PLS model with campaign 2 data. The trajectory of the consecutive observations in the plot reveals the process evolution inside the control ellipse which defines the 95% confidence limit of the joint distribution of the scores. When an important change is detected, such as the shift between observations 23 and 24, responsible variables can be easily identified by using contribution plots. Figure 12 (middle) shows the score contribution plot for the difference between observations 23 and 24 signaling the ethylene flow as responsible variable. This can be confirmed by looking at the database where it can be appreciated that the level of ethylene flowing into the polymerization reactor decreases sharply (Figure 12 (bottom)). Consequently the contribution plot may become an important diagnostic tool in this phase.

A PCA (Principal Component Analysis) model would be an alternative to the PLS when the objective is to study the process evolution and the overview of the data

set. Figure 13 shows a *score* plot of the PCA model built from the output *APRE* and the same input variables used in the previous PLS model for campaign 2. In the plot it can be seen how the process progresses through time. Additionally, no clusters or grouped observations, which may invalidate the use of a single PLS model of the data set for the whole campaign, can be appreciated. Moreover, the use of these multivariate methodologies could help to determine the need for some kind of intervention analysis in specific time points of the campaigns, so that they could become useful complementary tools in the application of the Box-Jenkins methodology.

Figure 13 (here)

When outliers are present in our data, PLS time series methodology allows coping with them in a straightforward way. As PLS is based on least squares regression methodology, there is no need to adjust the data to the previous or following observations or to use any kind of intervention analysis as it is needed in the Box-Jenkins methodology. In the pre-treatment of the data each observation is expanded to include the dynamic behavior of the process. Thus, a single outlier observation spreads its effect into the neighboring observations, increasing its chances to be considered as an outlier during the analysis of the data set. Anyway, the latter does not imply that the information of these additional excluded normal observations is completely lost because it is included in the lagged variables of other normal expanded observations.

For instance, the expansion for an observation t with m input variables and k lags depends on observations t-1, t-2,…,t-k:

Observation t Expanded observation t

$$(y_t, x_{1,t}, x_{2,t} \ldots\ldots x_{m,t}) \Rightarrow (y_t, x_{1,t}, x_{1,t-1}, \ldots\ldots x_{1,t-k}, x_{2,t}, x_{2,t-1}, \ldots x_{2,t-k} \ldots\ldots x_{m,t}, x_{m,t-1}, \ldots x_{m,t-k})$$

where:

Observation t-1

$$(y_{t-1}, x_{1,t-1}, x_{2,t-1} \ldots\ldots x_{m,t-1})$$

……..

Observation t-k

$$(y_{t-k}, x_{1,t-k}, x_{2,t-k} \ldots\ldots x_{m,t-k})$$

It is also important to check for the stationarity of the series. As commented in Section 3.2.1 if the data series are not stationary, the cross-correlation functions (CCF´s) cannot be estimated. It was also shown (see Figures 3 and 4) that a first difference in the variables was sufficient to achieve stationarity. The CCF's are very useful in the PLS-TS methodology in order to determine the number of lags to be considered in the initial tentative model. These functions also permit the detection of cyclic patterns and control loops in the process.

.

4.1.2 Data Pre-treatment

As it was previously commented the PLS-TS is a variation of PLS which takes into account the dynamic behavior of the series by including lagged input variables in the model. In this case the pre-treatment of the data consists of three steps: matrix

expansion (according to Figure 14), differentiation and classical pre-treatment methods in PLS models:

a) Matrix expansion: the matrix of the input variables T_{t-1} and E_{t-1} is expanded with 5 new lagged variables for each original input variable from *t*-2 to *t*-6. So, after expanding the input matrix there will be 12 input variables. This lagging process is carried out as shown in Figure 14, leading to the loss of 5 observations.

Figure 14 (here)

b) Differencing: the variables are differenced in order to make them stationary. This will make possible the comparison of results between the Box-Jenkins and PLS-TS methodology.

c) Classical pre-treatment in PLS model estimation: centering and unit variance scaling will be applied to the expanded $\boldsymbol{X}$ matrix and to the response variable. Centering variables (to relocate the origin of coordinates to the centre of gravity of the cloud of observations) will serve to facilitate the interpretation of the model. Since PLS is a regression method which is variance dependent, it is convenient to scale all the variables to unit variance. This way all input variables have the same opportunity to be expressed in the model.

4.1.3 Model Estimation

The PLS-TS model has been estimated with the SIMCA-P software. As the outputs are not correlated, we have estimated a separate model for each different output: ∇MI_t and $\nabla APRE_t$

$$\nabla MI_t = f(\nabla T_{t-1}\nabla T_{t-2}\nabla T_{t-3}\nabla T_{t-4}\nabla T_{t-5}\nabla T_{t-6}\nabla E_{t-1}\nabla E_{t-2}\nabla E_{t-3}\nabla E_{t-4}\nabla E_{t-5}\nabla E_{t-6})$$
$$\nabla APRE_t = f(\nabla T_{t-1}\nabla T_{t-2}\nabla T_{t-3}\nabla T_{t-4}\nabla T_{t-5}\nabla T_{t-6}\nabla E_{t-1}\nabla E_{t-2}\nabla E_{t-3}\nabla E_{t-4}\nabla E_{t-5}\nabla E_{t-6})$$

where ∇ is the differential operator.

Alternatively, a PLS model can be expressed by using the coefficients of a classical regression model. There are different expressions for the regression coefficients and in every phase of the study the most appropriate expression will be selected. The estimation starts with the selection of the most influential lags. For the lags selection the most appropriate is to use the β_{CS} regression coefficients of the model calculated on centered and scaled to unit variance variables according to the expression

$$\frac{\nabla APRE_t - m_{\nabla APRE_t}}{s_{\nabla APRE_t}} = \beta_1 \frac{\nabla E_{t-1} - m_{\nabla E_{t-1}}}{s_{\nabla E_{t-1}}} + \beta_2 \frac{\nabla T_{t-1} - m_{\nabla T_{t-1}}}{s_{\nabla T_{t-1}}} + \ldots \qquad (13)$$

Given that every variable is scaled by its standard deviation, the β_{CS} will serve to measure the importance of the variables in the model. According to the objective of

finding parsimonious models, it will be appropriate to select only the most influential lags for the model. The *PLS regression coefficients plot*, where the value of the β_{CS} corresponding to all the original and lagged variables is plotted in a single chart, becomes a very useful tool for this task. The candidates to be selected are the variables with major values of β_{CS} that, at the same time, have the same sign in the different campaigns under study. This plot, which corresponds to the impulse response function, could be used to facilitate the transfer function identification when used in combination with the Box-Jenkins methodology. Also, if the input variables were independent to each other, this plot would become similar to the cross-correlation function.

As an example, Figure 15 shows the *PLS regression coefficients plot* for ∇MI model in campaign 5. This figure is quite similar to the cross-correlation functions of ∇MI with ∇T and ∇E, jointly represented in Figure 16 (already shown in Figures 7 and 8). Therefore, the shape of the coefficients bars in the *PLS regression coefficients plot* could be used to identify the order of the transfer function.

Figure 15 (here)

Figure 16 (here)

The estimated β_{CS} and their 95% confidence intervals for the four campaigns and both PLS models (∇MI and $\nabla APRE$) are shown in Appendix 1 (Figures 17 and 18). The coefficients corresponding to the most influential lags are selected based on their statistical significance and their consistency in the different campaigns under study. In the ∇MI_t model the selected lags were ∇E_{t-1}, ∇T_{t-1} and ∇T_{t-2} because their β_{CS} coefficients are positive in the four campaigns under study, and ∇E_{t-2} because its β_{CS} coefficient is negative in campaigns 2 and 5, although in the campaigns 0 and 4 the sign is different but with small values. The selected lags in the $APRE_t$ model were ∇E_{t-1} and ∇T_{t-1} because their β_{CS} coefficients are positive in the four campaigns under study, and ∇T_{t-2} because its β_{CS} coefficient is positive in campaigns 0, 4 and 5, despite in the campaign 2 the sign is different but with a small value.

After the selection of the most influential lags for the models, both models can be pruned. This phase serves to refine the proposed model of the process and consequently obtain a more parsimonious one. With this purpose, the large quantity of highly correlated input variables considered at the beginning is substantially reduced in this step. In order to proceed with this stage we use the information contained in the different loading plots of the model that permit to identify the different clusters of highly correlated variables which provide the same information to the model. In our model, there was no need to proceed with this operation since these models have been created with a small number of input variables.

After pruning the model the next step is the final model estimation. In order to compare the estimated PLS models with the corresponding Box–Jenkins estimated models, the PLS models will be expressed in the form of a classical regression model (11). The chosen form for the regression coefficients is the ordinary coefficients β calculated according to the expressions (14) and (15). The study of the statistical significance of the ordinary coefficients is based on a jackknife procedure. Tables 3 and 4 provide the results of the estimation procedure for the four campaigns studied for the ∇MI_t and $\nabla APRE_t$ models, respectively. R^2_x and R^2_y indicate the fraction of the sum of

squares of all the inputs and output explained by the model, respectively. Q^2 is the fraction of the total variation of the output variable that can be predicted by the model.

∇MI_t model:

$$\nabla MI_t = \beta_1 \nabla T_{t-1} + \beta_2 \nabla T_{t-2} + \beta_3 \nabla E_{t-1} + \beta_4 \nabla E_{t-2} + a_{1t} \quad (14)$$

$\nabla APRE_t$ model:

$$\nabla APRE_t = \beta_1 \nabla T_{t-1} + \beta_2 \nabla T_{t-2} + \beta_3 \nabla E_{t-1} + a_{2t} \quad (15)$$

5. Comparison of the Estimated Models

The estimated models for ∇MI_t from Box-Jenkins and PLS-TS methodologies are described in Tables 1 and 3, respectively. Both models have the same structure and the estimated parameters present similarities and some discrepancies. The estimated coefficients for the variable ∇T_{t-1} are consistent (positive) in all campaigns for both methodologies, although this is not statistically significant in campaign 4 in both methodologies, and in campaign 5 in the Box-Jenkins model. A similar behavior applies to the coefficients associated to variable ∇T_{t-2} except for campaign 4, where this coefficient is somewhat smaller. The estimated coefficients for variable ∇E_{t-1} are consistent (positive) in all the estimated models, although this is not statistically significant in campaigns 0 and 2 in both methodologies, and in campaign 4 in the PLS-TS model. Finally, the estimated coefficients for variable ∇E_{t-2} are quite similar, almost always negative except for the campaign 0 in PLS-TS model, where it is not statistically significant. The estimated residual variance is similar in both methodologies. Based on this comparison we can state that the two methodologies reach at quite similar results following different approaches.

For both methodologies the model structure for $\nabla APRE_t$ is also similar, except for the noise model. In Box-Jenkins methodology, the latter is an autoregressive AR (1) model, while in the PLS-TS methodology, it is white noise. The estimated coefficients for the first time lags ∇T_{t-1} and ∇E_{t-1} are fairly similar, whereas for ∇T_{t-2} have in general opposite signs. This discrepancy can be explained by the different structure of the estimated noise model in both methodologies. The estimated residual variance is smaller in the Box-Jenkins model, except for the campaign 2 where it is greater than in the PLS-TS model.

6. Conclusions

The use of multivariate time series methodologies of Box-Jenkins and PLS to build empirical models using historical data for an industrial polymerization process that produces high density polyethylene have been investigated and compared. The basic ideas behind the two approaches have been presented and their advantages and limitations have been discussed.

The Box-Jenkins methodology yields more parsimonious models because it provides the possibility to express the transfer function as a quotient of polynomials

coefficient matrices and it allows to explicitly represent the noise as an ARIMA vector model, whereas in the PLS-TS a non-parsimonious finite impulse response (FIR) function is directly identified rendering frequently over-parameterized models. Although by means of PLS-TS the process estimation is more straightforward, if the model has a large number of coefficients, it will yield greater uncertainty and lesser accuracy in forecasting. In spite of the greater effort made in the parsimonious transfer function identification required by Box-Jenkins methodology, the latter is recommended when it is expected that the structure of the process is reasonably simple and the inertia of the process is not very large.

PLS-TS provides a graphical tool kit very useful for the descriptive study of the process, allowing to follow its evolution and to detect outliers and process anomalies in an easy graphical way. Outliers can easily be detected using the PLS-TS methodology. The Box-Jenkins methodology uses more complex methods of outlier identification, estimation and removal of their effects.

The PLS regression coefficients plot facilitates the transfer function identification; its utility is similar to the cross-correlation function (used in Box-Jenkins methodology) when the input variables are independent. Box-Jenkins methodology seems to be more sensitive to the co-linearity between input variables. Both methodologies are successful in determining the variables and their lags to be considered in the model, since the results obtained by both methodologies were consistent.

Regarding the gaining of insight in process both methodologies are comparable and can be used to improve our understanding of the process.

As a general conclusion, Box-Jenkins methodology is considered the best approach to identify dynamic transfer function model in MIMO processes with a reduced number of independent inputs and few outputs. Nevertheless, even in this context the PLS-TS methodology can be used as a complementary tool for process understanding and outlier detection. However, when dealing with a high number of input and output variables with co-linearity problems and complex transfer functions, PLS-TS is the rational choice, given the enormous complexity (an even the unfeasibility) that would suppose the use of the Box-Jenkins methodology.

ACKNOWLEDGEMENTS

This research was partially supported by the Spanish Government (MICINN) and the European Union (RDE funds) under grant DPI2008-06880-C03-03/DPI.

REFERENCES

1. Ferrer A. Control Estadístico de Procesos con Dinámica: Revisión del Estado del Arte y Perspectivas de Futuro. *Estadística Española* 2004; **46**: 19-47. (In Spanish).
2. Box GEP, Jenkins GM, Reinsel GC. *Time Series Analysis. Forecasting and Control* (4th edn). Prentice Hall: Hoboken, NJ, U.S.A., 2008.
3. Reinsel GC. *Elements of Multivariate Time Series Analyses* (2nd edn). Springer: New York, 1997.

4. Liu L-M. *Time Series Analysis and Forecasting* (2nd edn). Scientific Computing Associates: Illinois, 2006.
5. Wold S. Exponentially Weighted Moving Principal Components Analysis and Projections to Latent Structures. *Chemometrics and Intelligent Laboratory Systems* 1994; **23**: 149-161.
6. Wise BM, Ricker NL. Identification of Finite Impulse Response Models by Principal Components Regression: Frequency-Response Properties. *Process Control and Quality* 1992; **4**: 77-86.
7. Wise BM, Ricker NL. Identification of Finite Impulse Response Models with Continuum Regression. *Journal of Chemometrics* 1993; **7**: 1-14.
8. Dayal BS, MacGregor JF. Identification of Finite Impulse Response Models: Methods and Robustness Issues. *Industrial & Engineering Chemistry Research* 1996; **35**: 4078-4090.
9. Ferrer A, Aguado D, Vidal-Puig, S, Zarzo M. PLS: A versatile tool for industrial process improvement and optimization. *Applied Stochastic Models in Business and Industry* 2008; **24**:551-567. DOI:10.1002/asmb.
10. Hannan EJ. The Identification Problem for Multiple Equation System with Moving Average Errors. *Econometrica*,1971; **39**: 751-765.
11. Kohn R. Identification Results for ARMAX Structures. *Econometrica* 1979; **47**: 1295-1304.
12. Box GEP, Cox DR. An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B* 1964; **26** :211-252.
13. Chen C, Liu L-M.. Joint Estimation of Model Parameters and Outliers Effects in Time Series. *Journal of American Statistical Association* 1993a; **88**: 284-297.
14. Box GEP, MacGregor JF. The analysis of closed-loop dynamic stochastic system. *Technometrics* 1974; **16**: 391-398.
15. Liu L-M. *Forecasting and Time series Analysis using the SCA Statistical System,* vol 2. Scientific Computing Associates: Illinois, 1997.
16. Geladi P, Kowalski B. Partial Least Squares Regression: A tutorial. *Analytica Chimica Acta* 1986; **185**:1–17.
17. Helland IS. On the structure of partial least squares. *Communication in Statistics—Simulation and Computation* 1988; **17**(2):581–607.
18. Höskuldsson A. PLS Regression Methods. *Journal of Chemometrics* 1988; **2**: 211-228.
19. Wold S, Albano C, Dunn III WJ, Edlund U, Esbensen K, Geladi P, Hellberg S, Johansson E, Lindberg W, Sjöström M. Multivariate data analysis in chemistry. In *Chemometrics: Mathematics and Statistics in Chemistry*, Kowalski BR (ed). D. Reidel Publishing Company: Dordrecht, Holland, 1984; 17-95.
20. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. *Introduction to Multi- and Megavariate Data Analysis using Projection Methods (PCA & PLS)*. UMETRICS AB: Umea, Sweden, 2006.

Author's biography

Susana Barceló is Lecturer of Statistics in engineering and biotechnology, Ph.D. candidate at the Department of Applied Statistics, Operation Research and Quality, and member of the Multivariate Statistical Engineering Research Group of the Universidad Politécnica de Valencia, Spain. She holds a M.Sc. degree in Agricultural Engineering and a M.Sc. degree in Science and Food technology from this university. Her main research interest is currently in the field of Multivariate Statistical Process Control and Biotechnology. She is a member of the European Network for Business and Industrial Statistics (ENBIS) and the Spanish Statistical Association.

Santiago Vidal-Puig is Lecturer of Statistics, Ph.D. candidate at the Department of Applied Statistics, Operation Research and Quality, and member of the Multivariate Statistical Engineering Research Group of the Universidad Politécnica de Valencia (Spain). He holds a M.Sc. degree in engineering. His main research interest is currently in the field of Multivariate Statistical Process Control.

Alberto Ferrer is Professor of Statistics at the Department of Applied Statistics, Operation Research and Quality, and Head of the Multivariate Statistical Engineering Research Group of the Universidad Politécnica de Valencia (Spain). He holds a M.Sc. in Agricultural Engineering and a Ph.D. in Statistics. His research focuses on statistical techniques for quality and productivity improvement, especially those related to multivariate statistical projection methods. Prof. Ferrer is currently an associate editor of Technometrics, a member of the editorial board of Quality Engineering, member of the Council of the International Society for Business and Industrial Statistics (ISBIS), and a member of the European Network for Business and Industrial Statistics (ENBIS), Spanish Statistical Association and Spanish Chemometrics Network. He is also active as consultant on Industrial Statistics, Six Sigma and Process Analytical Technology (PAT).

CAMP	$\hat{\omega}_{11,1}$	$\hat{\omega}_{11,2}$	$\hat{\omega}_{12,1}$	$\hat{\omega}_{12,2}$	$\hat{\sigma}_1^2$	R_1^2
0	0.24 (0.08)	0.23 (0.08)	0.003 $(0.02)^{NS}$	-0.002 $(0.02)^{NS}$	0.004	54
2	0.14 (0.04)	0.20 (0.04)	0.008 $(0.01)^{NS}$	-0.05 (0.01)	0.005	36
4	0.08 $(0.05)^{NS}$	0.12 (0.05)	0.095 (0.02)	-0.008 $(0.02)^{NS}$	0.006	30
5	0.068 $(0.09)^{NS}$	0.30 (0.08)	0.11 (0.03)	-0.083 (0.03)	0.003	51

Table 1. Estimated parameters of the ∇MI_t model (equation (9)) for the four campaigns studied (standard errors of the estimations in parentheses). R^2: goodness of fit (%). NS: statistically non-significant, *p*-value>0.05.

CAMP	$\hat{\omega}_{21,1}$	$\hat{\omega}_{21,2}$	$\hat{\omega}_{22,1}$	$\hat{\phi}$	$\hat{\sigma}_2^2$	R_2^2
0	0.47 (0.13)	-0.24 $(0.13)^{NS}$	0.82 (0.04)	-0.63 (0.08)	0.013	98
2	0.53 (0.15)	-0.36 (0.14)	0.91 (0.05)	-0.46 (0.10)	0.099	95
4	0.10 (0.05)	-0.33 (0.04)	0.77 (0.05)	-0.33 (0.08)	0.05	98
5	0.80 (0.31)	-0.08 $(0.32)^{NS}$	0.67 (0.13)	-0.37 (0.12)	0.057	88

Table 2. Estimated parameters of the $\nabla APRE_t$ model (equation (10)) for the four campaigns studied (standard errors of the estimations in parentheses). R^2: goodness of fit (%). NS: statistically non-significant, *p*-value>0.05.

CAMP	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}_1^2$	R^2_x	R^2_y	Q^2
0	0.18	0.19	0.032^{NS}	0.013^{NS}	0.004	29.8	13.4	1.87
2	0.13	0.12	0.031^{NS}	-0.037^{NS}	0.006	29.5	36.7	17.2
4	0.02^{NS}	0.09	0.058^{NS}	-0.002^{NS}	0.007	64.1	20.8	12.8
5	0.18	0.27	0.086	-0.033^{NS}	0.003	35.9	38.7	31.9

Table 3. Estimated parameters of the ∇MI_t model (equation (14)) for the four campaigns studied. R^2_x: goodness of fit of the inputs. R^2_y: goodness of fit of the output. Q^2: goodness of prediction of the output. NS: statistically non-significant, *p*-value>0.05.

CAMP	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}_2^2$	R^2_x	R^2_y	Q^2
0	0.83	0.49	0.60	0.022	40.9	63.9	57.5
2	0.19^{NS}	$0,12^{NS}$	0.85	0.015	66.5	96.5	96
4	0.23^{NS}	0.38	0.61	0.095	37	40.2	28
5	0.61	0.46^{NS}	0.41	0.064	44.7	22.4	15.2

Table 4. Estimated parameters of the $\nabla VAPRE_t$ model (equation (15)) for the four campaigns studied. R^2_x: goodness of fit of the inputs. R^2_y: goodness of fit of the output. Q^2: goodness of prediction of the output. NS: statistically non-significant, *p*-value>0.05.

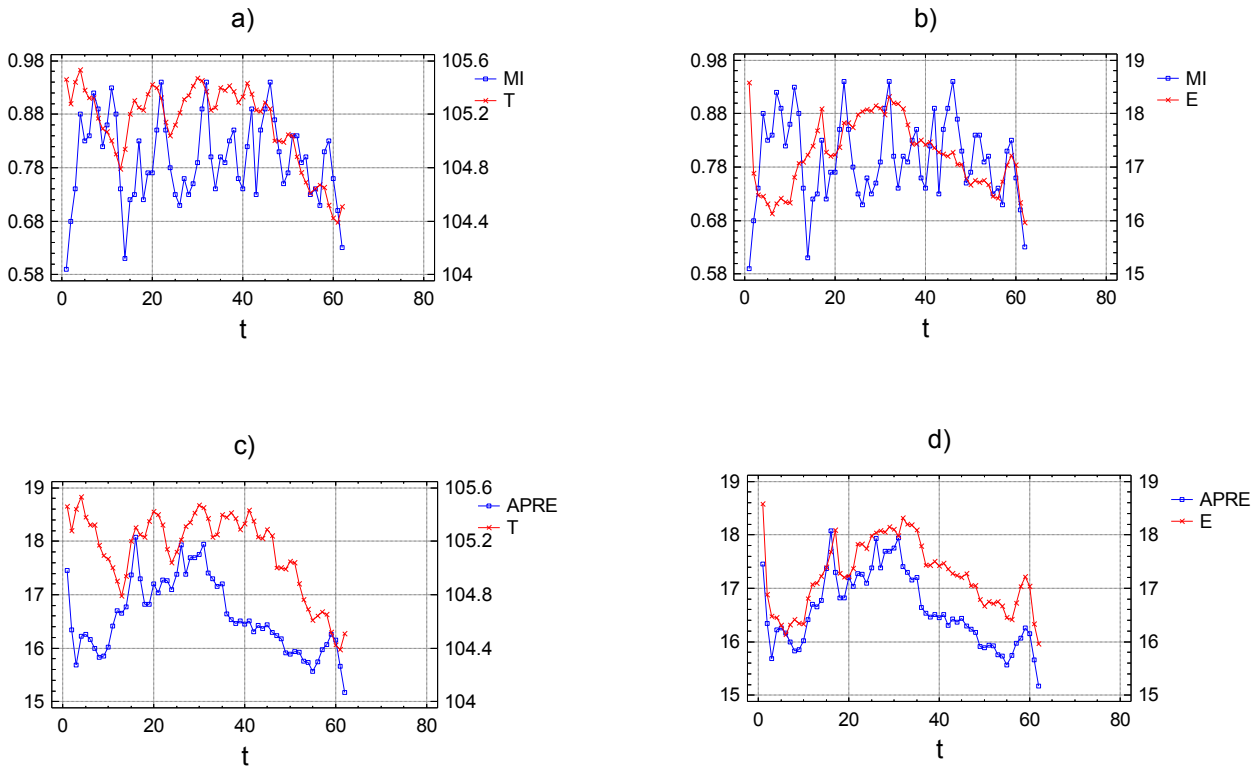


Figure 1. Example of input/output dataset. The characteristics of the outputs, polymer viscosity, measured by melt index (MI_t), and productivity, worked out by energy balance ($APRE_t$), are collected every two hours. The input variables are the averages of the reactor temperature (T_{t-1}) and ethylene flow (E_{t-1}) during the two hours before t. a) T_{t-1}, MI_t; b) E_{t-1}, MI_t; c) T_{t-1}, $APRE_t$; and d) E_{t-1} $APRE_t$.

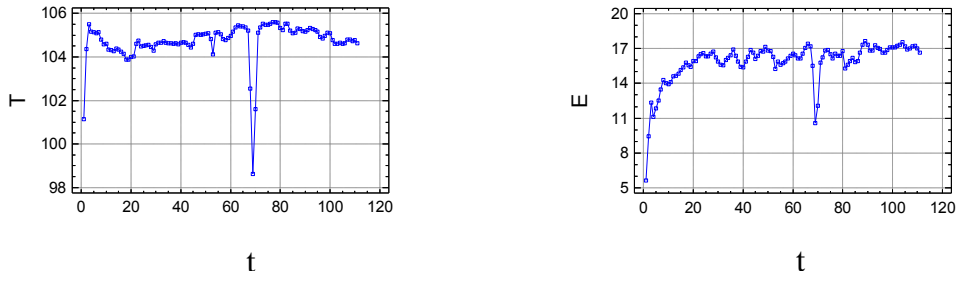


Figure 2. Time series plot for input variables T_t (left) and E_t (right) in Campaign 4.

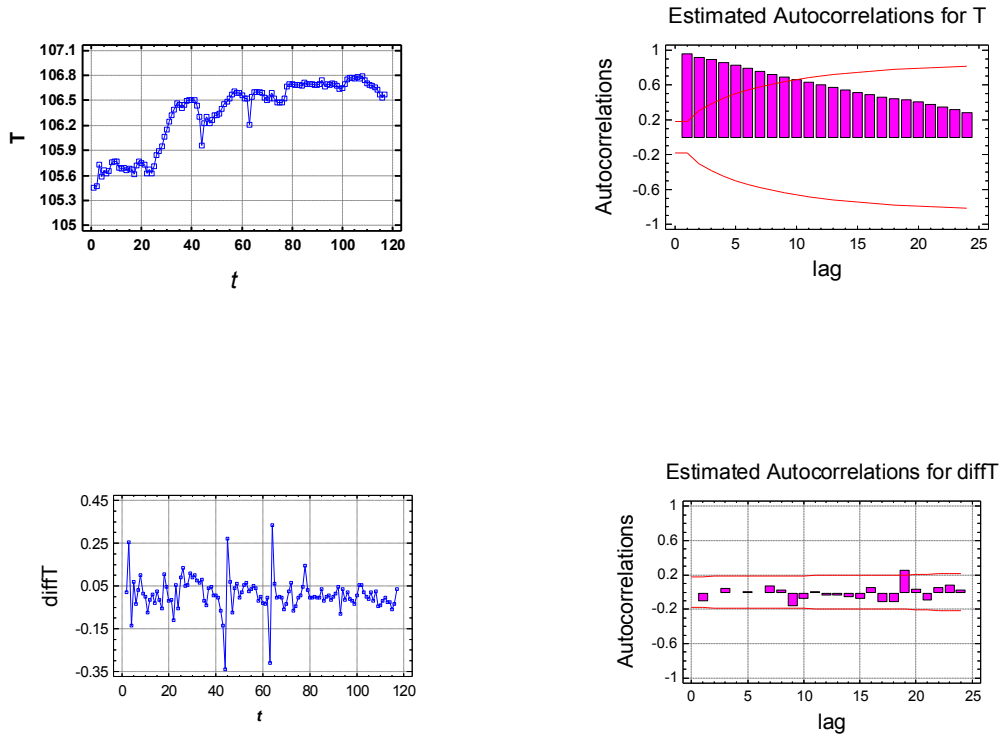


Figure 3. Time series plot (left) and estimated autocorrelation function (right) for T_t (top) and ∇T_t (bottom) in Campaign 0.

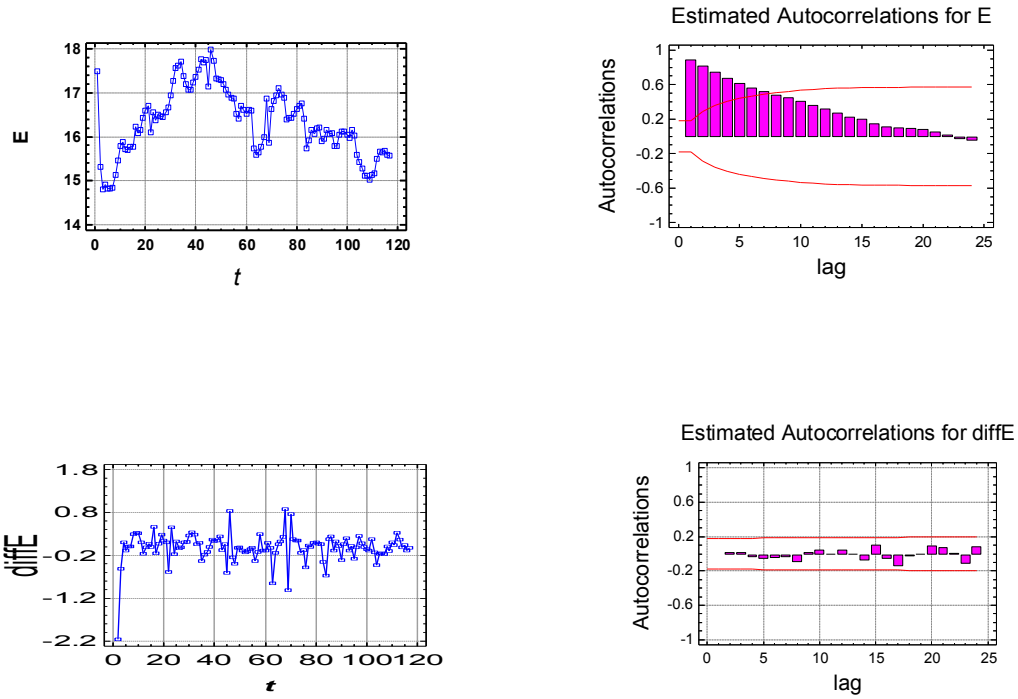


Figure 4. Time series plot (left) and estimated autocorrelation function (right) for E_t (top) and ∇E_t (bottom) in Campaign 0.

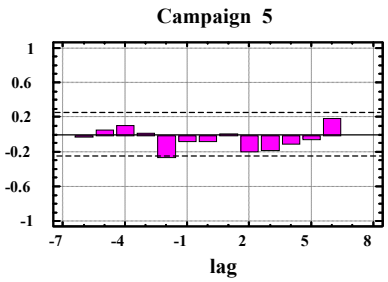


Figure 5. Estimated cross-correlation function between ∇VMI_t and $\nabla VAPRE_t$ in campaign 5.

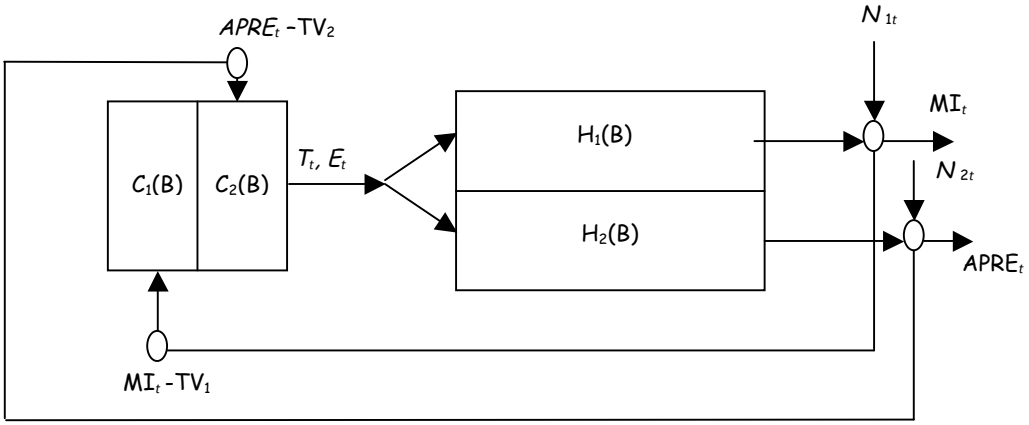


Figure 6. General closed-loop process scheme.

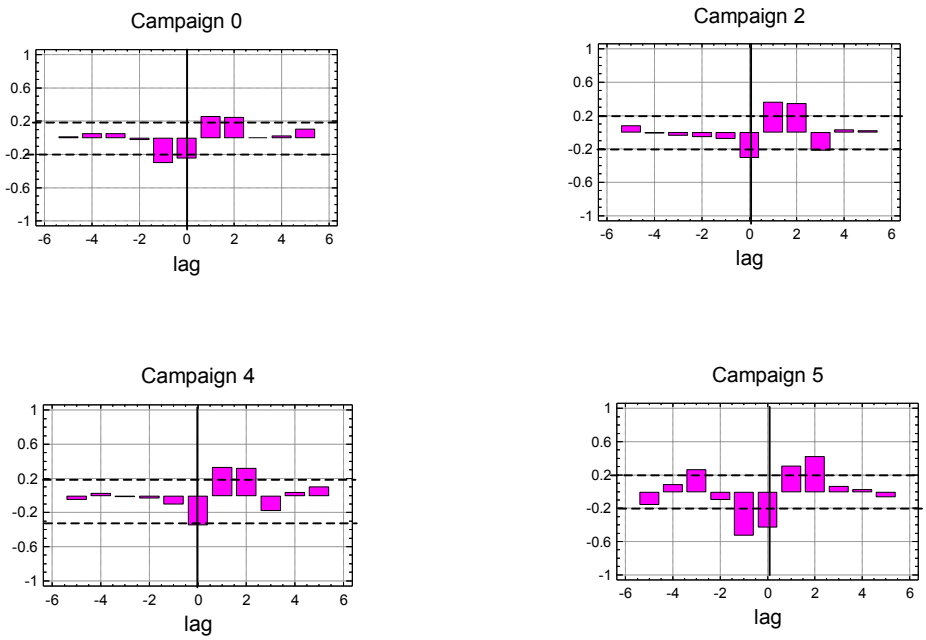


Figure 7. Estimated cross-correlation functions between *VMI* and *VT* for the different campaigns. Coefficients beyond limits (dashed lines) are statistically different from zero (*p*-values<0.05).

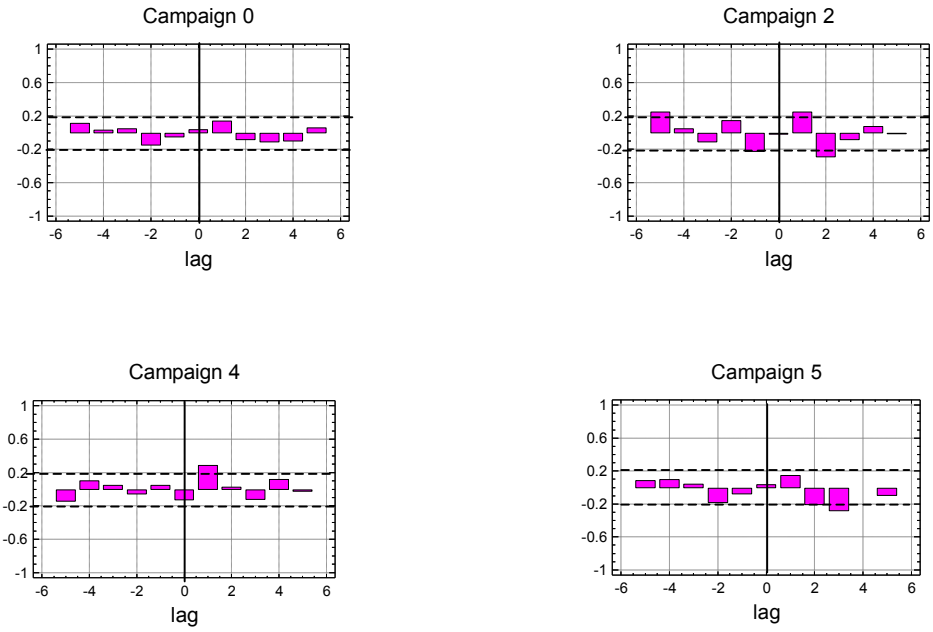


Figure 8. Estimated cross-correlation functions between ∇MI and ∇E for the different campaigns. Coefficients beyond limits (dashed lines) are statistically different from zero (p-values<0.05).

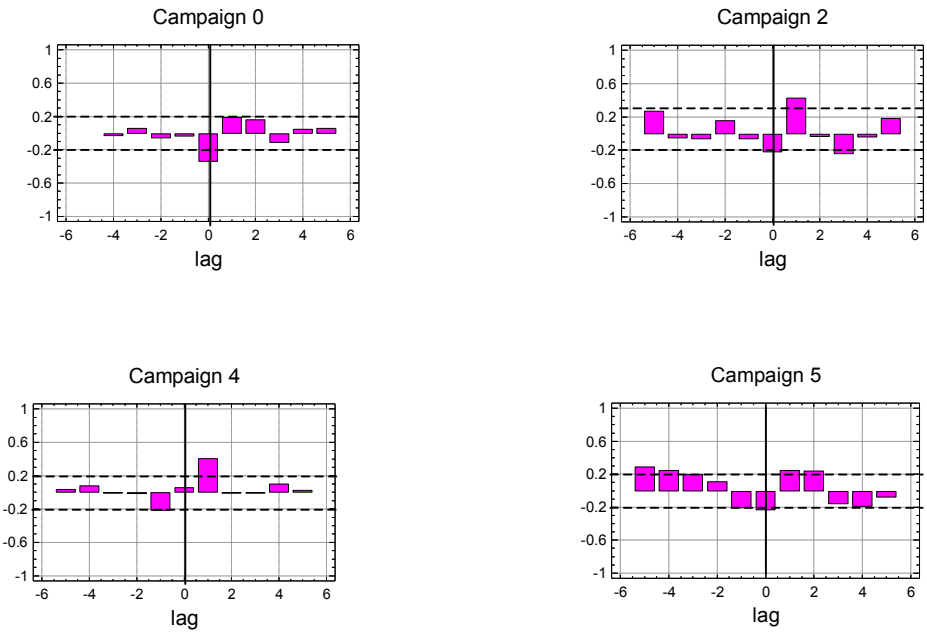


Figure 9. Estimated cross-correlation functions between *∇APRE* and *∇T* for the different campaigns. Coefficients beyond limits (dashed lines) are statistically different from zero (p-values<0.05).

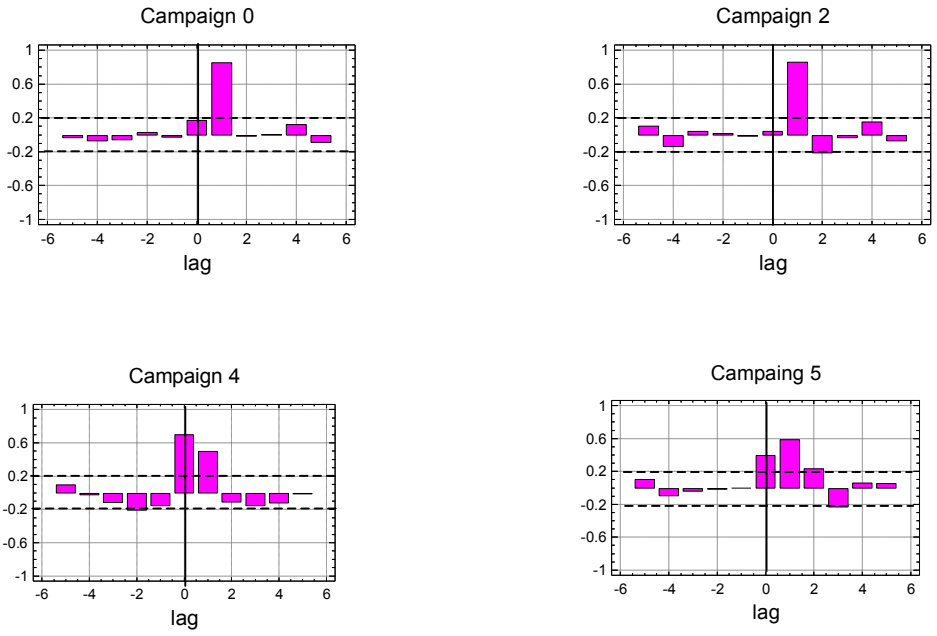


Figure 10. Estimated cross-correlation functions between $\nabla APRE$ and ∇E for the different campaigns. Coefficients beyond limits (dashed lines) are statistically different from zero (p-values<0.05).

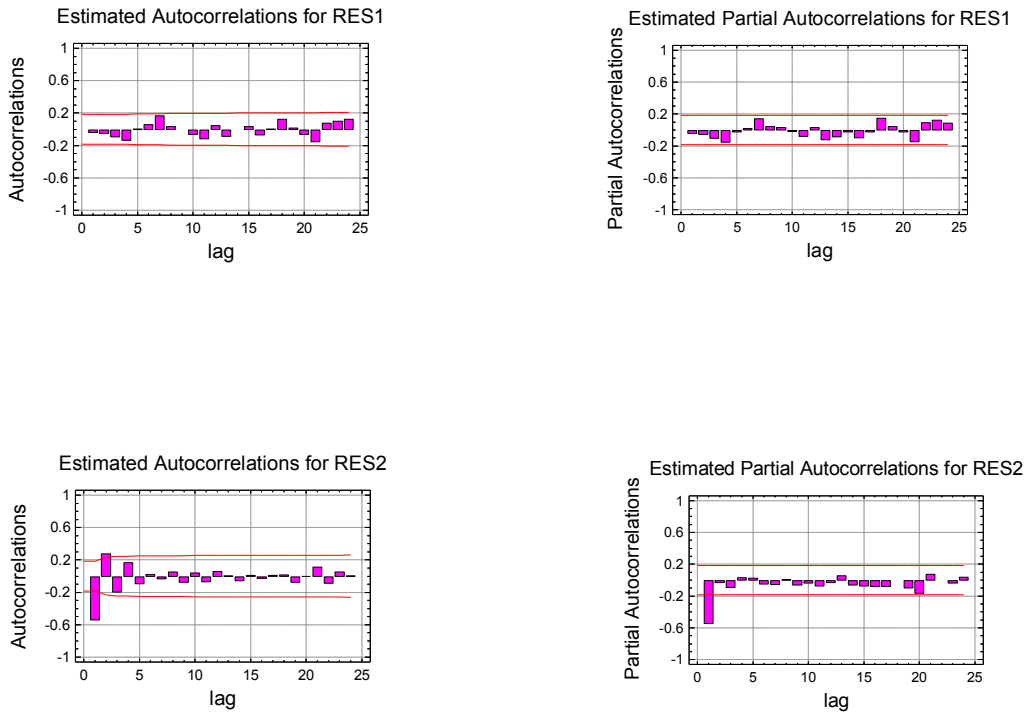
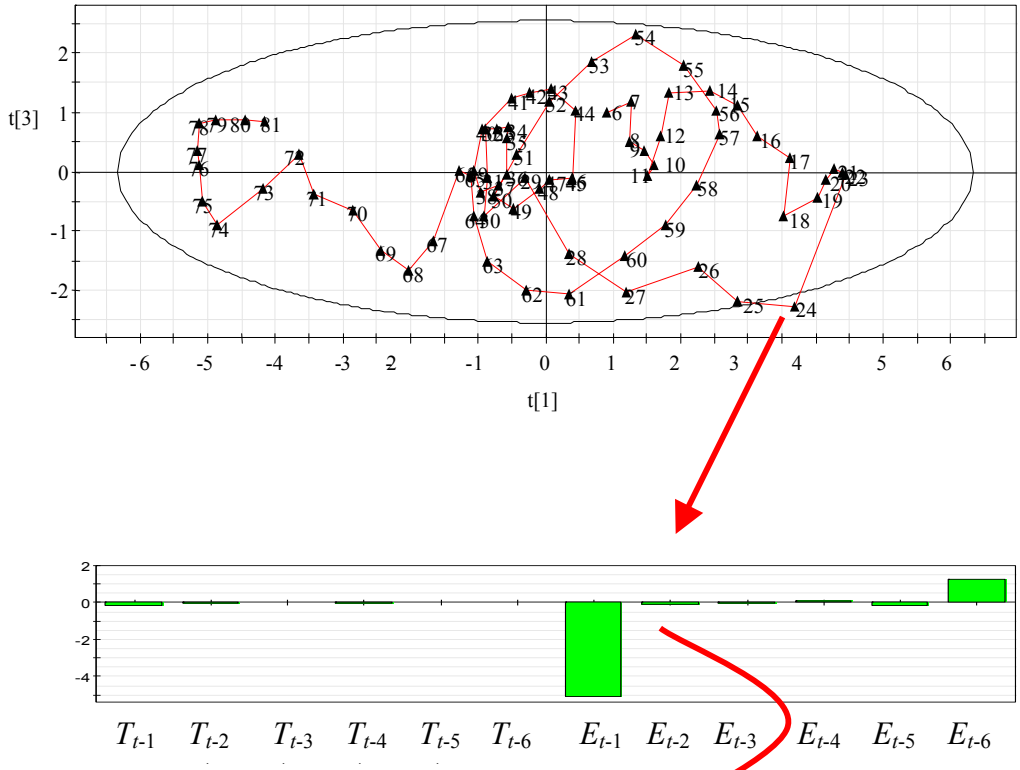


Figure 11. Estimated autocorrelation (left) and partial autocorrelation (right) functions of the residual series of the preliminary estimated dynamic regression model. Model ∇VMI (top) and model $\nabla VAPRE$ (bottom).



Observation	Ethylene flow
23	18,1045
24	14,9735
25	15,6845

Figure 12. *Score* plot for the first and third components of the *APRE* PLS model with campaign 2 data (top); score contribution plot for the difference between observations 24 and 23 (middle); data base detailed (bottom).

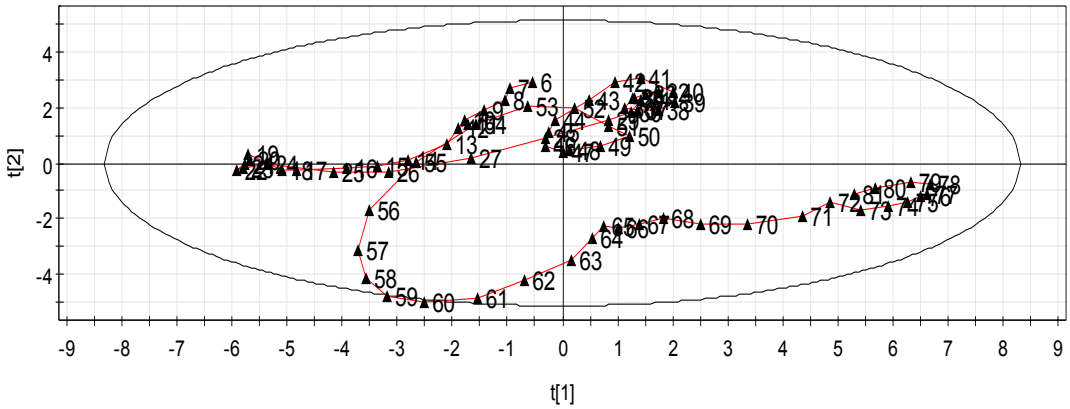


Figure 13. *Score plot* for the first and second components of the PCA model built from the output *APRE* and the same input variables used in the PLS model for campaign 2.

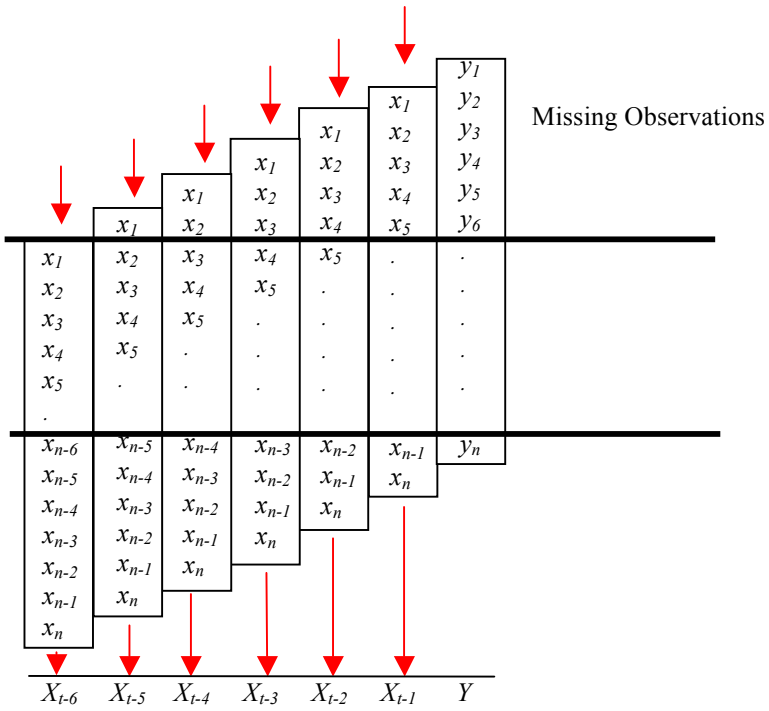


Figure 14. Lagging of the original input x_{it} variables

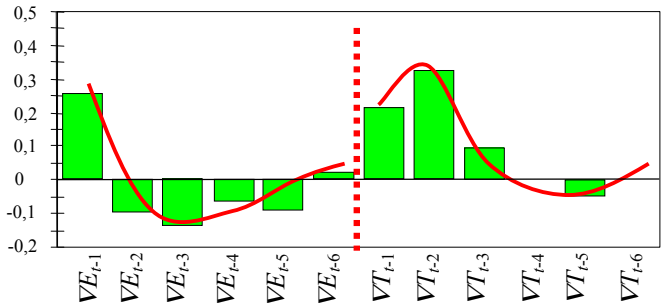


Figure 15. Estimated *PLS regression CoeffCS plot* for *VMI* model in campaign 5.

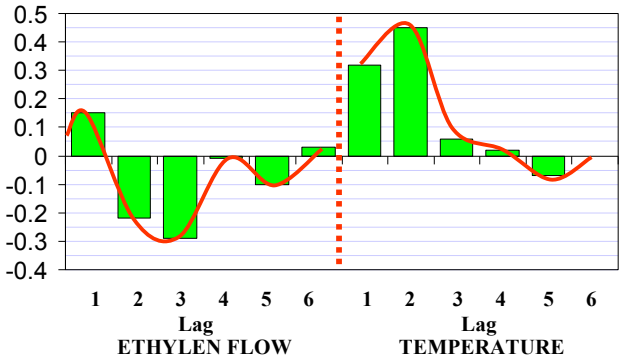


Figure 16. Estimated cross-correlation functions of *VMI* with *VE* and *VT* in campaign 5.

APPENDIX 1. *PLS regression CoeffCS plots.*

∇MI_t Model

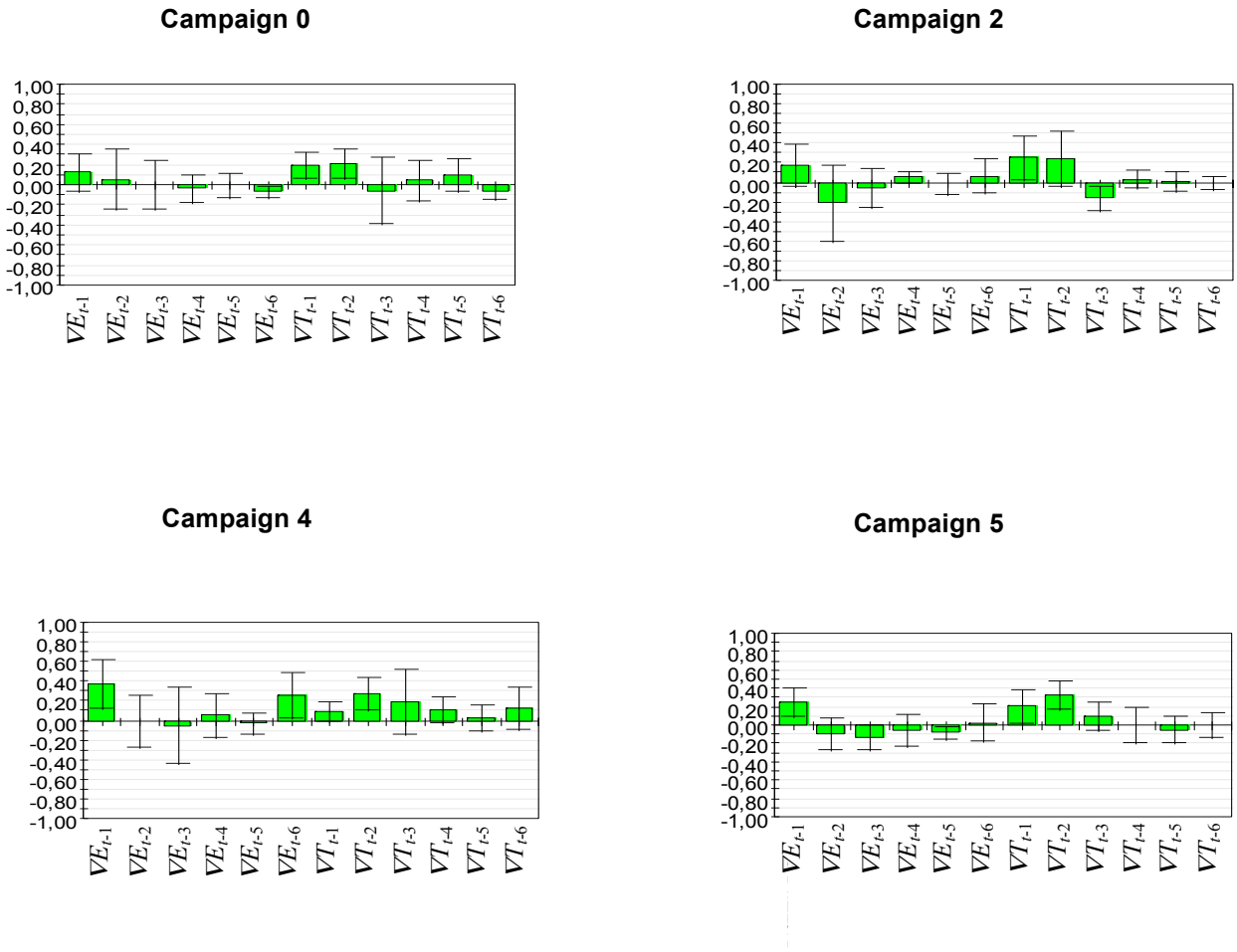


Figure 17. Estimated *β PLS regression coefficients (CoeffCs) plot* for the ∇MI_t model in the different campaigns (95% confidence intervals).

$\nabla APRE_t$ Model

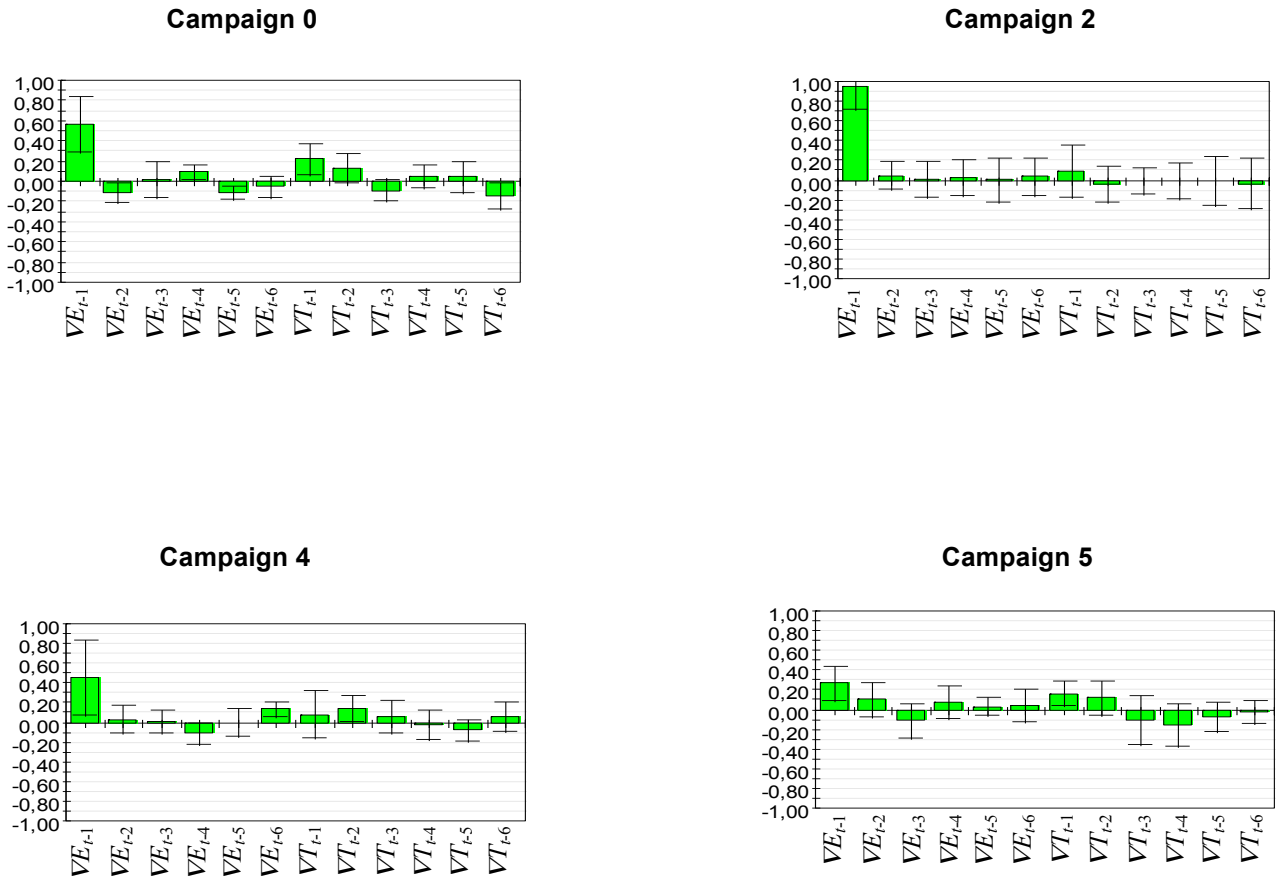


Figure 18. Estimated *β PLS regression coefficients (CoeffCs) plot* for the $\nabla APRE_t$ model in the different campaigns (95% confidence intervals).