IMPROVEMENTS IN TRANSFER ENTROPY CALCULATIONS FOR CAUSALITY DETECTION IN TIME SERIES

GUILHERME M GARCIA[1],CELSO MUNARO[1].

1. *LCI, Department of Electrical Engineering, Federal University of Espírito Santo*
E-mails: guifiles@gmail.com, munaro@ele.ufes.br

Abstract— The discovery of cause-effect relationships in signals from industrial processes is a challenging problem. A data-driven method to achieve this relation is the transfer entropy, a method based on the conditional probability density functions that measures directionality of variation. This method requires several parameters that must be properly chosen to avoid misleading results. In this work, the analysis of these parameters in the transfer entropy calculations is performed, and a methodology is proposed for their selection. The utility of the proposed approach is illustrated by several examples including the analysis of routine operating data in an industrial case study.

Keywords— Transfer entropy, fault diagnosis, time series prediction.

1 Introduction

Detection and diagnosis of plantwide abnormalities and disturbances are major problems in process industry. To isolate a fault in large-scale complex systems is particularly challenging because of the high degree of interconnections among different parts in the system. A simple failure may propagate along information and affect other parts of the system. To determine the root cause of certain abnormality, it is important to capture the process connectivity and find the connecting pathways.

Investigation of cause-effect relationships among variables, events or objects have been the fundamental questions of most natural and social sciences over the history of human knowledge (Shindler, 2007). The discovery of cause-effect relationships in signals from industrial processes is useful to confirm and discover known and unknown signal flow paths and subsequently use this information to find the root cause of process faults (Marques, 2013). Causality can be understood in terms of a flow among processes, expressed and analysed mathematically. Current statistics understands causal inference as one of its most important problems (Shindler, 2007).

Data-driven methods provide one of the ways to find the causal relationships between process variables. A few data-based methods are capable of detecting the causal relationships for linear processes. In the frequency domain, directed transfer functions and partial directed coherence are widely used in brain connectivity analysis. Other methods such as Granger causality, path analysis, and cross-correlation analysis with lag-adjusted variables are commonly used (Duan, et al., 2013). Granger causality analysis (Granger, 1969) is used mainly in the areas of econometrics and neuro-sciences. This technique has also been applied recently to find the root cause of plantwide oscillations from an industrial data. (Yuan et al., 2012) Methods based on Transfer entropy (Schreiber, 2000) have also been used for the same purpose. This method exploits conditional probabilities to determine cause and effect relationships in process data. Recently (Duan, et al., 2013) proposed the direct transfer entropy concept, an improvement to detect whether there is a direct information or not. This approach requires the choice of the size of the embedded vector and also the prediction horizon. If the prediction horizon is smaller then an existing time delay, no causality is detected. Since data is quantized for prediction, some care is also required to the number of bins.

This paper is organized as follows. In Section 2, an overview of transfer entropy method and the parameters choice is presented. Section 3 describes the implementation and shows some examples, followed by concluding remarks in Section 4.

2 Causality via transfer entropy

In Information Theory, for a system consisting of more than one component, important information on its structure can be obtained by measuring to which extent the individual components contribute to information production and at what rate they exchange information among each other. In Schreiber(2000) transfer entropy is proposed as the measure that shares some of the desired properties of mutual information but takes the dynamics of information transport into account. With minimal assumptions about the dynamics of the system and the nature of their coupling, one will be able to quantify the exchange of information between two systems, separately for both directions, and, if desired, conditional to common input signals.

Suppose two systems that generates events (x and y), the entropy rate that is defined as the amount of additional information required to represent the value of the next observation of one of the systems:

$$h_1 = -\sum_{x_{n+1},x_n,y_n} p(x_{n+1},x_n,y_n). \log_a p(x_{n+1}|x_n,y_n) \quad (1)$$

Suppose that the observation x_{n+1} was not dependent on the current observation y_n:

DOI: 10.5540/03.2013.001.01.0099

$$h_2 = -\sum_{x_{n+1},x_n,y_n} p(x_{n+1},x_n,y_n)\log_a p(x_{n+1}|x_n) \quad (2)$$

Now, the quantity h_1 represents the entropy rate for the two systems, and h_2 represents the entropy rate assuming that x_{n+1} is independent of y_n. Thus, the transfer entropy is

$$h_2 - h_1 = -\sum_{x_{n+1},x_n,y_n} p(x_{n+1},x_n,y_n)\log_a p(x_{n+1}|x_n) + \sum_{x_{n+1},x_n,y_n} p(x_{n+1},x_n,y_n)\log_a p(x_{n+1}|x_n,y_n) =$$

$$\sum_{x_{n+1},x_n,y_n} p(x_{n+1},x_n,y_n)\log_a \frac{p(x_{n+1}|x_n,y_n)}{p(x_{n+1}|x_n)} \quad (3)$$

There are actually two equations for the transfer entropy, because of its inherent asymmetry. The prediction horizon can be extended and become a parameter h

$$T_{y\rightarrow x} = \sum_{x_{n+h},x_n,y_n} p(x_{n+h},x_n,y_n)\log(\frac{p(x_{n+h}|x_n,y_n)}{p(x_{n+h}|x_n)}) \quad (4)$$

$$T_{x\rightarrow y} = \sum_{y_{n+h},y_n,x_n} p(y_{n+h},y_n,x_n)\log(\frac{p(y_{n+h}|y_n,x_n)}{p(y_{n+h}|y_n)}) \quad (5)$$

Substituting the joint probabilities (6) and (7)

$$p(x_{n+\mathrm{h}}|x_n,y_n) = \frac{p(x_{n+h},x_n,y_n)}{p(x_n,y_n)} \quad (6)$$

$$p(x_{n+\mathrm{h}}|x_n) = \frac{p(x_{n+h},x_n)}{p(x_n)} \quad (7)$$

the transfer entropy equations becomes

$$T_{y\rightarrow x} = \sum_{x_{n+h},x_n,y_n} p(x_{n+h},x_n,y_n). \log\left(\frac{p(x_{n+h},x_n,y_n).p(x_n)}{p(x_n,y_n).p(x_{n+h},x_n)}\right) \quad (8)$$

$$T_{x\rightarrow y} = \sum_{y_{n+h},y_n,x_n} p(y_{n+h},y_n,x_n). \log\left(\frac{p(y_{n+h},y_n,x_n).p(y_n)}{p(y_n,x_n).p(y_{n+h},y_n)}\right) \quad (9)$$

Joint PDFs for two stationary variables sequential in time are denoted by $p(x_{n+1},x_n)$ with the same PDF for x_n, x_{n+1}, because of stationarity, that is, $p(x_n) = p(x_{n+1})$, where 1 is the prediction horizon of x_n (one step ahead) and will be substituted by parameter h. The generalization of this joint PDF is the joint PDF for k + l variables giving $p(x_n,y_n)$, where $x_n = [x_n, x_{n-1}, \dots, x_{n-(k-1)}]$ and $y_n = [y_n, y_{n-1}, \dots, y_{n-(l-1)}]$ are embedded vectors. The parameters k and l are referred to as the embedding dimension of x and y, respectively (Bauer, et al., 2007).

A special case is when $k = 0$ and $l \neq 0$, so that the transition probability $p(x_{n+h}|y_n)$ measures the causal relationship between x and y in the sense that y can be identified as the cause or driver of x. So the transfer entropy becomes

$$T_{y\rightarrow x} = \sum_{x_{n+h},y_n} p(x_{n+h},y_n)\log(\frac{p(x_{n+h},y_n)}{p(x_{n+h}).p(y_n)}) \quad (10)$$

$$T_{x\rightarrow y} = \sum_{x_{n+h},y_n} p(y_{n+h},x_n)\log(\frac{p(y_{n+h},x_n)}{p(y_{n+h}).p(x_n)}) \quad (11)$$

Small values of transfer entropy suggest no causality or direction of influence while large values do. A threshold using a Monte Carlo method with surrogate data was proposed in (Bauer, et al., 2007). In this work, the same threshold is used and the values above are called significant values.

2.1 Kernel Estimation

Before calculating the causality value from transfer entropy, joint PDFs and transition probabilities have to be constructed from the time series. Estimation of the PDF from time series x and y is most commonly done with histograms, but due to the high order of the joint PDFs (k+l+1 dimensions), the number of samples required for the construction via histograms is extremely large. Therefore, the use of kernel estimation (Silverman 1986), was proposed in (Bauer, et al., 2007).

The Kernel method gives a more precise estimation of the PDF than histograms by considering the exact values of a time series x. A Kernel function K is centered at every sample point and summed to give an estimate $\hat{p}(x)$

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} K(x - x_i) \quad (12)$$

and considering a Gaussian Kernel function,

$$K(x - x_i) = \frac{1}{\sqrt{2\pi\theta}}\exp(-\frac{(x - x_i)^2}{2\theta^2}) \quad (13)$$

where θ is the estimator width which is adjusted to the number of samples N and the standard deviation of time series x (Silverman 1986).

For more than one dimension the Kernel can be extended to a vector valued data $x \in \Re^d$. This method as known as Fukunaga method. The estimate $\hat{p}(x)$ can be rewritten as

$$\hat{p}(x) = \frac{1}{n(2\pi)^{d/2}|\Sigma|^{1/2}}\sum_{i=1}^{n}\exp\left[-\frac{1}{2}(x - x_i)'\Sigma^{-1}(x - x_i)\right] \quad (14)$$

where Σ is the covariance matrix and d is the dimension of vector x. One can approximate the estimate $\hat{p}(x)$ by a product kernel

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{k=1}^{d} K_{\theta_k}(x^{(k)} - x_i^{(k)})\right] \qquad (15)$$

This approximation used by many researchers bring some issues: the parameter θ is calculated for each dimension (k+l+1) independently and when multiplied they loose their desirable effect, causing the sum of probabilities to be smaller than 1, impacting in the TE values. This problem becomes formidable with the dimension d. The Fukunaga method uses the covariance matrix to mitigate this problem, and will be used in this work.

The equations presented are continuous, however, for implementation in a computer the signals must be quantized in a discrete grid. The number of amplitude bins is denoted by q and can be set independently. The increase of q bring a better approximation, but in order to achieve a better result the number of bins should be limited, since with the increase of bins they tend to have probability zero, causing the sum of probabilities to be smaller than 1. A threshold will be proposed to select q and to limit this effect.

2.2 Selection of the parameters

The three parameters required for TE approach are: the prediction horizon h and the embedding dimensions k and l. Since these parameters greatly affect the calculation of the transfer entropy, a systematic method is needed to determine their values. In the seminal work Schreiber (2000), the values $k = l = h = 1$ were proposed. Unfortunately, this fixed choice is not suitable in some cases, as shown in the future examples. In (Bauer, et al., 2007) the values $k = 0, l = 2$ were proposed and h was selected as a function of the process dynamics. If the process dynamics is known, the parameter can be set accordingly. If a dead time is detected between two signals, the optimum value of h is equal to the dead time. However, if the process dynamics is unknown, small values such as $h < 4$ should give good results. More recently, (Duan, et al., 2013) used the same procedure to determine prediction horizon h, but different methods to select the parameters k and l. The embedding dimension k, is determined as the minimum nonnegative integer above which the change of entropy rate for the future signal given only his past (e. g. $H^c(y_{i+h}|y_i)$) decreases significantly. Based on the values of k and h, the embedding dimension l, is determined as the minimum positive integer above which the change of entropy rate decreases significantly. No considerations are made about signal quantizing.

In this work, we use the maximization of TE to choose all parameters. First, the parameter h is calculated with minimum effort, i.e., for $l = 1$ and $k = 0$, and the value of h that maximizes the TE will be chosen. The importance in selecting h initially comes from the need of its value to be higher to an existing dead time, very common in industrial processes. If the dead time exists and is known, the initial value of h is chosen with its value. Second, with the value of $h, k = 0$ and l is varied from 1 to 3, selecting the value that maximizes the TE. Finally, with parameters h and l chosen, embedding dimension k is varied from 0 to 2, to maximize the TE.

With the increment of dimension is very difficult to calculate the kernel. For a dimension higher than 3 is very difficult to achieve the goal that the sum of all probabilities over all bins is equal to 1.

3 Applications

In these examples, the Transfer Entropy is computed using routines implemented in Matlab. The calculated values are compared to those computed with fixed values $k = l = h = 1$ from (Schreiber, 2000) and $k = 0, l = 2$ with h varying from 1 to 4 from (Bauer, et al, 2007).

3.1 Example 1

This example was used in (Ding, 2008), where two models were analysed to show the possibility of detecting direct and indirect causality: y affects z directly and x indirectly in the first model (Fig.1a)) and affects both z and x directly in the second model (Fig.1b)).

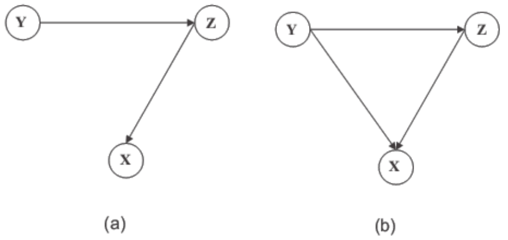

Figure 1. The two topologies for Example 1

First, the number of bins is determined in order to assure that the sum of probabilities be greater than 0.98. This threshold was determined according to the author's experience in the examples shown.

Table 1. Effect of number of bins over the sum of probabilities

Bins	$\sum p(x)$	$\sum p(y)$	$\sum p(z)$
10	0,9932	0,9858	0,9941
20	**0,9833**	**0,9819**	**0,9840**
30	0,9768	0,9797	0,9797
40	0,9723	0,9736	0,9721
50	0,9699	0,9680	0,9655

With the signals quantized, the next step is to choose the parameters h, l and k for each pair of signals. In Figure 3 the TE values are shown for parameter h varying from 1 to 5 for all pairwise

DOI: 10.5540/03.2013.001.01.0099

analysis. The notation TE$x \rightarrow z$ denotes the transfer entropy from x to z.

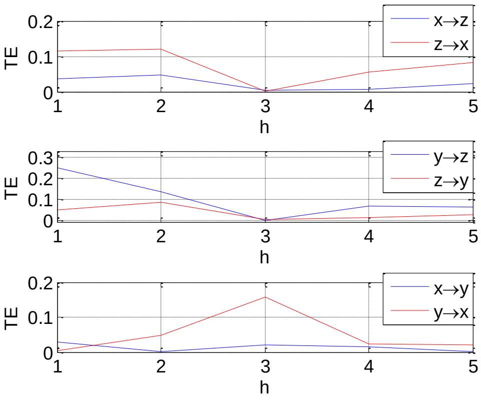

Figure 2 - TE values for *k=0*, *l=1* and *h* varying.

The maximum for TE$x \rightarrow z$, TE$z \rightarrow x$ and TE$z \rightarrow y$ is for $h = 2$; TE$y \rightarrow z$ is maximized for $h = 1$; for TE$_{x \rightarrow y}$ and TE$_{y \rightarrow x}$ the maximum value is for $h = 3$.

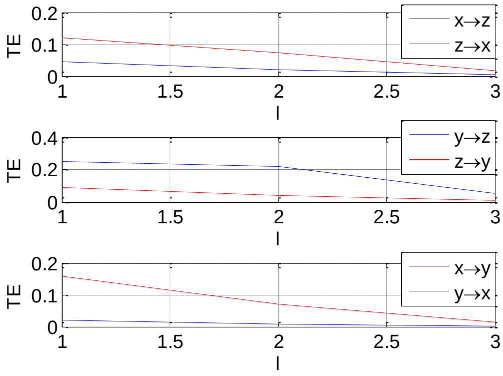

Figure 3 - TE values for *h* chosen, *k=0* and *l* varying.

With h chosen, the parameter l is selected using Figure 3: for all pairs $l = 1$ maximizes the TE values. With values for l and h selected, Figure 4 is then used to choose k: the maximum TE value is obtained for $k = 0$ for all pairs.

The result for the selected parameters is shown in Table 2, first line in all columns. The second line is for $q = 100$, in order to compare with $q = 20$ used for first line. The third and forth lines contain the values using parameters according to (Schreiber, 2000) and (Bauer, et al., 2007), respectively. The bold values are those which are significant.

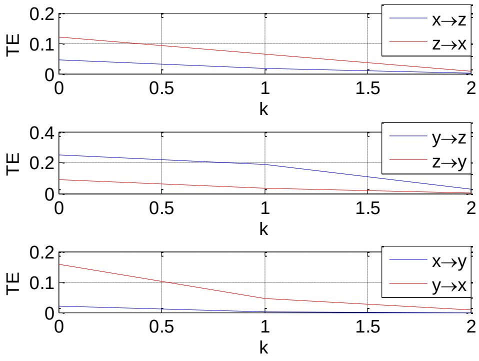

Figure 4 - TE values for *l* and *h* chosen and *k* varying.

Table 2. Results for first model and different choice of parameters and number of bins

TE$_{row\text{->}col}$	x	y	z

x	-	0.0288;
		0.0068;
		0.0021;
		0.0200
			0.0463;
			0.0118;
			0.0298;
			0.0238

y	**0.1587;**
	0.0431;
	0.0023;
	0.0743
		-	**0.2513;**
			0.0882;
			0.1899;
			0.2167

z	**0.1201;**
	0.0343;
	0.0713;
	0.0825
		0.0873;
		0.0231;
		0.0543;
		0.0420
			-

The comparison of values in first and second line in Table 2 shows that when the number of bins is increased from 20 to 100 the TE values drop significantly. The proposed choice of parameters allow achieving significant TE values for all causal relations previously known. The other choices only appoint the causal relation for TE$_{y \rightarrow z}$.

The same procedure was applied to the second model of example 1 from (Ding, 2008), with the topology shown in Figure 1.b: the results compared with the other two approaches are presented in Table 3. The maximum for TE$_{x \rightarrow z}$, TE$_{y \rightarrow z}$ is for $h = 1$; TE$_{z \rightarrow x}$ and TE$_{z \rightarrow y}$ is maximized for $h = 2$; for TE$_{x \rightarrow y}$ and TE$_{y \rightarrow x}$ the maximum value is for $h = 3$. With h chosen l and k are selected, the maximum for all pairs is obtained for $k = 0$ and $l = 1$.

Table 3. Results for second model and different choice of parameter

TE$_{row\text{->}col}$	x	y	z

x	-	0.0272;
		0.0018;
		0.0260
			0.0512;
			0.0432;
			0.0299

y	**0.2003;**
	0.0062;
	0.1236
		-	**0.2471;**
			0.1699;
			0.1963

z	**0.1686;**
	0.1174;
	0.1318
		0.0691;
		0.0529
		0.0368
			-

The parameters chosen allowed achieving significant values for all causal relations previously known. The parameters chosen according to (Schreiber, 2000) appointed only the causal relations TE$_{y \rightarrow z}$, TE$_{z \rightarrow x}$, while the parameters chosen according to (Bauer, et al., 2007) detected all causal relations, but with smaller TE values when compared to the proposed scheme. The value of TE$_{y \rightarrow x}$ is higher for second model, compared with first model with indirect causality from TE$_{y \rightarrow x}$. This difference is

DOI: 10.5540/03.2013.001.01.0099

helpful to detect direct or indirect pathways as proposed in (Duan, et al., 2013). Also, greater values of TE are valuable because they must be higher than a threshold to be significant.

3.2 Example 2

This example uses routine operating data from three control loops from a thermoelectric power plant (Figure 5). The level loop from a steam drum (LIC400) adjusts the set point for the water flow loop (FIC408). This flow comes from a reservoir (deaerator) whose level is given by LIC430.

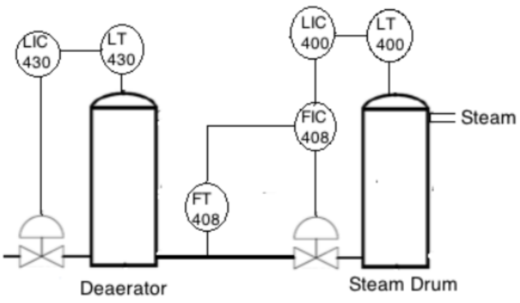

Figure 5 – Diagram of the thermoelectric power plant

LIC400 affects FIC408: when the level decreases, the flow is increased. Although the flow FIC408 comes from the tank with level LIC430, it is difficult to establish a relationship between them. This may happen because the level LIC430 depends on other variables and the controller tuning on this level loop is detuned or very slow, as should be the case. The signals (Figure 6) were differenced for use in the algorithms (in order to be covariance stationary).

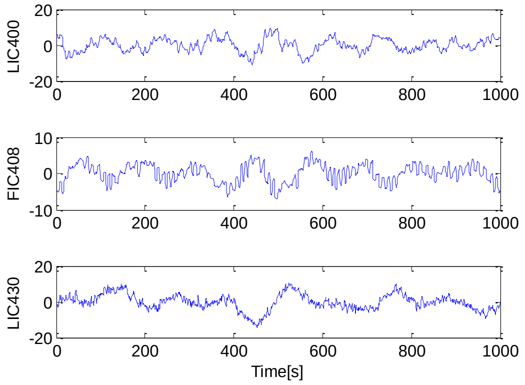

Figure 6 – Signals from the thermoelectric power plant

As in the other examples, the number of bins is determined using the same threshold of 0.98, with different values of q for the variables (Table 4).

Table 4. Effect of number of bins over the sum of probabilities

Bins	$\sum p(LIC400)$	$\sum p(\text{FIC408})$	$\sum p(\text{LIC430})$
q = 10	**0,9897**	0,9942	0,98576
q = 20	0,9778	0,9914	**0,98088**
q = 30	0,9679	0,9869	0,97588
q = 40	0,9591	0,9867	0,97308
q = 50	0,9614	0,9839	0,96189
q = 60	0,9585	0,9826	0,95474
q = 70	0,9475	**0,9807**	0,94899

The procedure to choose parameters k, l, h is shown in Figures 7, 8 and 9.

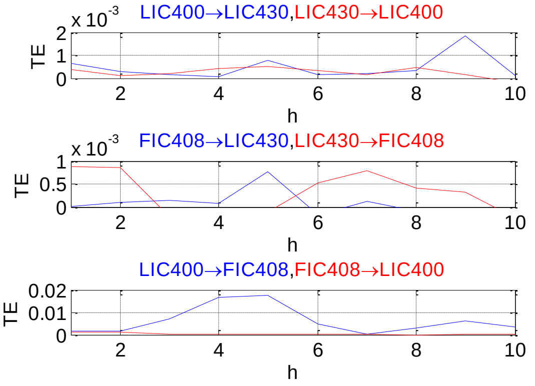

Figure 7 - TE values for *l=1*, *k=0* and *h* varying.

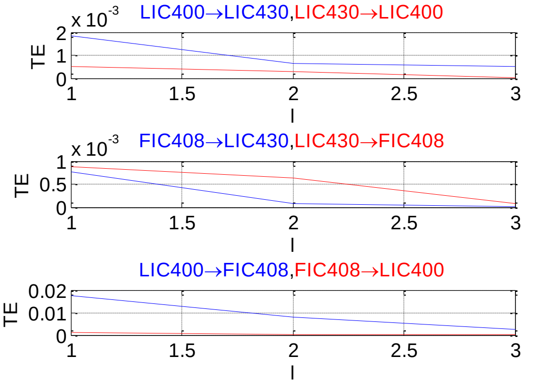

Figure 8 - TE values for *h* chosen, *k=0* and *l* varying.

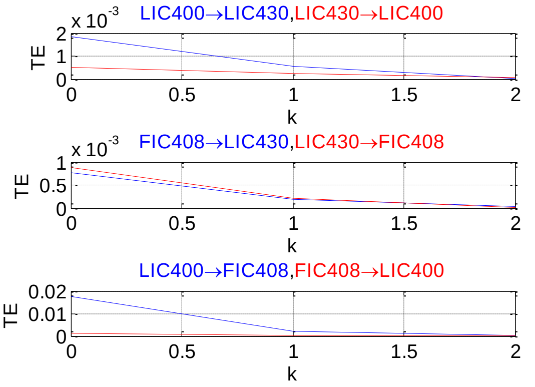

Figure 9 - TE values for *l* and *h* chosen and *k* varying.

The maximum for $\text{TE}_{\text{LIC400}\rightarrow\text{FIC408}}$, $\text{TE}_{\text{FIC408}\rightarrow\text{LIC430}}$ and $\text{TE}_{\text{LIC430}\rightarrow\text{LIC400}}$ is for $h = 5$; $\text{TE}_{\text{LIC400}\rightarrow\text{LIC430}}$ is maximized for $h = 9$; for $\text{TE}_{\text{LIC430}\rightarrow\text{FIC408}}$ and $\text{TE}_{\text{FIC408}\rightarrow\text{LIC400}}$ the maximum value is for $h = 9$. With h chosen embedded dimensions k and l are selected, the maximum for all pairs is obtained for $k = 0$ and $l = 1$.

The results are presented in Table 5. The relation between LIC400 and FIC408 is the only one with a

DOI: 10.5540/03.2013.001.01.0099

significant value. For the other approaches to chose the parameters, only the one according to (Bauer, et al., 2007) detect the causal relation. These examples confirm that the values for l and k are small, in general. Thus, $k = 0$ and $l = 1$ is a good initial guess for selecting h, with the additional benefit of a lower computational effort. If h is smaller then the dead time, the TE will be very small. If TE value increases and then decreases, the examples have shown this maximum to be a suitable choice. Otherwise, the search should continue until a given limit (example, maximum dead time in the signals, obtained from knowledge about the process under analysis).

Table 5. Results for example 2 and different choice of parameters

$TE_{row\text{->}col}$	LIC400	FIC408	LIC430

LIC400	-	**0.0174;**
		0.0000;
		0.0111
			0.0019;
			0.0003;
			0.0009

FIC408	0.0011;
	0.0004;
	0.0004
		-	0.0008;
			0.0000;
			0.0000

LIC430	0.0005;
	0.0002;
	0.0005
		0.0009;
		0.0002;
		0.0006
			-

To illustrate the computational effort, the time elapsed to calculate the $TE_{LIC400 \rightarrow FIC408}$ for $k = 0$ and l varying is shown in Table 6, using a core i3 with 3.07 Ghz and 4 GB of DDR3.

Table 6. Time elapsed to calculate the $TE_{LIC400 \rightarrow FIC408}$ for k = 0 and l varying

l	**Time(s)**
1	71.61
2	73.74
3	78.22
4	84.43
5	87.79

4 Conclusion

A proposal to choose the parameters to be used in the transfer entropy method, in order to find causal relations between variables was here presented. Several examples were used to illustrate the methodology and it was clear that higher values for the transfer entropy were obtained, compared to other methods. These higher values allow the use of more robust thresholds, which is important in order to avoid false positives results.

The proposed method, choosing initially the prediction horizon, brings an interesting result: the dimensions of the embedded vectors can be smaller. This choice reduces the kernel dimension, reducing the computational effort to maximize the value of the transfer entropy.

Acknowledgements

The authors are grateful to Arcelor-Mittal Tubarão for the data provided for Example 2.

References

C. W. J. Granger, “Investigating causal relations by econometric models and cross-spectral methods”, Econometrica, vol. 37, no. 3, pp. 424-438, August 1969. DOI: 10.2307/1912791

M. Bauer, J. W. Cox, M. H. Caveness, J. J. Downs, and N. F. Thornhill, “Finding the direction of disturbance propagation in a chemical process using transfer entropy”, IEEE Trans. Cont. Syst. Tech, vol. 15, no. 1, 2007.

P. Duan, F. Yang, T. Chen, and S. L. Shah, “Direct causality detection via the transfer entropy approach”, in press with IEEE Transactions on Control Systems Technology, 2012.

M. Ding, Y. Chen, and S. L. Bressler, “Granger causality: Basic theory and application to neuroscience," Elsevier Science, 2008.

B. W. Silverman, “Density Estimation for Statistics and Data Analysis”. New York: Chapman & Hall, 1986, pp. 34–48. DOI: 10.1007/978-1-4899-3324-9

T. Schreiber, “Measuring information transfer”, Phys. Rev. Lett., vol. 85, no. 2, pp. 461–464, 2000. DOI: 10.1103/PhysRevLett.85.461

V. M. Marques, C. J. Munaro, and S. L. Shah, “Data-based Causality Detection from a System Identification Perspective”, to appear in ECC 2013, Zurich-CH.

K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis." Physics Reports 441, no. 1 (2007): 1-46. DOI: 10.1016/j.physrep.2006.12.004

T. Yuan and S. J. Qin, “Root cause diagnosis of plant- wide oscillations using granger causality,”