PATH ANALYSIS FOR PROCESS TROUBLESHOOTING

Biao Huang [*,1] **Nina Thornhill** [**] **Sirish Shah** [*]
Dave Shook [***]

[*] *Dept. of Chemical and Materials Engg., University of Alberta, Edmonton, AB, Canada, T6G 2G6*
[**] *Department of Electronic and Electrical Engg., University of College London, London, U.K. WC1E 7JE*
[***] *Matrikon Inc., Edmonton, AB, Canada, T5J 3N4*

Abstract: In this paper, a model-free data-driven approach to process troubleshooting is proposed. The method is simple and can handle both univariate and multivariate processes. The only information needed for such an analysis is the data. The objective is to identify possible source of variability/oscillation from all interacting variables. To achieve this objective, a model-free method known as path analysis is used. In this paper, we will summarize the theory and algorithms developed for such an analysis. An industrial case study is presented to demonstrate the feasibility of the proposed method.

Keywords: Process monitoring, troubleshooting, data mining, path analysis

1. INTRODUCTION

If a control loop has no potential to improve performance by tuning the controller, then one obvious choice is to trace the source of the upset and reduce the disturbances/oscillations in the source. One therefore has to search for, among many loops which interact with the loop of concern, the source of the disturbances/oscillations. This can be a forbidding task for a large scale process without an appropriate analysis tool. The method of path analysis was developed by the geneticist to explain causal relations in population genetics (Johnson and Wichern, 1982). The goal of path analysis is to provide plausible explanations of observed correlations by constructing models of cause-and-effect relations variables. In this study we will explore this method further and develop it for process troubleshooting applications.

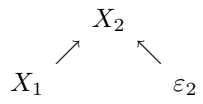
2. PATH DIAGRAM

2.1 *What is path analysis?*

The concept of path analysis is explained in this subsection according to (Johnson and Wichern, 1982). For a more comprehensive discussion on the path analysis method, readers are referred to (Johnson and Wichern, 1982) and references therein.

It is well known that a significant correlation between two variables does not imply a causal relationship. For example, the variation in both variables may be introduced by a third variable. Or one of the two variables may affect the second variable through a third variable or many other variables.

When one variable X_1 precedes another variable X_2 in time, it may be postulated that X_1 causes X_2. The relation can be represented, in the path analysis, as $X_1 \rightarrow X_2$. Taking into account the error ε_2, the path diagram may be presented as



The diagram may be written as a linear model

$$X_2 = \beta_0 + \beta_1 X_1 + \epsilon_2$$

where X_1 is considered to be a causal variable that is not influenced by other variables. The

[1] Corresponding author; biao.huang@ualberta.ca

notion of a causal relation between X_1 and X_2 requires that all other possible causal factors be ruled out. Statistically, we specify that X_1 and ε_2 be uncorrelated, where ε_2 represents the collective effect of all unmeasured variables that could conceivably influence X_1 and X_2.

To offset the influence of variable units, the regression equation is written in the standardized form as

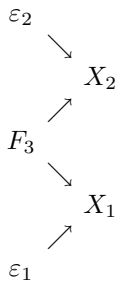
$$\frac{X_2-\mu_2}{\sqrt{\sigma_{22}}}=\beta_1\sqrt{\frac{\sigma_{11}}{\sigma_{22}}}(\frac{X_1-\mu_1}{\sqrt{\sigma_{11}}})+\sqrt{\frac{\sigma_{\varepsilon\varepsilon}}{\sigma_{22}}}\frac{\varepsilon_2}{\sqrt{\sigma_{\varepsilon\varepsilon}}}$$

or written in a compact form:

$$Z_2=p_{21}Z_1+p_{2\varepsilon}\varepsilon$$

Note that all variables including the error ε_2 now have the same variance of 1 and mean of 0. The error ε also has a coefficient. The parameters, p, in the standardized model, are defined as path coefficients.

Mathematically, it is equally logical to postulate that X_2 causes X_1 or to postulate a third model that includes a common factor. In the latter case the correlation between X_1 and X_2 is spurious and not a cause-effect correlation. The path diagram is now



where we again allow for errors in the relationship. In terms of standardized variables, the linear model implied by the path diagram above becomes

$$Z_1=p_{13}F_3+p_{1\varepsilon_1}\varepsilon_1$$
$$Z_2=p_{23}F_3+p_{1\varepsilon_2}\varepsilon_2$$

where the standardized errors ε_1 and ε_2 are uncorrelated with each other and with F_3.

A distinction is made between variables that are not influenced by other variables in the system (exogenous/input variables) and those variables that are affected by others (endogenous/output variables). With each of the latter output variables is associated a residual. Certain conventions govern the drawing of a path diagram. Directed arrows represent a path. The path diagram is constructed as follows.

(1) A straight arrow is drawn to each output (endogenous) variable from each of its source.

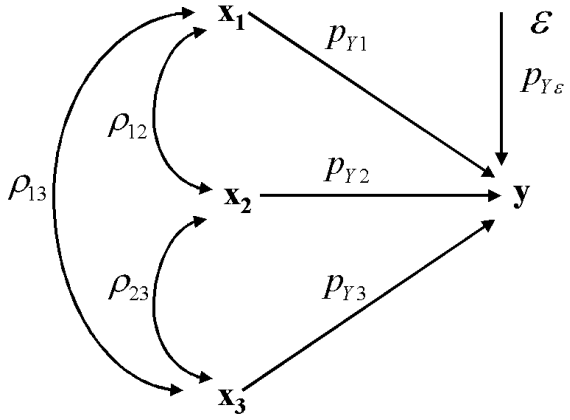


Fig. 1. *An example of path analysis*

(2) A straight arrow is also drawn to each output variable from its residual.
(3) A curved, double-headed arrow is drawn between each pair of input (exogenous) variables thought to have nonzero correlation.

The above procedure is illustrated in Fig. 1.

To calculate the coefficients for the path diagram, we use standardized variable, i.e. all variables have mean 0 and variance 1. If a regression model of original variables is given by

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_rX_r+\varepsilon$$

Then a (multivariate) regression model of the normalized variables can be constructed as

$$\frac{Y-\mu_Y}{\sqrt{\sigma_{YY}}}=\beta_1\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{YY}}}(\frac{X_1-\mu_1}{\sqrt{\sigma_{11}}})+\beta_2\frac{\sqrt{\sigma_{22}}}{\sqrt{\sigma_{YY}}}(\frac{X_2-\mu_2}{\sqrt{\sigma_{22}}})+$$
$$\cdots+\beta_r\frac{\sqrt{\sigma_{rr}}}{\sqrt{\sigma_{YY}}}(\frac{X_r-\mu_r}{\sqrt{\sigma_{rr}}})+\frac{\sqrt{\sigma_{\varepsilon\varepsilon}}}{\sqrt{\sigma_{YY}}}(\frac{\varepsilon}{\sqrt{\sigma_{\varepsilon\varepsilon}}})$$

or

$$Y_s=p_{Y1}Z_1+p_{Y2}Z_2+\cdots+p_{Yr}Z_r+p_{Y\varepsilon}\varepsilon_s(1)$$

The coefficients, $p_{Yk}=\beta_k\sqrt{\sigma_{kk}}/\sqrt{\sigma_{YY}}$ and $p_{Y\varepsilon}=\sqrt{\sigma_{\varepsilon\varepsilon}}/\sqrt{\sigma_{YY}}$ are the path coefficients or the direct effects.

An example of the path diagram is shown in Fig. 1, where $p_{Y1}, p_{Y2}, p_{Y3}, p_{Y\varepsilon}$ are the path coefficients (direct effect coefficients); ρ_{ij} is the correlation coefficient between X_i and X_j.

2.2 *Path analysis*

It is interesting to see that the correlation coefficient between Y and X_i can be constructed from the path diagram. This is shown below.

$$\rho_{YX_i}=Corr(Y,X_i)=Cov(Y_s,Z_i)$$

Using (1),

$$Cov(Y_s, Z_i) = Cov(\sum_{j=1}^{r} p_{Yj} Z_j, Z_i) = \sum_{j=1}^{r} p_{Yi}\rho_{ij}$$

which is weighted sum of the path coefficients. This correlation may be interpreted as the total effects from X_i to Y through all possible paths, and therefore this total effect is nothing but the correlation coefficient between X_i and Y. The difference between the direct effect and correlation coefficient is evident through this analysis.

Another interesting fact is the variance decomposition. Note that the following equation exists:

$$\begin{aligned} 1 &= Var(Y_s) = Var(\sum_{i=1}^{r} p_{Yi} Z_i + p_{Y\varepsilon}\varepsilon) \\ &= \sum_{i=1}^{r}\sum_{k=1}^{r} p_{Yi}\rho_{ik}p_{Yk} + p_{Y\varepsilon}^2 \\ &= \sum_{i=1}^{r} p_{Yi}^2 + 2\sum_{i=1}^{r}\sum_{k=i+1}^{r-1} p_{Yi}\rho_{ik}p_{Yk} + p_{Y\varepsilon}^2 \\ &= v_d + v_i + v_u \end{aligned}$$

This equation may be interpreted as

(Total variance of the output)
= (Contribution from direct effects)
+(Contribution from indirect effects)
+(Contribution from unknown source)

Two useful indices can be defined:

- Completeness index of the selected variables is defined as $\gamma_c = v_d + v_i$ which is bounded from 0 to 1. $\gamma_c = 0$ indicates that the selected input (independent/exogenous) variables have no effect at all on the output (dependent/endogenous) variables, while $\gamma_c = 1$ indicates that the selected input variables are complete and explain all variability in the output variables. $\gamma_c = 0.5$ indicates that 50% variance of the output variables can be explained by the selected input variables. Therefore, $\gamma_c \approx 0.5$ or $\gamma_c < 0.5$ is a typical indication that additional input variables may need to be selected for a meaningful analysis.
- Significance index of the direct effect is defined as $\gamma_d = 1 - \frac{|v_i|}{|v_d|}$. $\gamma_d = 1$ indicates that all effects are from the direct path, the input variables are mutually independent and the source of variability can be identified easily. Therefore, $\gamma_d < 0.5$ is typical indication that the source of the variability may not be isolated even though the selected input variables are sufficient to explain the variability in the output.

2.3 *Asymptotic property of path analysis*

Consider a model given by

$$y = a^T X_1 + e \tag{2}$$

where y is the variable of concern (output variable), X_1 is an input variable that directly affects y, and e is a disturbance variable that is independent of X_1. The problem of interest is to isolate the source variables X_1 from a group of input (plausible source) variables. That is, to isolate X_1 from a set of input variables, X.

Among this set of input variables, some are also affected by the same source X_1 and therefore a strong correlation apparently exists between these variables and y as well; the remaining variables are irrelevant to y. Accordingly we partition X into X_1, X_2 and X_3. X_2 is the set of input variables that are directly affected by X_1 and described by the following model

$$X_2 = FX_1 + \varepsilon \tag{3}$$

where F is a coefficient matrix of an appropriate dimension and ε with $Cov(\varepsilon) \neq 0$ is a disturbance variable vector. The posed condition $Cov(\varepsilon) \neq 0$ ensures X_2 not to include any variable that is exactly the same as one of the variables in X_1 or a linear combination of X_1. Physically, this tells us that we should not include any two or more input variables, which are exactly the same or have exact linear relationship, into the input variables. Numerically, this condition will avoid the collinearity problem in regression analysis. X_3 is a set of input variables that do not affect y and may be represented by the following model

$$X_3 = v \tag{4}$$

where v is a disturbance variable vector and is independent of both X_1 and X_2. In addition, e, ε and v are mutually independent. Now suppose we build a model of y by including all possible input variables as the input:

$$\hat{y} = l_1^T X_1 + l_2^T X_2 + l_3^T X_3 \tag{5}$$

where l_1, l_2 and l_3 are model coefficients of appropriate dimension. All variables, y, X_1, X_2 and X_3 have been normalized, namely $EX_1X_1^T = I$, $EX_2X_2^T = I$, $EX_3X_3^T = I$. Using model (5), one would like to know if the estimated model can converge to the true model (2) in the limit.

Substituting eqns (3) and (4) into eqn.(5) yields

$$\begin{aligned} \hat{y} &= l_1^T X_1 + l_2^T(FX_1 + \varepsilon) + l_3^T X_3 \\ &= (l_1^T + l_2^T F)X_1 + l_2^T\varepsilon + l_3^T X_3 \end{aligned} \tag{6}$$

Subtracting eqn.(2) by eqn.(6) yields

$$y - \hat{y} = (a - l_1 - F^T l_2)^T X_1 - l_2^T \varepsilon - l_3^T X_3 + e \tag{7}$$

Taking mean square value on both sides of eqn.(7) results in

$$\begin{aligned} & E(y-\hat{y})^2 \qquad (8) \\ & = (a - l_1 - F^T l_2)^T E[X_1 X_1^T](a - l_1 - F^T l_2) \\ & + l_3^T E[X_3 X_3^T] l_3 + l_2^T E\varepsilon\varepsilon^T l_2 + Eee^T \\ & = (a - l_1 - F^T l_2)^T (a - l_1 - F^T l_2) \\ & + l_3^T l_3 + l_2^T E\varepsilon\varepsilon^T l_2 + Ee^2 \\ & \geq Ee^2 \qquad (9) \end{aligned}$$

The equality is achieved if and only if $l_1 = a$, $l_2 = 0$ and $l_3 = 0$. The minimum of $E(y - \hat{y})^2$ is achieved in the limit by least squares. Therefore, the least squares estimation can asymptotically converge to the true model (2) even though a number of redundant/irrelevant variables have been included in the model. The implication of this result is that if X_1 is the source of the variability in y among all selected input variables, then this source can be correctly identified by checking the estimated coefficients of all input variables. The one which is statistically nonzero is likely to be the sources of variability. Other input variables (with zero coefficients), although they are also correlated with y, are in fact the response to X_1 but not the source of y, and can therefore be ruled out through this analysis. One potential problem in the calculation is the collinearity of the input variables. If two or more of input variables are highly correlated, then the regression analysis may fail. In this case, PCA/PLS based regression analysis may be applied.

Obviously, the path analysis proposed so far is limited to steady state analysis. Process dynamics such as time delay may affect the result if the disturbances are relatively fast. Thus, the algorithm discussed so far can only be applied to trace slow disturbances. Extension to dynamic application such as for oscillation detection will be discussed in the next section.

2.4 *An example on path analysis*

The following example illustrates the path analysis method. Four variables, x_1, x_2, x_3 and y are used for analysis where y is the quality variable of the concern (output variable). All four variables are highly correlated with the correlation coefficients shown in Table 1 (the last row of the table). According to this simple correlation analysis all input variables seem to have a strong correlation with y (with the minimum correlation coefficient 0.77). However, the path analysis shown in Fig.2 clearly distinguishes x_1 from others and indicates

Table 1. Correlation coefficients

	x_1	x_2	x_3	y
x_1	1			
x_2	0.76	1		
x_3	0.91	0.72	1	
y	0.94	0.77	0.83	1

that it is the real source of the variation in y. The two indices can be calculated as

$$\gamma_c = 0.89$$
$$\gamma_d = 0.90$$

Both numbers are close to 1, indicating that the selected variables are able to explain most of the variability in y and the source of the variability can be easily identified. $\gamma_c = 0.89$ also indicates that 89% variability in y can be explained by the selected variables and $\gamma_d = 0.90$ indicates that the direct effect dominates the indirect effect and therefore it is fairly easy to isolate the source of the variability.

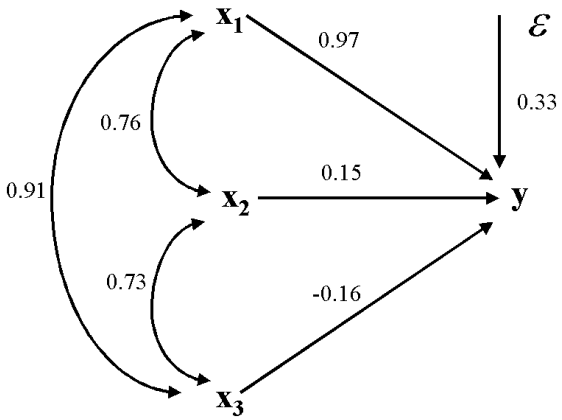


Fig. 2. *An example of path analysis*

If there are a large number of variables involved in the analysis, the graphic representation may not be efficient. The direct effect table can be constructed which lists the direct effect coefficients. For example, the direct effect of the path analysis figure shown in Fig.2 can be equally represented by Table 2. Another table is known as total effect table which shows the total effect from a input variable to the output variable by combining direct effect and indirect effect. For example, the total effect from x_1 to y, according to Fig.2, can be calculated as

$$0.97 + 0.76 \times 0.15 + 0.91 \times (-0.16) = 0.94$$

while the total effect from x_2 to y can be calculated as

$$0.15 + 0.76 \times 0.97 + 0.73 \times (-0.16) = 0.77$$

The total effect from this analysis is given in Table 3, which is exactly the same as the correlation coefficients between the input variables and the output variable y as has been discussed above.

Table 2. Direct effect table

	x_1	x_2	x_3	ε
y	0.97	0.15	-0.16	0.33

Table 3. Total effect table

	x_1	x_2	x_3	ε
y	0.94	0.77	0.83	0.33

3. APPLICATION OF PATH ANALYSIS FOR OSCILLATION DETECTION

One of the most important applications of the path analysis is for oscillation detection and tracing the source of the oscillation. Oscillation is a dynamic behavior of the process and is determined by its amplitude, frequency and phase. While the amplitude and frequency can be captured by the static path analysis, the phase lag or time delay clearly nullifies the static approach.

For oscillation detection or tracing the oscillation, one is not interested in finding the phase information of the oscillation as long as the frequency of the oscillation is captured. Autocovariance or spectrum of a time series captures the oscillation characteristics including amplitude and frequency but is independent of the phase. Therefore, applying the path analysis to the autocovariance of data will circumvent the problem of phase lag or time delay.

Thornhill et al. (2001)(Thornhill *et al.*, 2001*a*) have presented a MATLAB function to calculate filtered autocovariance and spectrum of time series. The same algorithm is used here for dynamic path analysis. A set of data (courtesy of a SE Asian refinery) has been used for oscillation detection in Thornhill et al. (2001)(Thornhill *et al.*, 2001*b*). The same set of data is revisited by applying the path analysis to the autocovariance of the data. The process diagram is shown in Fig.3.

The question is to trace the source of the oscillations. The path analysis is applied to the autocovariance of the data to search for the source, which constitutes the following five steps:

Step 1: Draw a control volume around PSA unit (see Fig.3 for control volume 1). Tag 11 and 34 as output variables; Tag 19 and 20 as input variables. Tag 3 is not an independent input as it is found to

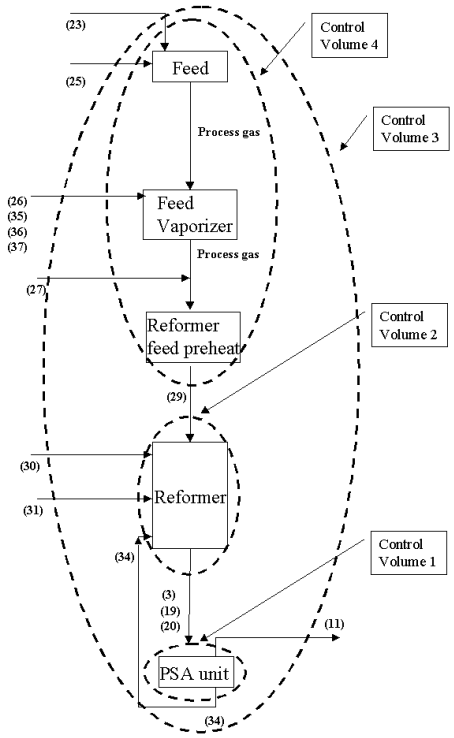


Fig. 3. *Schematic of process and control volumes*

Table 4. Summary of indices

	γ_c	γ_d
Tag 11	0.98	0.94
Tag 34	0.85	0.36

have an identical shape as Tag 20. Path analysis yields the following results:

(1) The completeness index and significance index are calculated and shown in Tab4: The first column of the table shows that Tag 19 and 20 can explain most variability of Tag 11 and 34. The second column shows that the source of Tag 11's variability can be easily identified while the source of Tag 34 variability may not be isolated easily.

(2) The direct effect table is shown in Tab.5. The

Table 5. Direct effect table

	Tag 19	Tag 20	ε
Tag 11	0.92	0.07	0.14
Tag 34	0.26	0.71	0.25

first column clearly indicates that Tag 19 is the source of Tag 11. The second column shows that Tag 20 is possibly the main contributor to Tag 34 but Tag 19 also has a considerable contribution. Therefore, a unique source of Tag 34 can not be identified.

Step 2: Draw a control volume around Reformer (see Fig.3 for control volume 2). Tag 19 and 20 as output; Tag 29, 30, 31, 34 as inputs. In this step, we will analyze Tag 19 only. The analysis with Tag

20 as output will be performed in the next step. Path analysis yields the following results:

(1) The two indices are calculated as $\gamma_c = 0.93$ and $\gamma_d = 0.92$. These results indicate that most of the variability in Tag 19 can be explained by the selected inputs and in addition the source can be easily identified.
(2) The direct effect table is summarized in Tab.6. This result clearly indicates that Tag 34 is the source of Tag 19.

Table 6. Direct effect table

	Tag 29	Tag 30	Tag 31	Tag 34	ε
Tag 19	0.04	0.06	0.03	0.93	0.26

Step 3: Continuation of Step 2 with Tag 20 as output.

(1) The two indices are calculated as $\gamma_c = 0.94$ and $\gamma_d = 0.97$. These results indicate that most of the variability in Tag 20 can be explained by the selected inputs and the source can be easily identified.
(2) The direct effect table is summarized in Tab.7. This result clearly shows that Tag 34 is the source of Tag 20.

Table 7. Direct effect table

	Tag 29	Tag 30	Tag 31	Tag 34	ε
Tag 20	-0.02	0.03	0.01	0.95	0.25

Comments: Step 2 and 3 indicate that Tag 34 is actually the source of both Tag 19 and 20. This result explains why Tag 34 can not find its source from Tag 19 or 20 in Step 1.

Step 4: Draw a control volume around the whole process as shown in the flowchart (see Fig.3 for control volume 3). Tag 11 is the output (Tag 34 is a recycle stream and not an output); Tag 23, 25, 26, 27, 35, 36, 37, 30, 31 are the inputs. Some inputs such as light naphtha flow rate is not available and has not been included in the analysis. Path analysis yields the following result shown in Table 8. Due to the space limit, direct path coefficients with small values are omitted from the table. The two indices $\gamma_c = 0.96$ and $\gamma_d = 0.90$ indicate that the selected inputs are sufficient to explain the output variability and the source of the variability can be easily identified. The direct path coefficient from Tag 25 to Tag 11 clearly shows that Tag 25 is the source of the oscillation in Tag 11. Combining with the results obtained in previous steps, now the question is which one of Tag 25 and Tag 34 is the source of the oscillation. If there is no recycle from Tag 34 to Tag 25, then the result obtained in this step clearly shows that Tag 25 is the real source and Tag 34 is actually a response to Tag 25. However, if there is a recycle from Tag 34 to Tag 25, then Tag 34 could be the source, a result obtained in Thornhill et al.(2001)(Thornhill *et al.*, 2001*b*).

Table 8. Direct effect table

	γ_c	γ_d	Tag 25	ε
Tag 11	0.96	0.90	1.0	0.20

Step 5: Draw a control volume around Feed unit, Feed vaporizer/superheater unit and Reformer feed pre-heat unit (see Fig.3 for control volume 4). Tag 29 is taken as output; Tag 23, 25, 26, 35, 36, 37, 27 as inputs. Path analysis yields $\gamma_c = 0.37$, the selected input variables are not sufficient to explain the variability in Tag 29. Therefore, the source of the Tag 29 oscillation can not be identified from the given tags.

4. CONCLUSIONS

In this paper the path analysis is proposed for process troubleshooting by tracing the source of variability/oscillation. Path analysis is similar to correlation analysis in terms of its simplicity but it provides a directional correlation information. That is, a correlation analysis reveals all possible correlation between two variables, direct and indirect, while path analysis reports the direct relation of two variables. It is shown in this paper that path analysis can be used to trace the source of process variability. The result has also been extended to tracing the oscillation by applying the path analysis to autocovariance data. An industrial case study is presented to illustrate the effectiveness of the proposed algorithms.

5. REFERENCES

Johnson, R.A. and D.W. Wichern (1982). *Applied Multivariate Statistical Analysis*. Prentice-Hall.

Thornhill, N., B. Huang and H. Zhang (2001*a*). Detection of multiple oscillations in control loops. *To appear in Journal of Process Control*.

Thornhill, N.F., S.L. Shah and B. Huang (2001*b*). Detection of distributed oscillations and root-cause diagnosis. In: *Proceedings of CHEMAS*. Cheju Island, Korea.