DETERMINATION OF PHYSIOLOGICAL MODES IN *SACCHAROMYCES-CEREVISAE* CULTURE USING SEQUENTIAL DATA ANALYSIS.

J. Ph. Cassar, V. Guillou

LABEM, Polytech'Lille - USTL
Bd. P. Langevin 596555 Villeneuve d'Ascq Cedex France
cassar@univ-lille.fr

Laboratoire de Synthèse et Physicochimie de Molécules d'Intérêt Biologique (SPCMIB) UMR 5068
Université Paul Sabatier 118 Route de Narbonne F.31062 Toulouse Cedex 4
guillou@chimie.ups-tlse.fr

Abstract: Detecting the physiological mode of microorganism during a culture may allow to better handle behaviour modifications that can affect the productivity. This paper proposes a new sequential method that utilises a classification of the directions that generate the subspace orthogonal to the pre-processed measurements. Existence of the orthogonal subspace and the nature of pre-processing are derived from theoretical modelling of the culture. An application to a *Saccharomyces* Cerevisae culture shows the ability to well isolate main phases of the process.

Keywords: data analysis, monitoring, biological application, and physiological states.

1. INTRODUCTION

This paper deals with the monitoring of cultures that involve a single strain of microorganism. The biological reactions that act in this process include microbial growth, maintenance and production reactions (Bastin and Dochain, 1990). In these types of reaction, substrates are consumed and are transformed either into biomass or into products. Modelling of biological process is for a long time the purpose of many studies. Its interest comes from the necessity of acquiring information about the behaviour of microorganism during the cultivation to better understand its dynamics. In another hand, it intends to predict some information as microorganism or substrate concentration.

In most of the proposed methods, one assumption is made about the set of reactions that is actually acting at each instant and an activation function has to be set for each reaction. On the contrary, in (Cassar *et al.*2004, Cassar and Guillou 2004) basic reactions and physiological states are first defined from a dynamical model introducing yield matrix coefficients already introduced in (Bastin and Dochain, 1990, Chen and Bastin, 1996). A set of relations to be verified by the measured indicators is associated with each physiological state. The monitoring purpose is then to detect the changes from one state to other one.

The proposed approach aims at establishing a link between such a dynamical modelling of fermentation process and data analysis techniques. Data analysis techniques have been widely used for monitoring processes – see (Kourti and MacGregor, 1995) for a good review. A fuzzy classification procedure is utilised in (Waisman, 2000) to identify the current physiological state from the processing of the whole measurements set. In a similar approach (Cassar and

Guillou, 2004), a hierarchical classification is applied using PCA projections to better distinguish the different modes.

The limits of such approaches rests on the fact that the weight of one mode depends on its relative points number. So sequential methods have to be used to better distinguish short lasting events. Stephanopoulos (and al. 1997) proposed a pattern recognition approach that is performed after signals decomposition into symbols. In this paper, a multivirate signal decomposition is performed by using projections derived from application of PCA techniques on sliding time windows.
An example, using the well-known yeast *Saccharomyces Cerevisiae* fermentation, is used to illustrate all the concepts that are developed in this paper. In a first part, the modelling of the culture process is introduced to define from measurements which data must be used by PCA techniques. The second part introduces the principle of PCA methods and establishes the link with the models. A sequential data analysis is proposed to classify the measurements into different states. The last part presents and discusses the results obtained from a batch culture of *Saccharomyces Cerevisiae.* The conclusion exhibits the interests and the perspectives of this work.

2. PHYSIOLOGICAL MOTIVATIONS.

2.1. Fermentation process modelling

The fermentation in a bioreactor involves products and microorganisms. Regarding the evolution of product concentrations in the liquid phase and microbial growth, the dynamic model rests on the balance equations that are applied on each product present in the liquid phase. Thus, fermentation processes are usually described by dynamical model (1) (Bastin and Dochain, 1990):

$$\frac{d}{dt}\xi = -D\xi + \mathbf{F} + \mathbf{Q}(\xi) + \mathbf{K}^{*}.\mathbf{r}(\xi) \qquad (1)$$

The n_p dynamical state vector ξ is the concentration vector of products, included ions H^{+}, involved in the reactions. In this model, the matrix $\mathbf{K}^{*} = \left[k_{ij}\right]$ is the yield coefficient matrix that expresses the stoichiometry of the reactions involved by the culture. By convention negative yield coefficients are associated with substrates in the reaction while positive ones are associated with product. A zeroed value indicates the component is not involved in the reaction. $\mathbf{F}$ and $\mathbf{Q}(\xi)$ are respectively the input flows relative to the volume for the liquid and for the gas. The notation $\mathbf{Q}(\xi)$ expresses here the influence of the value of the continuous state vector ξ on flow exchange rate between the gas the liquid phase. The evolutions of the production or consumption rates $\mathbf{r}(\xi)$ are also functions of this vector.

As a batch process is considered in this study, the fact of neglecting the influence of the basic flow on volume makes it possible to cancel the dilution term D. The term $\mathbf{F}$ only concerns the base flow used to regulate the pH value. Relation (1) is then broken up into two relations according to the gas or liquid mass balances.

$$\begin{aligned}\frac{d}{dt}\xi_L &= \mathbf{F} + \mathbf{K}_L.\mathbf{r}(\xi)\\ \frac{d}{dt}\xi_G &= \mathbf{Q}(\xi) + \mathbf{K}_G.\mathbf{r}(\xi)\end{aligned} \qquad (2)$$

2.2. Measurements

In relation (2), some terms are measured. Measured pH-value allows calculating $\left[H^{+}\right]$-concentration. Gas consumption $d_{[O_2]}$ and production $d_{[CO_2]}$ are measured too. Base flow rate is governed by pH control. It provides information on ions flow rates $d_{\left[H^{+}\right]}$ and $d_{\left[NH_4^{+}\right]}$. In the following relations, $\mathbf{C}_1$ and $\mathbf{C}_2$ are appropriated selection matrices and measurement are written as:

$$\begin{aligned}\mathbf{y}_G &= \mathbf{Q}(\xi) = \begin{bmatrix} d_{[CO_2]} \\ d_{[O_2]} \end{bmatrix}\\ \begin{bmatrix} 1 \\ -1 \end{bmatrix} y_L &= k\,\mathbf{F} = \begin{bmatrix} d_{\left[NH_4^{+}\right]} \\ -d_{\left[H^{+}\right]} \end{bmatrix}\\ y_{\left[H^{+}\right]} &= \mathbf{C}_1\xi = \left[H^{+}\right]\end{aligned} \qquad (3)$$

The introduction of measurements (3) into model (2) leads to the following formulation with respect to the measured species.

$$\frac{d\left[NH_4^{+}\right]}{dt} = y_L + \mathbf{C}_2\mathbf{K}_L\mathbf{r}(\xi) \qquad (a)$$

$$\frac{dy_{\left[H^{+}\right]}}{dt} = -y_L + \mathbf{C}_1\mathbf{K}_L\mathbf{r}(\xi) \qquad (b) \qquad (4)$$

$$\frac{d\xi_G}{dt} = \mathbf{y}_G + \mathbf{K}_G\mathbf{r}(\xi) \qquad (c)$$

As $\left[NH_4^{+}\right]$ and gas concentrations are not measured, relations (4).b and (5).c can only be used if their derivatives are supposed to be equal to zero. That is mostly verified for gas, and supposes that $\left[NH_4^{+}\right]$ is completely consumed by the growth of biomass. In this case, (4) can be written as:

$$\mathbf{y} = \mathbf{K}\,\mathbf{r}(\xi) \qquad (5)$$

$$\text{with } \mathbf{y} = \left[-y_L \quad \left(\frac{dy_{[H^+]}}{dt} + y_L \right) \quad -\mathbf{y}_G{}^T \right]^T$$

$$\text{and } \mathbf{K} = \left[\mathbf{C}_3\mathbf{C}_1\mathbf{K}_L \quad \mathbf{C}_2\mathbf{C}_1\mathbf{K}_L \quad \mathbf{K}_G \right]^T$$

2.3. Physiological modes

In relation (5), $\mathbf{r}(\xi)$ expresses all the biochemical reactions that can act during the fermentation. In (Cassar and al., 2004), we propose a writing of these reactions such that a way that each $\mathbf{K}$-column can be uniquely defined from the stoichiometry matrix of the involved products. Only a subset of these reactions is active at a given time. This subset is induced by the physiological mode of the strain. In the sequel of this paper, the physiological mode I is formally defined by the corresponding set of reactions $\mathbf{r}^{(I)}$.

At a given time k, relation (5) becomes:

$$\mathbf{y}_k = -\mathbf{K}^{(I)}\,\mathbf{r}^{(I)}(\xi_k) \qquad (6)$$

where $\mathbf{K}^{(I)}$ contains the $\mathbf{K}$-columns associated with the active reaction of the biological mode I.

In this study, a *Saccharomyces Cerevisae* strain is considered. Nine basic biochemical reactions that involve glucose, ethanol and acetic acid as main products are given in (Cassar and Guillou, 2004). They combination leads to nine physiological modes according to the available substrates.

In (Cassar and al., 2004), it is shown how biological modes can be associated with different sets of relations between the available indicators. In the sequel of this paper, we are interested in finding intervals of time during which that relations may be verified.

3. METHOD PRÉSENTATION

3.1. Objectives

If the microorganism is supposed to be in a given physiological state, then the $\mathbf{y}$-vector components are generated from the reaction rates by a linear combination expressed by the $\mathbf{K}^{(I)}$-matrix in relation (6). They are thus linked by a linear combination that doesn't change while the microorganism remains in the same physiological state.

If a single physiological mode is considered to be active in a given time window, relation (6) can be written as:

$$\mathbf{Y}_i = \mathbf{K}^{(I)}\,\mathbf{R}_i + \mathbf{E}_i \qquad (7)$$

$\mathbf{Y}_i = [\mathbf{y}_k]_i^{i+n} = [\mathbf{y}_i \quad \mathbf{y}_{i+1} \quad \cdots \quad \mathbf{y}_{i+n}]$ gathers the $\mathbf{y}_k$-vectors. $\mathbf{R}_i$ is built in the same way with the reaction rates $\mathbf{r}^{(I)}$ and $\mathbf{E}_i$ contains the stochastic components that have to bee introduced into the model.

The first objective aims at determining the $\mathbf{r}^{(I)}$-vectors dimension – i. e. the number of active reactions - that defines the dimension of the sub-space spanned by the $\mathbf{y}_k$-vectors.

The second objective is the decomposition of the available data $\mathbf{Y}_i$ into two terms: the matrix that expresses the linear combination $\mathbf{K}^{(I)}$ and the reactions rate signals $\mathbf{R}_i$. However, this decomposition is not unique as shown in (Liao and al., 2003). Indeed, relation (7) can always be written as:

$$\mathbf{K}^{(I)}\,\mathbf{R}_i = \mathbf{K}^{(I)}\mathbf{U}'\mathbf{U}\,\mathbf{R}_i = \mathbf{A}\,\mathbf{F}_i \qquad (8)$$

$\mathbf{U}$ is an orthogonal matrix such that $\mathbf{U}'\mathbf{U} = \mathbf{I}$, $\mathbf{A}$ and $\mathbf{F}$ are an alternate decomposition of the signals $\mathbf{Y}$. However, let $\mathbf{W}$ be a matrix left orthogonal to $\mathbf{K}^{(I)}$. This matrix is also orthogonal to $\mathbf{A}$ and doesn't thus depend on the decomposition. The changes in $\mathbf{K}^{(I)}$ can be checked as a change of the orthogonal space generated by $\mathbf{W}$. A segmentation of the $\mathbf{Y}$-signal is now to be found where each segment corresponds to a different orthogonal subspace. Providing a time segmentation based on this segmentation that can be interpreted as sequence of physiological states constitutes the third objective of our study that.

3.2. Orthogonal subspace dimension determination.

A time window i is characterised by the dimension of the $\mathbf{Y}_i$ orthogonal sub-space and by the base that generates this sub-space. These both data are provided by a PCA analysis. The number of eigenvalues of $\mathbf{Y}_i'\mathbf{Y}_i$ that can be considered as near from zero gives the orthogonal sub-space dimension. The base of the orthogonal subspace is given by the corresponding eigenvectors.

Let λ be the eigenvalue associated to an eigenvector $\mathbf{w}$. $(\hat{\sigma})^2 = \dfrac{\lambda}{n-1}$ estimates the value of the $\mathbf{w}'\mathbf{Y}_i$ component variance. An eigenvalue will be considered as close to zero if this estimated variance doesn't differ in a significant way from the expected variance deduced from the stochastic term $\mathbf{E}_i$. This expected value is not *a priori* known. The proposed approach estimates it by the error variance of a polynomial curve fitting applied on the time window.

Let the hypothesis H_0 indicates the model (7) holds. Then the expected eigenvalues are defined as:

$$(\sigma)^2 = \|\mathbf{w}'\mathbf{E}\| \qquad (9)$$

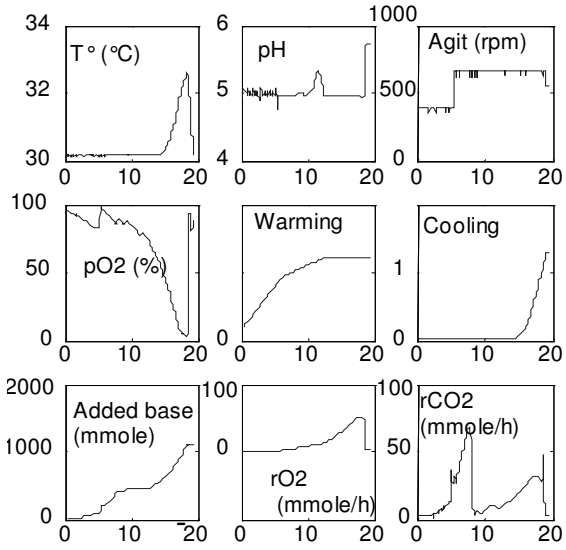

Figure 1: On line measurement (time in hours)

Let H_1 be the hypothesis the model (7) does not hold. Supposing the stochastic terms follow a Laplace Gauss distribution, the Fisher's test is used to distinguish the hypotheses:

$$\frac{(\hat{\sigma})^2}{(\sigma)^2} \underset{H_0}{\overset{H_1}{\gtrless}} F_{lim} \qquad (10)$$

Whether the H_1 hypothesis is chosen, then **w**-vector can not be considered to be orthogonal to the available data. The number of eigenvalues that match the H_0 hypothesis then gives the dimension of the orthogonal subspace.

3.3. Classifying the eigenvectors directions.

As PCA that provides the eigenvectors is performed independently on each time window, the eigenvectors may express different directions while the generated orthogonal subspace remains the same. Let $\mathbf{W}_i$ be the matrix of eigenvectors that generates the orthogonal subspace on the actual time window.

Each subspace is linked with a base $\mathbf{W}^{(I)}$. It has to be checked whether one or several directions of this base can be generated from the matrix $\mathbf{W}_i$. These directions are a linear combination $-\mathbf{W}_i\mathbf{n}$ - of the $\mathbf{W}_i$ directions. Let $\mathbf{W_P}$ be the matrix that gathers the already encountered directions. For a given direction of $\mathbf{W_P}$ - named $\mathbf{w}_p$ – the linear combination $\mathbf{n}$ that leads to the closest direction has to be found: $\mathbf{n} = \arg\min_{\mathbf{n}}(d(\mathbf{w}_P, \mathbf{Wn}))$ An angular distance is chosen, as the directions are normalised vectors. That leads to projection maximisation: $\max_{\mathbf{n}}\left(\mathbf{w}_p{}'\mathbf{Wn}\right)$ under the constraint

$$\|\mathbf{Wn}\| = 1 \qquad (11)$$

The solution is (see demonstration in appendices): $\mathbf{n} = -\frac{\mathbf{W}_i{}'\mathbf{w}_p}{\sqrt{\mathbf{w}_p{}'\mathbf{W}_i\,\mathbf{W}_i{}'\mathbf{w}_p}}$. The criterion maximal value is then $c_{p,i} = \|\mathbf{w}_p\|_{\mathbf{W}_i\mathbf{W}_i'}$ that is the $\mathbf{W}_i\,\mathbf{W}_i{}'$ weighted norm of $\mathbf{w}_p$. This criteria is used by a classification whose algorithm is given table 1.

Each biological state is defined by the given set of directions that gives the $\mathbf{W}^{(I)}$ matrix.

Table 1: Directions classification algorithm

FOR each time window
  Determine the D dimension of the orthogonal subspace.
  The D greatest values of c_i are kept. Let $I = \{p | c_{p,i} > c_{lim}\}$ be the set of the indices of the corresponding directions that belongs to the subspace generated by $\mathbf{W}_i$.
  IF card (I) = D
    I indices give the directions associated to an already found mode.
    The directions in $\mathbf{W_P}$ are actualised
  ELSE
    The directions given by I are kept.
    New directions in the subspace generated by $\mathbf{W}_i$ are calculated to be orthogonal to the already kept directions.
    This new directions are added to $\mathbf{W_P}$
  ENDIF
ENDFOR

4. APPLICATION

4.1. Data presentation

The classification has been applied on a data set obtained from a batch culture of Saccharomyces Cerevisiae performed in the INSA LBB laboratory in Toulouse (France) (Poilpré, 2002) In this experiment high glucose concentration is used.

Figure 1 exhibits culture measured temperature, measured pH, stirring speed (Agit), dissolved oxygen, cooling action (act frd), amount of added base (ajout base), oxygen consumption and carbon dioxide production. pH regulation only involves base addition and is thus limited when acidification is needed.

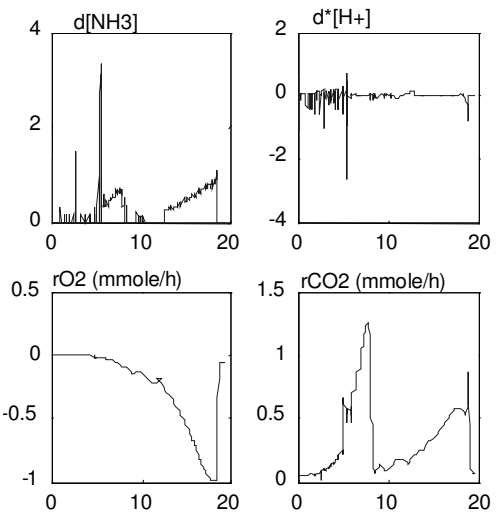

Figure 2: Combined variables used for classification

4.2. Measurements pre processing and orthogonal subspace determination

The four **y** components defined in (5) are presented Figure 2. In order to avoid scaling effects, a normalisation constraint 90% of each variable value to be in the range [-1,1]. Effects of few extra range values on the normalisation are thus avoided.

Figure 3 gives the logarithm of the criterion ratio (10) . The logarithm allows to better exhibit the values very close to zero. The threshold value is fixed to 0,69 (log (2)) according to the F distribution law with a risk chosen at 0,99.

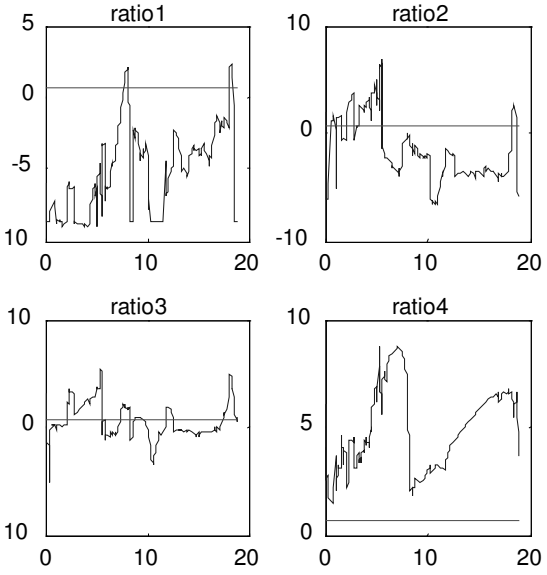

Figure 3: Logarithms of the variance ratios and decision thresholds.

4.3. Classification results.

Applying the procedure described in the previous section - with a window size n fixed to 15 - allows distinguishing 10 different directions that allow generating all the orthogonal subspaces. A given mode is associated with a set of directions that defined a base of the associated orthogonal subspace. The limit threshold on the angular distance c_{lim} was fixed to 0.9. Figure 4 shows the time evolution of the indexes of the recognised directions . The number of directions that act at a given time gives the dimension of the orthogonal subspace. This dimension expresses the number of eigenvalue whose ratio remains below the threshold. The number of acting relations is thus determined to be at least equal to (4 – this dimension).

In figure 5, segmentation of the culture time into modes is presented. Seven significant modes (size upper than 9) are generated. They gather 88% of the whole set of data (502). Seven periods can be detected in this figure. The two first one (mode 1 and 2) corresponds to the beginning of the culture with exponential growth on glucose. Next period (mode 3) is induced by the increase in agitation rate and gas flow rate whose effect is a better oxygen transfer to the yeast. At this stage, growth is very efficient with strong oxygen requirement. First part of mode 4 exhibits a strong decrease in yeast growth, as depicted by carbon dioxide production, due to the beginning of glucose depletion. The second part exhibits additionally a small acid consumption as indicated by the pH free evolution as the pH regulation can't act above the set point value.

A strong acid consumption is observed in mode 5 and depicted by a high increase of pH value. The yeasts depleted of glucose can't metabolise directly ethanol, the main product of glucose fermentation but metabolise the acid produced during this stage. This period and the small period between modes 4 and 5 correspond to Diauxie that is the delay required by yeast to adapt its metabolism to the new substrate ethanol.

Next period (mode 6) is the beginning of yeast oxydative growth on ethanol. It goes on with a strong exponential growth (mode 7) until substrates are depleted (expressed by mode 4).

5. CONCLUSION

A segmentation method of a culture of *Saccharomyces Cerevisae* has been proposed and shown to be efficient on an example of such culture. This approach is based on a classification of the directions that generates the subspace orthogonal to the pre-processed measurements. Nature of the pre processing and existence of the orthogonal subspace are deduced from the theoretical modelling of the fermentation process. The obtained time decomposition is deeply related to biological

interpretation of the deferent stage of the batch process. As the structure of the constraints imposed by the stoichiometry doesn't match the structural conditions given by (Liao *et al.,* 2003) it will not be possible to estimate the reaction rates without matching each observed mode with an expected physiological mode. That constitutes the perspective of this work

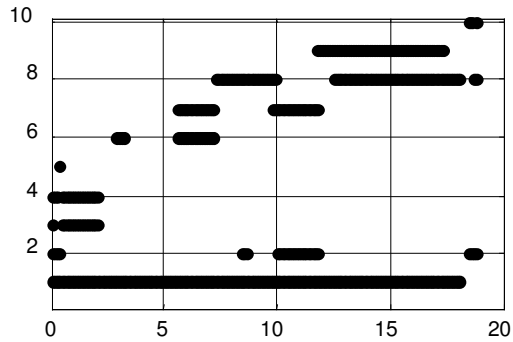

Figure 4: Recognised directions

REFERENCES

Bastin G., Dochain D. (1990). On Line Estimation and Adaptative Control of Bioreactors. **Amsterdam**, Elsevier

Cassar J. Ph., Guez J. S., Jacques P. (2004 a). Stoechiometric Modelling for Physiological State Changes Monitoring in Bioprocesses. *CAB04 Computer Advanced in Biotechnology,* **Nancy, France**

Cassar J. Ph., Guillou V. (2004 b). Hierarchical Data Analysis Based Biological States Recognition. *ICEF9 International Conference Engineering and Food,* **Montpellier, France**

Cassar J. Ph., J. S. Guez, P. Jacques (2003). A Hybrid Approach for fhe Monitoring of Physiological State Changes. *ADSH03,* **St Malo, France**, 93-98

Chen L., G. Bastin (1996). Structural Identifiability of the Yield Coefficients in Bioprocess Models when Reaction Rates are Unknown. *Mathematical Biosciences*, **132**, 35-67

Liao J. C., Boscolo R., Yang Y., Tran L. M., Sabatti S. and Roychowdhury V. P. (2003). Network component analysis: Reconstruction of regulatory signals in biological systems. *PNAS,* **100-** 26, 15522—15527

Poilpre E. (2002). Mécanisme d'adaptation rapide de Saccharomyces cerevisiae en métabolisme oxydatif. *INPT PhD Thesis,* **Toulouse, France**, 164 pages

Raich A., Cina A. (1997). Diagnosis of process disturbances by statistical distance and angle measures. **Vol. 21, No. 6**, pp. 661Pergamon

Stephanopoulos G., & al (1997). Fermentation Database Mining by Pattern Recognition. *Biotechnology and Bioengineering*, **53**, 443-452

Waissman J (2000). Construction d'un modèle comportemental pour la supervision de procédés : application à une station de traitement des eaux. *INPT Thesis,* Toulouse, France

APPENDICE

The Jacobean associated to the optimisation problem is: $J = (\mathbf{w}_P{}'\mathbf{W}_i\,\mathbf{n}) + \lambda(\mathbf{n}'\mathbf{W}_i{}'\mathbf{W}_i\,\mathbf{n} - 1)$

$$\frac{dJ}{d\mathbf{n}} = \mathbf{W}_i{}'\mathbf{w}_P + 2\lambda(\mathbf{W}_i{}'\mathbf{W}_i\,\mathbf{n}) = 0$$

As $(\mathbf{W}_i{}'\mathbf{W}_i) = \mathbf{I}$ because the $\mathbf{W}_i$ columns are normalised and orthogonal, n is given by:

$$\mathbf{n} = -\frac{\mathbf{W}_i{}'\mathbf{w}_P}{2\lambda}$$

$\frac{dJ}{d\lambda} = \mathbf{n}'\mathbf{W}_i{}'\mathbf{W}_i\,\mathbf{n} - 1 = \mathbf{n}'\mathbf{n} - 1 = 0$. So, we can write by substituting the **n** value: $\frac{\mathbf{w}_P{}'\mathbf{W}_i\,\mathbf{W}_i{}'\mathbf{w}_P}{4\lambda^2} = 1$ and

$$\lambda^2 = \frac{\mathbf{W}\mathbf{P}_i\,\mathbf{W}'\mathbf{W}\,\mathbf{W}\mathbf{P}_i{}'}{4}$$

Then $\mathbf{n} = -\frac{\mathbf{W}_i{}'\mathbf{w}_p}{\sqrt{\mathbf{w}_p{}'\mathbf{W}_i\,\mathbf{W}_i{}'\mathbf{w}_p}}$

The criteria maximal value is then equal to:

$$-\frac{\mathbf{w}_P{}'\mathbf{W}_i\,\mathbf{W}_i{}'\mathbf{w}_p}{\sqrt{\mathbf{w}_p{}'\mathbf{W}_i\,\mathbf{W}_i{}'\mathbf{w}_p}} = \sqrt{\mathbf{w}_p{}'\mathbf{W}_i\,\mathbf{W}_i{}'\mathbf{w}_p} = \|\mathbf{w}_P\|_{\mathbf{W}_i\,\mathbf{W}_i{}'}$$

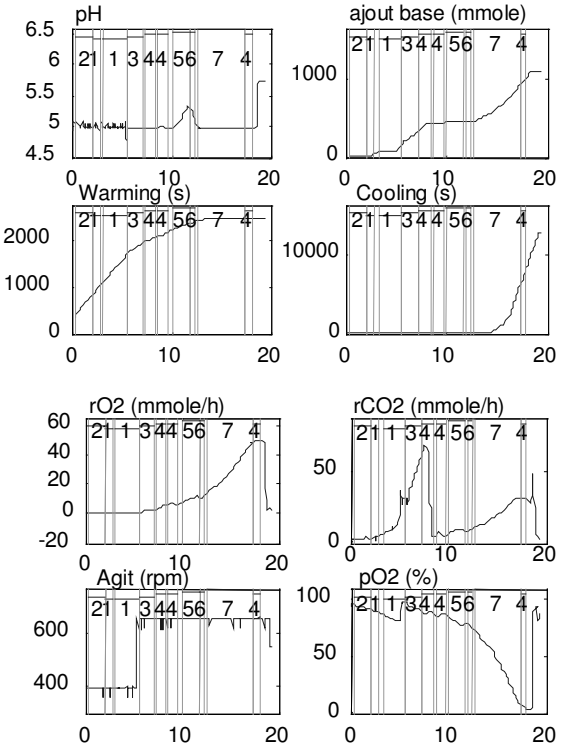

Figure 5: Time repartition of the recognised modes