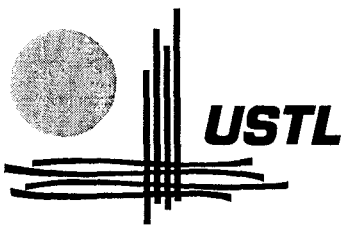


U.F.R. I.E.E.A. Numéro d'Ordre : 2659

# Apprentissage de Modèles de la Dynamique pour l'Aide à la Décision en Monitorage Clinique

THÈSE

présentée en vue de l'obtention du

**Doctorat de l'Université des Sciences et Technologies de Lille**
**spécialité Automatique et Informatique Industrielle**

par

Daniel CALVELO AROS


Soutenue le 13 décembre 1999 devant le jury formé par :

| | | |
|---|---|---|
| *Président :* | Marcel STAROSWIECKI | Université de Lille 1 |
| *Directeur de Thèse :* | Max DAUCHET | Université de Lille 1 |
| *Rapporteurs :* | Catherine GARBAY | Institut Albert Bonniot - Grenoble |
| | Stéphane CANU | Institut National des Sciences Appliquées de Rouen |
| | Boutaïb DAHHOU | Université Paul Sabatier - Toulouse |
| *Examinateurs :* | Marie-Christine CHAMBRIN | INSERM - Université de Lille 2 |
| | François FOURRIER | Université de Lille 2 |
| | Denis POMORSKI | Université de Lille 1 |

Laboratoire d'Automatique et Informatique Industrielle de Lille UPRES A 8021

# Remerciements

Je tiens à remercier avant tout l'équipe Aiddiag, pour la confiance, le soutien, les merveilleuses pauses-café. Marie : l'âme, la tête et les mains ; tu nous fabriques la raison. Pierrot, vole toujours par-dessus nos pauvres matérialismes. Christian, continue d'assurer comme une bête. Nathanaël, les choses reposent et se reposent sur toi. Alex, la science te rattrapera où que tu te caches. Alors, Annick, pas envie de refaire des folies par ici ? Krys, Martin, Laurent, Luc, Nico, les autres, vous reviendrez ?

Merci Denis et Max pour l'opportunité. Les opportunités.

Merci aux labos de math et de physique tout entiers pour tout et pour ces vendredis mémorables.

Merci à Mme Garbay et MM. Canu et Dahhou pour avoir accepté de rapporter ce travail, pour leur perspective et les discussions.

Merci Claude, François, Marcel pour votre capacité à donner aux autres un aperçu de l'intelligence.

Ya, y ahora lo sentimental. Gracias a la familia por la confianza y el apoyo, aunque suene huachafo. La familia chica y la grande. Lea, un espacio para tu mirada gigante. Y no te digo más porque no me da el sitio y estas cosas formales, a mí tú sabes. Los de más lejos pero cerquísima ¿qué querés que les diga ? Prefiero abrazarlos de otra manera que será la misma pero diferente, tres o diez años más tarde. $S^n$, mon concentré d'amitié en heures-clope, merci.

Mes directeurs de thèse sont responsables de l'ambition et la rigueur que vous pouvez trouver dans ce travail. Je plaide coupable pour les défauts.

Enfin, merci à ceux qui font pour d'autres sans attendre de réciproque. Nous ne vous donnons jamais assez la pareille.

Ce travail a été réalisé dans le cadre d'une collaboration entre le Laboratoire d'Informatique Fondamentale de Lille, le Laboratoire d'Automatique et Informatique Industrielle de Lille et l'Unité 279 de l'Inserm. Il a bénéficié d'une bourse du Ministère de la Recherche et de l'appui du CH&U de Lille par le PHRC 9501. Les travaux ont été réalisés dans les locaux du Laboratoire de Biomathématiques de la Faculté de Pharmacie de Lille et de l'Institut de Technologie Médicale du CH&U de Lille.

# Table des matières

# Table des figures

*Les ordinateurs sont inutiles : ils ne donnent que des réponses.*

Attribué à Pablo Ruiz Picasso

# 1

# Introduction

Où les problèmes sont posés et les voies de solution proposées.

## Sommaire

La problématique générale de l'aide à la décision en monitorage médical a été abordée depuis deux perspectives majeures. D'un côté, par le traitement du signal à partir des différentes mesures que l'on peut effectuer sur les données vitales du patient. De l'autre, par des méthodologies de modélisation de la connaissance, à partir des connaissances et des discours médicaux. Il s'agit des deux extrêmes d'une chaîne de traitement liant les *paramètres physiologiques*, mesurés par des capteurs à travers les moniteurs, à l'interprétation du personnel soignant.

Le positionnement médian dans cette chaîne est exploré depuis peu, au sein de plusieurs projets [CPM96, LHRG96, MD96, MHPP96, Sha94, DPG+97, Ste96, HK96]. Ces recherches portent essentiellement sur l'aide à la décision en Réanimation (adulte et néonatale) et en Anesthésie. Il s'agit en effet des domaines où le monitorage est le plus intensif, en matériel et en variables de monitorage.

Ces systèmes d'aide à la décision clinique peuvent remplir trois fonctions majeures [Coi94] :
- L'aide au monitorage et à la surveillance, par le biais d'"alarmes intelligentes".
- L'appui à l'investigation clinique, par le suivi de protocoles [CRC+ss].
- L'assistance à l'interprétation des cas cliniques, par les possibilités de résumé historique.

De façon générale, l'approche de l'aide à la décision peut être exprimée en terme de transformation de signaux en données puis en information, enfin en connaissance. Cette chaîne est une chaîne d'*abstraction.*

Nous entendons par *signal* la suite temporelle des valeurs d'une certaine mesure.[1] Il s'agit d'une valeur numérique ou catégorique brute, interprétable uniquement en termes physiques.

La *donnée* se place à un niveau d'abstraction légèrement supérieur. La donnée *caractérise* le signal, mais reste une valeur numérique (ou catégorique). C'est le concept de donnée qui est appelé "information" dans la Théorie de l'Information, par exemple. Nous nous plaçons pour l'instant dans une perspective davantage cognitive qu'informatique.

Le passage au niveau d'*information* implique l'introduction d'un *contexte* d'interprétation. Le passage de la donnée ou du signal à l'information implique déjà la prise en compte du passé, de l'environnement et de la fonction de la donnée.

Le niveau de *connaissance* prend en compte la *structure* des informations. Nous entendons structure dans le sens le plus large possible : interrelations entre informations, historique de ces interrelations, hiérarchisation et classification,...

Le passage d'un niveau d'abstraction au suivant implique de mettre en œuvre des moyens de *transcodage* pour effectuer le passage d'un *système de représentation* à un autre.

Cette chaîne d'abstraction, qu'on retrouve déclinée en variantes [Coi94, MPCP93] peut être schématisée, à la manière de la Fig. 1.1, en un processus qui comprend :

1. L'acquisition des signaux ;
2. Un traitement univarié des signaux :

[1] Nous utiliserons la nomenclature proposée ici par la suite. Nous réservons le vocable *variable* pour une quantité quelconque, signal, donnée ou valeur générale indépendante du schéma d'aide à la décision. Nous abuserons éventuellement de ce sens, pour parler de *variable symbolique* pour des variables aux valeurs qualitatives.

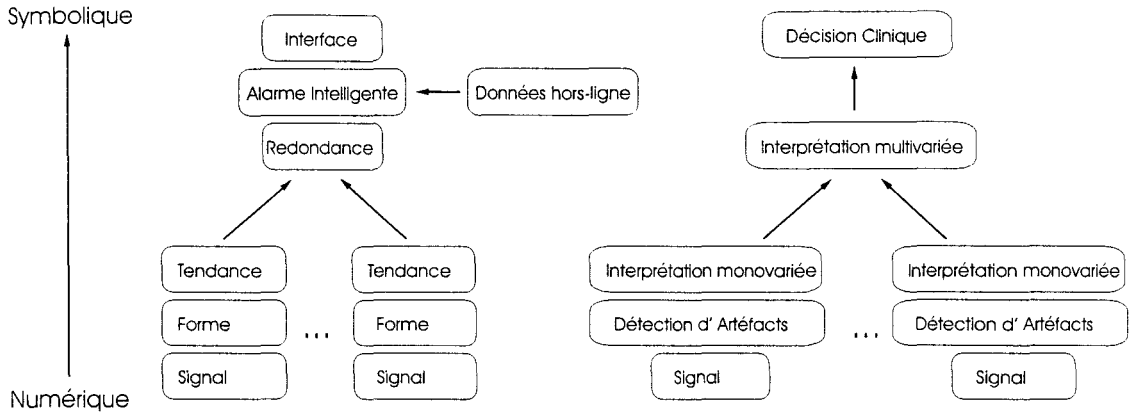



FIG. 1.1 – Schémas généraux d'un système d'aide à la décision médicale
A droite, un schéma général adapté de [Coi94] pour l'aide à la décision clinique. A gauche, le schéma de [MPCP93], qui est une instantiation du premier au problème de la génération d'alarmes intelligentes. Le patient, les capteurs et appareils de monitorage se placent dans le bas du schéma, le personnel soignant dans le haut. La connaissance du domaine n'a pas de place spécifique dans ces modèles. On se place du côté des données, pour essayer d'aller vers le domaine symbolique.

- L'extraction de caractéristiques des signaux, les données ;
- La validation de ces données par rapport aux artefacts possibles ;
- La détermination de "tendances" d'évolution.

3. Un traitement multivarié :
   - La mise en évidence des relations entre les paramètres déterminés précédemment ;
   - L'exploitation de la redondance entre paramètres ;
   - La reconnaissance de patrons-type.
4. L'intégration de la connaissance du domaine, l'évaluation de l'état du patient.
5. La présentation des différentes informations au personnel soignant.

A ce stade, on est parvenu au niveau de l'information définie ci-dessus.

Dans ces différentes phases et fonctions, la connaissance nécessaire *a priori* intervient de façon différenciée. En effet, du côté des signaux, les méthodes d'extraction de caractéristiques font surtout appel à des techniques de traitement du signal. Il en est de même pendant l'essentiel de la phase de traitement univarié. Par contre, plus l'information devient abstraite, plus l'intervention de la connaissance du domaine est explicite [VRC+00].

Dans ce schéma général, nous nous plaçons du côté des données, en essayant de repousser le plus en aval possible de la chaîne de traitement l'introduction de la connaissance d'expert. De plus, nous nous plaçons dans la chaîne de traitement *après* les appareils de monitorage usuels, donc après les fonctions d'extraction de caractéristique et de filtrage des signaux.

Ce parti-pris part d'un constat simple : le personnel soignant *se sert habituellement de ces données*, mais il n'en exploite *a priori* ni la masse ni, de façon immédiate, les interrelations.

Ce travail se situe dans le cadre du projet Aiddiag [RCJ+94].

Ce projet cherche à établir un système d'aide au diagnostic et à la décision médicale dans

l'environnement d'un service de Réanimation adulte. Outre les objectifs généraux des systèmes d'aide à la décision médicale signalés ci-dessus, le projet Aiddiag présente des objectifs de généralisation à des domaines non-médicaux, d'intégration de données hybrides (imagerie notamment), et d'implémentation abordable. Aiddiag se propose à plus long terme de devenir une station d'acquisition et de modélisation des pratiques de terrain, visant la modélisation de la connaissance du domaine.

Pour cela, la plate-forme Aiddiag remplit les fonctions de :
- centralisation des données en provenance des moniteurs usuels ;
- visualisation de l'ensemble des données ;
- acquisition des actions effectuées sur le patient [JRC+97] ;
- acquisition, par l'interface utilisateur, des modalités d'utilisation de la plate-forme, dans l'objectif de modélisation cognitive ;
- construction progressive d'une base de données d'enregistrements de terrain.

Nous nous plaçons dans le travail présent dans la perspective de l'abstraction des données vers l'information. Nous proposons une chaîne de traitement en ce sens, qui recouvre les objectifs de visualisation et d'aide à l'interprétation *a posteriori* [CCVP99].

Nous détaillons par la suite les caractéristiques du système étudié et celles des données que nous observons. Nous aborderons ensuite le problème du choix des méthodes d'analyse de ces données, avant de passer aux propositions méthodologiques et au détail du plan de ce document.

## 1.1 Le système étudié

Dans la Fig. 1.2 nous avons schématisé ce que nous essayons d'englober dans un système. A partir des signaux mesurés sur le patient, les moniteurs cardio-vasculaires, respiratoires ou autres (mesures de $CO_2$ par exemple) élaborent les données dont nous nous occupons. Cette partie de la chaîne d'acquisition [Cru92] n'est pas parfaitement identifiée, mais elle peut l'être potentiellement. Plus "loin" du côté des capteurs se trouve le patient : un système biologique ouvert. On sait *a priori* ce système non-linéaire, auto-régulé, quasi-décomposable (au moins en sous-systèmes représentatifs des grandes fonctions physiologiques, cardiaque, respiratoire,...), présentant des échelles de comportement multiples (autant temporelles que spatiales). Cette connaissance *a priori* va guider nos démarches du côté de la précaution : tout modèle extrait devra être soumis à un essai intensif de falsification.

Agissent sur le patient, parmi les paramètres potentiellement disponibles : l'aide ventilatoire et le traitement pharmaceutique. L'aide ventilatoire est observée à travers le ventilateur lui-même. L'information pharmacologique ne peut être mesurée automatiquement et dépend d'une introduction manuelle des données. Il en est de même pour d'autres paramètres de surveillance tels que les résultats d'examens faits à partir de prélèvements. Ainsi, les résultats d'analyse des gaz du sang et de la bactériologie ne sont pas monitorisés mais introduits *a posteriori* par le corps médical.

L'enregistrement des données se fait dans les conditions du terrain : aucune action n'est envi-

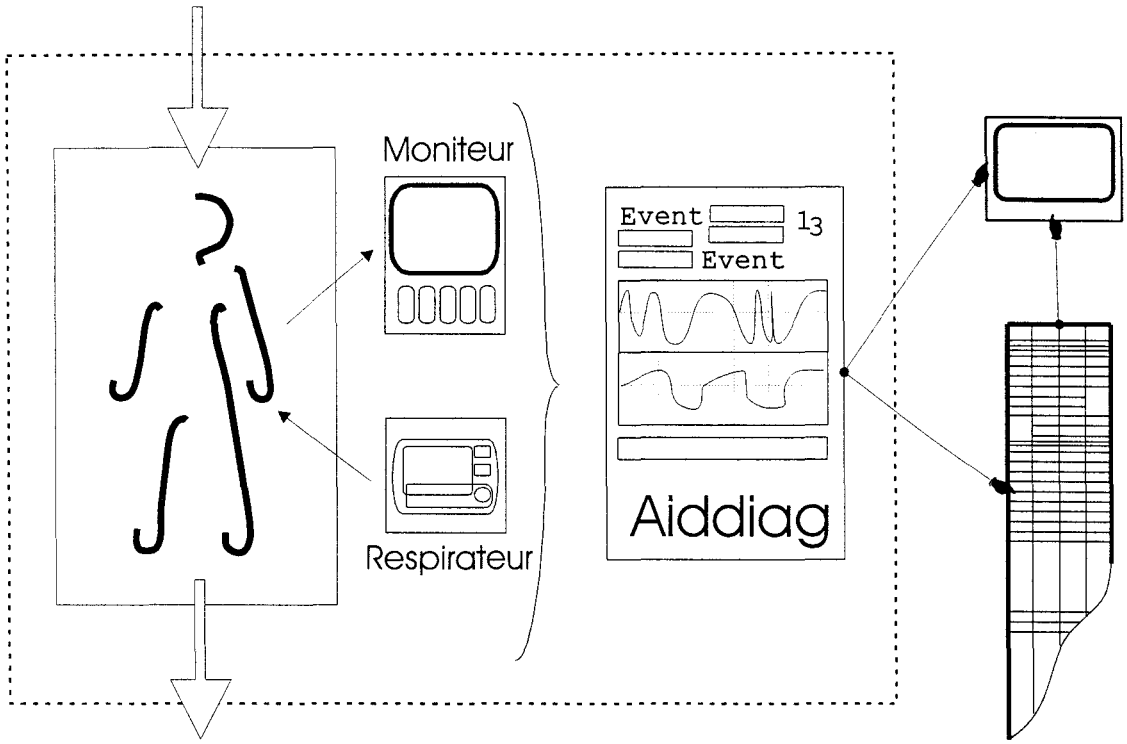



FIG. 1.2 – Le système étudié

Le patient est surveillé en conditions de terrain. Les actions que l'on enregistre effectivement ne constituent qu'une partie de l'ensemble des actions qu'encourt le patient. Les divers moniteurs et appareils thérapeutiques produisent un rapport de leur action. Celui-ci est centralisé par la station Aiddiag, qui fournit des ensembles de données pour le traitement en-ligne ou hors-ligne. De même, des moyens de visualisation (représentée par ▭) sont disponibles à plusieurs niveaux.

sageable dans un but non thérapeutique. Nous ne disposons pas de dispositif expérimental au-delà des capteurs. Les moyens classiques d'identification des paramètres d'un modèle *a priori* ne sont donc pas applicables.

Un certain nombre de "lois" ont été établies par la science médicale et par le génie biomédical pour modéliser certains comportements. L'introduction de telles connaissances dans un protocole d'analyse devra se faire avec la plus haute suspicion : nous nous trouvons devant des conditions physio-pathologiques, et non pas simplement physiologiques.

## 1.2 Des données significatives ?

Le problème de la mesure pour la modélisation dans un système aussi complexe que celui décrit ci-dessus est difficile.

Du point de vue méthodologique, se pose un problème double d'échelle de mesure : la mesure locale que l'on effectue sur le patient est-elle vraiment significative de l'état global ? Et inversement, jusqu'où l'évaluation de l'état global permet-elle de déceler des pathologies localisées ? Typique-

ment, la fièvre est une mesure de l'état global du patient, mais elle ne permet pas de spécifier la source de l'infection. De l'autre côté, la mesure par des capteurs optiques de l'oxymétrie de pouls appliqués au doigt perd de sa pertinence en cas d'anomalies circulatoires des membres périphériques.

D'un point de vue concret, seules des données fiables permettent une modélisation "correcte". Autrement, les analyses sont dénuées de sens. Avons-nous les données suffisantes pour caractériser les phénomènes que nous observons ? Sont-elles suffisamment propres par rapport aux bruits ? Permettent-elles de répondre aux questions concrètes ? La période d'échantillonnage est-elle suffisamment importante ? Répondent-elles aux hypothèses de base des méthodes d'analyse que nous essayons d'appliquer ?

Pour nous et dans un premier temps, on considérera les données comme valides en terme de période d'échantillonnage, bruit et représentativité, même si dès les premières analyses on retrouve des limitations sévères : résolution, variations extrêmes,...Les résultats fournis permettront de caractériser chaque variable en terme d'adéquation aux traitements.

### 1.2.1 Un flux de données fragmentées

La mise en situation réelle de l'outil de mesure introduit des aléas dans la présence effective des données. Certaines données ne sont pas, par nature, toujours présentes. Certaines impliquent l'installation de capteurs particuliers, telles les mesures invasives de pression sanguine ou la capnographie (mesure de la teneur en $CO_2$). Toutes dépendent de la présence effective des appareils, ce qui n'est pas acquis dans le fonctionnement normal d'un service de réanimation (typiquement, les paramètres respiratoires sont mesurés lorsque le patient est placé sous ventilation artificielle, ce qui est le cas pour bon nombre de patients en soins intensifs, mais pas tous). De plus, des défauts de positionnement des capteurs surviennent de par l'agitation du patient, et mènent à des mesures aberrantes — en particulier pour les électrodes de monitorage de l'ECG et le capteur d'oxymétrie de pouls. Lorsque celles-ci sont détectées par les prétraitements des moniteurs, les données ne sont plus transmises.

En conséquence, les tableaux bruts de données sont "troués". On appellera *absences* par la suite, des mesures ou des séquences de mesures manquantes. Face à ce problème, deux techniques sont envisageables : conservatrice et libérale.

Dans la perspective conservatrice, il n'est pas possible de connaître ni les conditions exactes d'absence d'une donnée, ni les hypothèses suffisantes pour l'expliquer. Il n'est alors pas possible d'intégrer les absences dans les procédures subséquentes. Dans cette perspective, on aura recours par exemple à l'extraction de blocs contigus de données pour pouvoir poursuivre.

Dans la perspective libérale, on peut se permettre d'interpoler les variables sur un nombre raisonnable de points manquants. Cela implique de poser des hypothèses sur l'origine des absences : problèmes de synchronie dans la chaîne d'acquisition, défauts sur les lignes de transmission, valeurs aberrantes du fait de mauvais positionnement des capteurs.[2] Ces hypothèses sont difficiles à sélec-

[2]Dans ce cas, il vaudrait mieux encore éliminer les mesures autour de l'épisode de données manquantes, ce qui est systématiquement fait dans [MSHP99].

tionner et à valider *a posteriori.*

Il restera de toute façon à intégrer plus tard le problème des données incomplètes dans les outils de traitement définitifs. Le fait même de l'absence d'un paramètre peut être riche d'information, mais on peut supposer que cette information ne se situe pas sur le même plan que le résultat d'un traitement de type statistique.

En définitive, nous choisirons notre perspective plutôt du côté conservateur. Lorsque l'absence de données rendra difficile certaines opérations, nous aurons recours, de façon opportuniste, à des interpolations. Plus précisément, nous essaierons de remplacer les valeurs manquantes par des valeurs neutres pour chaque traitement, si cela nous permet de gagner en précision ; nous laisserons tomber les points manquants lorsqu'aucune hypothèse ne semblera justifiée ou justifiable pour l'interpolation ou la suppression.

### 1.2.2 Des données difficiles à exploiter

Même au sein de blocs contigus, les données enregistrées sur le terrain présentent des caractéristiques particulières qui rendent le système étudié relativement inaccessible à la modélisation. Les données sont en effet :

#### Dépendantes du temps

On enregistre les données à intervalles réguliers dans le temps ; c'est leur évolution autant que leur valeur qui sont importantes pour l'interprétation.

Il est en effet impossible de se rapporter à un état de référence dans lequel le patient peut être considéré normal. D'un côté, cet état non pathologique n'est pas défini avec précision : il n'a pas été mesuré pour le patient particulier, mais correspond uniquement à un état physiologique normal général. D'un autre côté, le patient pathologique n'a pas d'état de normalité défini, et son état de référence reste fonction de la pathologie et de la réaction (physiologique et thérapeutique) à la pathologie.

En conséquence, les informations "lues" par les médecins sur les données sont de nature relative et non pas absolue. Elles sont relatives à la pathologie, à laquelle dans l'immédiat nous n'avons pas accès. Elles sont relatives à l'état actuel du patient au cours de l'évolution de la pathologie en présence de la thérapeutique. Ce état est défini par l'évolution pathologique et thérapeutique depuis le début du séjour.

#### Non stationnaires

L'enregistrement se fait dans des conditions de terrain. Il n'est pas possible de forcer le système dans un état déterminé.

Par rapport à des modèles dynamiques préétablis, la différenciation entre transitoires et périodes *stationnaires* est difficile à réaliser — au moins dans les sens les plus immédiats de ces termes :

stationnarité au sens de la variance locale, ou de la corrélation entre variables, ou des avances et retards des paramètres, ou encore des réponses de chaque paramètre aux variations des autres.

Le système ne permet pas l'expérimentation, donc n'est pas *identifiable* suivant les méthodes classiques en automatique.

### Échantillonnées

La technique impose (et les critères thérapeutiques permettent) un échantillonnage avec une période de l'ordre de la dizaine de secondes. Les données sur lesquelles nous avons expérimenté ont été échantillonnées à 5s de période la plupart du temps ; certains enregistrements ont été réalisés à 15s. Cette période semble suffisante pour la détection d'anomalies au moins respiratoires et pour le déclenchement d'alarmes : le temps de réponse admis pour une alarme vitale est de l'ordre de la minute, avec des variations suivant le paramètre surveillé [KMSK95]. Il est communément admis que l'apparition de lésions cérébrales irréversibles en cas d'arrêt cardio-circulatoire surviennent à partir d'un délai d'hypoxie de l'ordre de trois minutes.

### Quantifiées

La résolution fournie par les capteurs puis par les appareils de mesure peut être faible : de l'ordre du pourcent dans les mesures les plus précises. Elle peut atteindre 20% dans les cas extrêmes, par exemple, les paramètres de réglage de $FIO_2$ (la fraction d'oxygène inspiré), de $PEEP$ (la pression en fin d'expiration), mais aussi, à certaines périodes, la valeur de $SpO_2$ (l'oxymétrie de pouls). L'annexe A résume ces caractéristiques.

### Prétraitées

Les données ont déjà subi un traitement dans les appareils de mesure : il ne s'agit pas de signaux mais plutôt de paramètres caractérisant un modèle implicite du signal. Ces traitements risquent d'être détectés fortuitement par les méthodes d'analyse subséquente. Toutefois, il n'est pas trivial d'identifier les filtres utilisés et les traitements subis. Ceci pour des questions d'opacité matérielle, mais aussi du secret industriel. Typiquement, les traitements consistent en la mesure, sur des signaux quasi-périodiques, de valeurs de maxima, minima, fréquence, moyenne, en plus d'autres paramètres typiques de certains signaux. En annexe B nous détaillons ce passage des signaux aux données.

Ce prétraitement introduit une difficulté supplémentaire : il n'est pas le même pour tous les appareils ni même pour tous les paramètres, ce qui rend les données hétérogènes dans un sens assez délicat à intégrer dans les méthodes de modélisation. Par exemple, du fait de filtrages différents et capteurs différents, la relation *a priori* évidente $f_R \cdot V_T = V_E$ (*c-à-d.* la fréquence respiratoire en $min^{-1}$ multipliée par le volume expiré en $l$ est égale à la ventilation-minute en $l \cdot min^{-1}$) n'est pas toujours vérifiée.

**Non synchronisées**

L'étude que nous avons menée du synchronisme des paramètres montre que les séries temporelles ne sont pas synchrones. Ceci veut dire que les évolutions conjointes de plusieurs paramètres, lorsqu'elles existent, se font avec des décalages temporels.

En plus du manque de synchronisme, les divers paramètres n'ont pas des temps de réponse égaux.

Par exemple, il est facile de repérer qualitativement sur les séries chronologiques les moments où des *aspirations* ont eu lieu. Cette pratique thérapeutique consiste à dégager les conduits trachéaux de mucosités qui ont pu s'y accumuler. Il s'agit d'une opération régulière sur les patients ventilés artificiellement.

Sur les séries temporelles et sur les distributions, l'aspiration induit des valeurs extrêmes, tant vers les minima que vers les maxima. Le patron observé est celui d'une chute (*resp.* remontée) brutale suivie d'une remontée (*resp.* chute) tout aussi brutale. La régularité assez claire de ce phénomène (toutes les 2h à une dizaine de minutes près) permet de distinguer ce patron d'autres perturbations éventuelles de même magnitude.

Ces perturbations étant de très forte magnitude, elles contribuent très fortement dans une approche linéaire (du fait de la notion de moindres carrés). Si on était capable d'isoler les aspirations des variations intrinsèques, alors les autocorrélations permettraient d'estimer les temps de réponse respectifs de chaque paramètre, pour le retour à des valeurs de référence repérées avant l'aspiration.

D'un autre côté, la mise en correspondance de l'enregistrement des événements ayant eu lieu dans l'entourage du patient, avec l'enregistrement des données fournit un moyen de repérer des perturbations extérieures du même type que l'aspiration.

Disposant de ce flux de données et d'une capacité de calcul déterminée, on cherche à élucider un certain nombre de questions.

## 1.3 Les problèmes posés

L'objet ultime d'un travail sur les méthodes d'analyse des données applicables à l'aide au diagnostic en réanimation serait de pourvoir une boîte à outils pour l'analyse la plus automatisée possible du flux de données. Un tel module ne peut être utilisable que s'il intègre explicitement ses limitations et ses hypothèses d'action. C'est pourquoi une étape d'analyse fine des limites d'application des approches de modélisation existantes est nécessaire.

Cette analyse se fait suivant une approche double : d'un côté, à partir des méthodes, par l'expérimentation sur les données et la recherche de rapprochements physiologiques des résultats obtenus ; d'un autre côté, à partir de préoccupations concrètes de physiologie, par la recherche de moyens de validation de modèles à partir des données.

Trois problèmes fondamentaux se posent de façon générale, du point de vue de l'analyse de

données :

**Le problème de visualisation** Il s'agit de trouver une représentation lisible d'un ensemble de données. Cette représentation doit mettre en évidence les relations fondamentales entre les données et permettre une interprétation aisée de ces relations.

En analyse de données classique, la visualisation passe essentiellement par la réduction de la dimension de l'espace de représentation. Elle comprend l'analyse factorielle et ses proches (l'analyse en composantes principales, l'analyse canonique). Il s'agit de repérer les combinaisons linéaires de variables qui apportent le plus d'informations à l'ensemble des mesures, afin d'utiliser le moins de dimensions possibles dans la représentation des données.

**Le problème de structuration** Il s'agit de pouvoir associer une mesure à une classe (ou un sous-ensemble d'appartenance, dans une approche possibiliste) sans que celle-ci soit définie *a priori*. On cherche ainsi à faire apparaître une structure du phénomène observé. Les classes sont établies par la méthode elle-même, c'est pourquoi on parle de classification *automatique* ou *non supervisée*.

En terme d'analyse de données classique, l'analyse discriminante et les méthodes de classification linéaire permettent de répartir en classes des mesures qui sont "proches" — dans un sens défini par la méthode en question. Cela est réalisé en pratique par la mesure de la corrélation entre les données, en tant que mesure de proximité.

D'autres mesures de proximité sont utilisées dans divers systèmes : mesures non-linéaires à travers des transformations non-linéaires [SSM96], mesures de proximité euclidienne dans l'espace des phases [KMK92], mesures de proximité des distributions (entropies, $\chi^2$), distance de Hausdorff,...

**Le problème d'explication** Il s'agit ici de pouvoir déterminer — et donc prédire — la valeur d'une ou plusieurs données (dites à expliquer) à partir d'un ensemble d'autres données (dites explicatives).

Les méthodes employées sont soit paramétriques (on possède un modèle sous-jacent du système) soit non paramétriques (on parlera alors d'apprentissage).

La méthode linéaire par excellence est la régression (*cf.* §2.2), qui sous sa forme la plus élémentaire est la recherche d'une relation linéaire entre deux variables.

D'autres méthodes peuvent être mises en avant, que l'on peut classifier suivant le système de représentation qu'elles adoptent : formules de régression, formules logiques d'ordre zéro, logique d'ordre un,...

Par exemple, basées sur la théorie de l'information, les méthodes d'Induction par Arbres de Décision (*cf.* §5) permettent d'établir les modalités de la donnée à expliquer à partir des modalités des données explicatives. Lorsque la donnée à expliquer est une classe (dans le cas où l'on puisse en déterminer une pour chaque mesure), l'Induction par Arbres de Décision peut être considérée comme une méthode de classification *supervisée*.

Dans ce travail, nous allons nous occuper des problématiques de l'aide à la décision, suivant ces trois axes. Du point de vue de la visualisation, nous proposons des moyens d'introduire le contexte *c-à-d.* le temps et l'"environnement" déterminé par l'ensemble des variables. Du point de vue de la structuration, nous proposons l'application de techniques d'apprentissage pour une abstraction tenant compte des propriétés des flux de données multivariés dans leur ensemble. Du point de vue de l'explication nous tâcherons de remplir les objectifs d'identification de séquences dans le déroulement d'un séjour en réanimation ; par ailleurs, nous utiliserons l'asymétrie de notre ensemble de données — nous disposons de données assimilables à des variables de sortie — pour proposer des modèles explicatifs.

Pour ces objectifs, nous avons dû proposer des méthodes non traditionnelles. En effet, les limites majeures des techniques classiques ont été assez vite mises en relief.

## 1.4 L'Analyse de données classique et ses limites

Les méthodes classiques d'analyse de données [Cib87, RR93, Sap90] permettent de s'attaquer aux trois problèmes fondamentaux, mais elles reposent sur l'arsenal mathématique de l'algèbre linéaire et sont par conséquent soumises à certaines hypothèses de base.

### 1.4.1 Hypothèses de base

Parmi ces hypothèses, une étude préliminaire [Cal96] a permis d'en dégager trois qui sont limitantes pour l'analyse de données physiologiques :

**Linéarité** Les méthodes classiques sont linéaires : la mesure de "proximité" entre des paramètres ou entre des mesures se fait par un calcul de corrélation. Celle-ci est une mesure de distance "droite", qui suppose une notion d'indépendance entre les mesures et en plus l'équivalence de toutes les mesures, indépendamment de leur unité (on peut corriger ceci en normalisant les données par rapport à leur moyenne, leur écart-type,...). Du fait de cette mesure de distance, les méthodes classiques ne peuvent détecter que des relations linéaires entre les données. Pour détendre cette limitation, on peut avoir recours à des transformations des variables (logarithmique, exponentielle,...), qui doivent être réalisées à partir de critères propres au problème et comportent donc une dose certaine de subjectivité, d'arbitraire.

**Staticité** Les méthodes classiques — et la plupart des méthodes issues des travaux statistiques d'analyse des données — sont éminemment statiques. Cela veut dire que vis-à-vis du calcul, les mesures sont interchangeables. Ainsi, le caractère temporel des phénomènes échappe intrinsèquement à ces méthodes.

Si l'on veut faire apparaître un caractère dynamique dans ces méthodes, il faut introduire explicitement des variables retardées — *c-à-d.* les mesures précédentes d'un paramètre. Il se pose

alors le problème de déterminer l'"ordre" du système, c'est-à-dire le temps dont il faut effectivement tenir compte pour observer le phénomène qui intéresse l'analyste.

Une autre possibilité est d'introduire des variables supplémentaires rendant compte de la dynamique de chacune ou de l'ensemble des variables en question. Les techniques de filtrage sont délicates à mettre en œuvre dans ce contexte : elles risquent d'introduire des relations dans les données qui ne sont que d'ordre technique et qui peuvent passer pour des relations intrinsèques au système.

**Stationnarité** Tout essai d'analyse à partir d'un ensemble de données suppose que celles-ci représentent le même phénomène — au système de représentation *c-à-d.* à la classe de modèles près. Or, dans son évolution au cours du temps, un système biologique peut changer d'état d'une façon telle que non seulement les paramètres physiologiques mais aussi les interrelations entre les systèmes physiologiques changent. Pour pouvoir établir une analyse de façon fiable, il faut pouvoir détecter ces changements profonds, et appliquer les méthodes d'analyse uniquement sur les durées où l'on peut considérer que le système observé reste inchangé.

Les données physiologiques mesurées au moyen des moniteurs couramment utilisés ne peuvent satisfaire ces hypothèses que dans des situations hélas rares.

Illustrons sur un exemple les limites des approches classiques.

### 1.4.2 Exemple

Nous disposons d'un fichier dans lequel l'outil d'acquisition Aiddiag a permis d'enregistrer, sur 10 heures 40 minutes et toutes les 15 secondes les mesures de treize paramètres fournis par le moniteur cardiaque (SpaceLabs) et le ventilateur (Dräger Evita-2). Parmi les nombreuses informations recueillies, notons que le patient est placé sous ventilation contrôlée, avec une consigne $f_R = 15$ tout au long du fichier.

Un schéma d'analyse élémentaire passe par une visualisation afin de repérer des relations entre des groupes de variables. Ces relations seront vues ensuite une à une dans le détail, pour essayer enfin de trouver un modèle mathématique qui renfermerait l'essentiel du comportement des données observées.

Une Analyse en Composantes Principales (ACP [Hot33]) est d'abord appliquée (*cf.* Fig. 1.3). Elle fournit des résultats à partir des données prises dans leur ensemble et de façon statique : aucune évolution n'est ici repérée, et les relations observées sont valides pour les presque onze heures de mesures.

En étudiant plus finement les relations, on se rend compte que l'hypothèse de linéarité fait défaut : la corrélation linéaire entre $V_E$ et $F_c$ n'est qu'une approximation de la relation réelle qui lierait ces paramètres. La distribution conjointe des mesures de ces deux paramètres (Fig. 1.4) montre deux modes nettement distincts. On retrouve ces deux modes dans la distribution conjointe de $V_E$ et $f_R$ (Fig. 1.5) où l'indépendance linéaire des deux paramètres est de même mise en défaut.

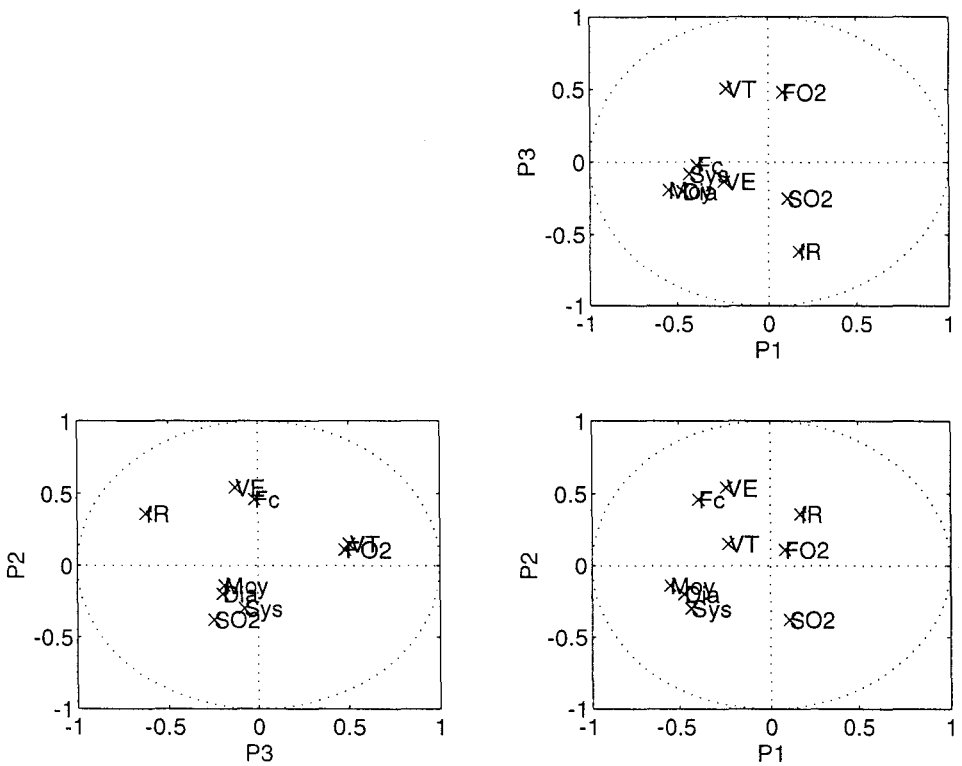



FIG. 1.3 – Analyse en composantes principales

On n'a représenté, pour des soucis de lisibilité, que neuf des treize paramètres sur les trois premiers axes principaux P1, P2, P3. L'ACP montre la corrélation entre les pressions artérielles Moy, Dia, Sys, une forte corrélation entre $V_E$ et $F_c$ et l'indépendance linéaire de $V_E$ et $f_R$. On ne peut tirer beaucoup d'autres renseignements, vu la distance des points au cercle des corrélations.

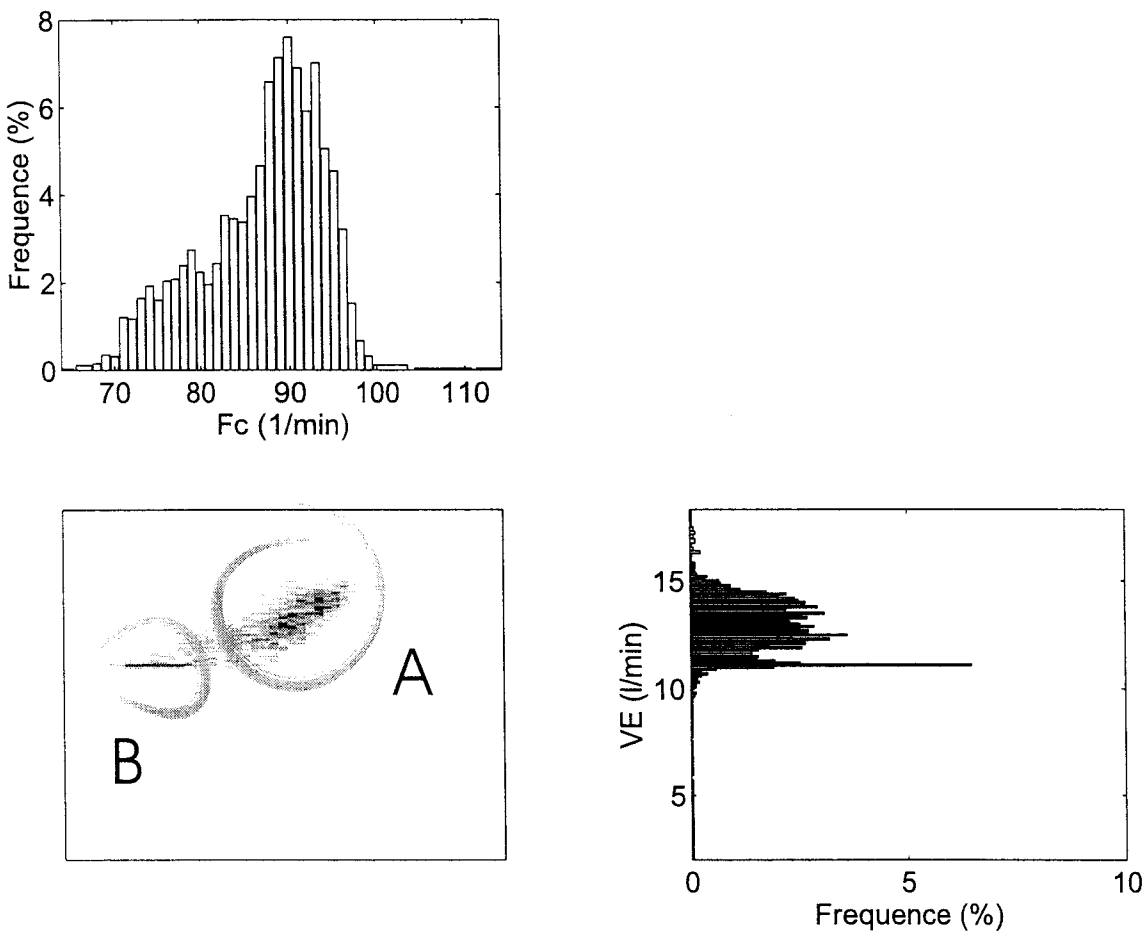



FIG. 1.4 – Distribution conjointe de $V_E$ et $F_c$
Le niveau de gris indique la fréquence dans la distribution : blanc pour nulle et noir pour maximale. Pour un mode (A) il existe effectivement une relation linéaire, mais un autre mode (B) met en évidence l'indépendance des deux paramètres pour une valeur fixe de $V_E$ .

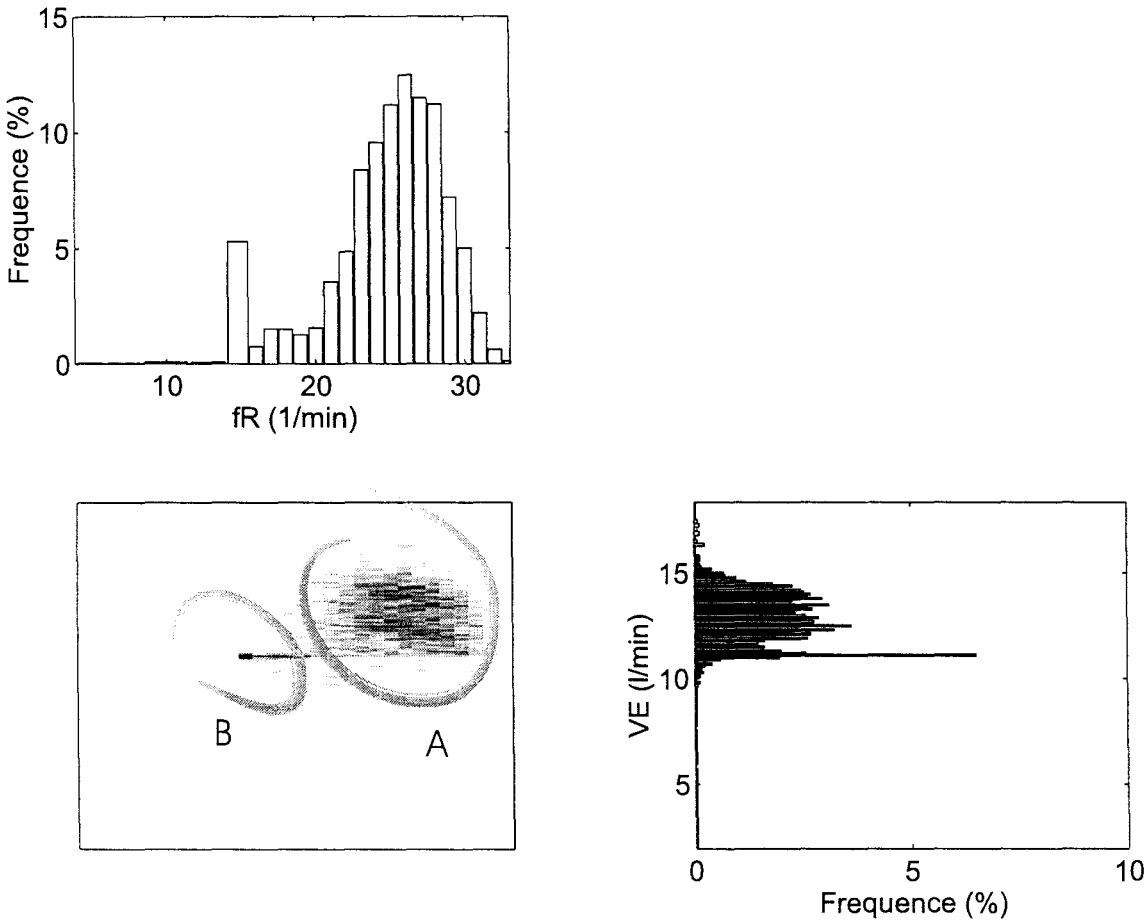



FIG. 1.5 – Distribution conjointe de $V_E$ et $f_R$
Un mode (B) est visible à $V_E = 11.1\, l.min^{-1}$ et $f_R = 15\, min^{-1}$ , un autre (A) est défini par l'indépendance des deux paramètres.

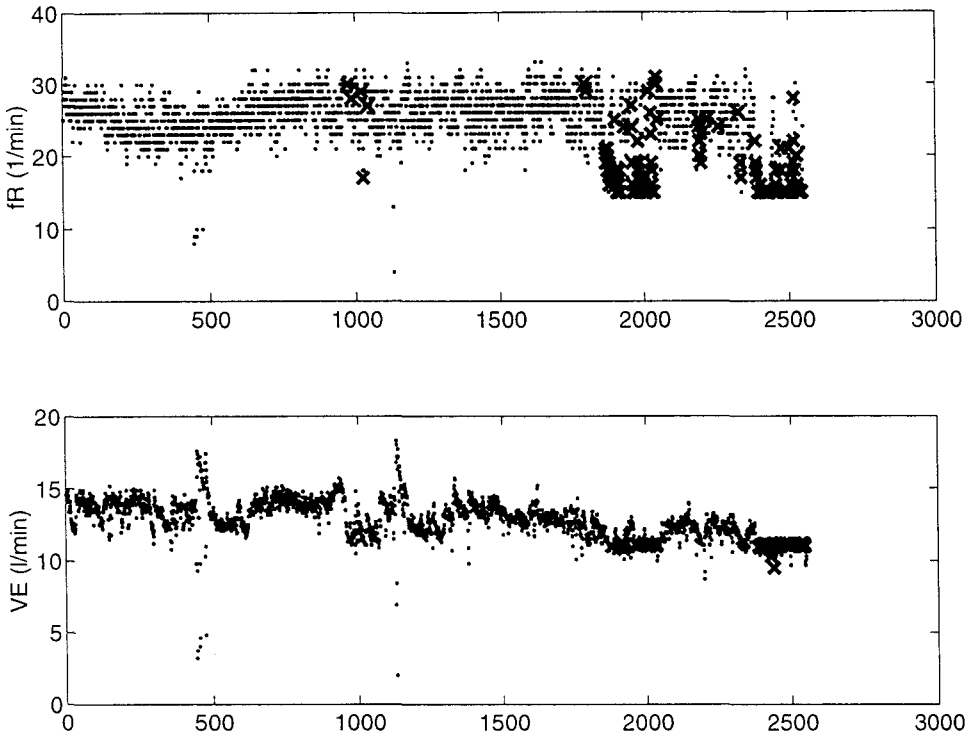



FIG. 1.6 – Évolution temporelle de $V_E$ et $f_R$
En abscisse le temps (1unité$\sim$ 15$s$), en ordonnée la mesure. Les croix indiquent le mode (B), les points le mode (A). Le mode (B) s'établit uniquement vers la fin du fichier. Pourtant, on ne peut pas dire que l'un des modes supplante l'autre à partir d'un moment donné. On peut par ailleurs remarquer les deux perturbations, aux abscisses $\sim$ 500 et $\sim$ 1100, correspondant à des aspirations.

Deux modes coexistent donc : un mode (A) où $V_E$ et $F_c$ sont corrélés et indépendants de $f_R$— ce qu'on repère par l'ACP — et un mode (B) où $V_E$ et $f_R$ oscillent autour d'une valeur fixe indépendamment de $F_c$. Pour déterminer le sens de cette coexistence de deux modes, il faut mettre en défaut l'hypothèse de staticité. En effet, lorsqu'on repère dans le temps chacun des deux modes (Fig. 1.6), il est clair que les deux modes sont localisés dans le temps.

Ainsi, nous mettons en évidence les limitations des techniques classiques linéaires : le produit scalaire comme base de la mesure de distance, l'interchangeabilité des mesures, la staticité des modèles — l'analyse fine que nous avons menée fait apparaître une interprétation *historique* mais non pas *dynamique*. De plus, nous n'avons pu intégrer que les données numériques. Les caractéristiques de celles-ci rendent suspecte l'application de méthodes issues de l'algèbre linéaire, en particulier la forte quantification.

## 1.5 Propositions méthodologiques

Pour pallier ces limitations majeures, un certain nombre de méthodes nouvelles d'analyse ont été développées au sein des communautés de l'apprentissage automatique, en particulier à partir des travaux sur la Théorie de l'Information [Cul72].

Deux idées fondamentales sont mises en œuvre, du point de vue technique :

1. Tout d'abord, on traite les données en tant que *distributions* et non pas en tant que valeurs effectives. Ceci permet d'introduire des approches probabilistes, plus à même de décrire des phénomènes de nature complexe. On relâche ainsi l'hypothèse de linéarité pour faire apparaître des relations plus subtiles entre les données. Par contre, on introduit une hypothèse supplémentaire : il faut un nombre appréciable de mesures pour pouvoir estimer une distribution et appliquer les outils des probabilités.
2. Ensuite, on ne cherche plus à décrire par la corrélation la relation entre des données ; on utilisera une mesure d'*entropie* pour déterminer l'écart entre deux distributions. Cette mesure peut être interprétée comme une quantité d'information contenue dans une donnée ou échangée entre plusieurs données.

Si ces deux principes conduisent à tenir compte de relations complexes entre les données physiologiques, ils restent néanmoins soumis aux hypothèses de staticité et de stationnarité.

Pour relâcher ces contraintes, on se place sous l'hypothèse de stationnarité *locale*. Cela veut dire que pour des "courtes" durées, on supposera le phénomène stationnaire et on pourra appliquer les méthodes d'analyse classique ou informationnelle. Pour cela, il faut découper l'observation en tranches de longueur judicieusement choisie. Ce découpage permet de considérer le phénomène physiologique comme une succession d'états stationnaires et d'états transitoires entre ceux-ci. Il reste, certes, à définir ce qu'on entend par courte durée. Cette durée doit être mesurée à partir des données, au vu de leur comportement temporel "typique". Il s'agit d'une grandeur en relation directe avec l'ordre du système considéré.

Pour traiter la question de la staticité, on utilisera des méthodes de prétraitement qui feront apparaître l'aspect dynamique (et non plus uniquement historique comme ci-dessus).

Du point de vue de l'utilisation finale, nous veillerons à employer des méthodes qui permettent la validation à chaque étape. Nous insisterons donc sur l'*intelligibilité* des modèles construits, en terme d'explicitation des résultats donc de simplicité volontaire des modèles. De même, des techniques de *visualisation* permettront de mettre en évidence, par l'exploitation de la redondance [KK93], les symétries présentes dans les données traitées.

### 1.5.1 Chaîne de traitement : des données aux modèles symboliques

Concrètement, nous proposons la chaîne de traitement suivante (*cf.* Fig. 1.7) :

1. Un traitement univarié pour l'extraction de séries chronologiques symboliques. Celui-ci passe par :

(a) La détermination d'un temps caractéristique séparant le court terme du long terme ;

(b) L'extraction de la tendance locale de chaque paramètre ;

(c) La conversion numérique-symbolique qui tiendra compte des notions de *tendance*, de *stabilité* et de *normalité* des valeurs ;

2. L'application de méthodes d'apprentissage automatique, afin de construire des modèles locaux ;
3. L'étude de la succession des modèles construits pour permettre la détermination des zones de stabilité de la configuration du rapport entre les paramètres.

Ces zones de stabilité, qui représentent le fait que les liaisons entre les paramètres restent semblables pendant une période de temps, feront l'objet d'une évaluation, axée sur :

**La connaissance du domaine.** Suivant l'interprétation qui peut être faite par l'expert des modèles construits par rapport à la physio-pathologie, et

**Les caractéristiques des données.** Pour affirmer cette notion de stabilité de rapports entre variables, d'autres techniques de mesure de dépendance devront exhiber des comportements analogues à ceux repérés par la méthodologie. En particulier, des zones de stabilité détectées par notre méthodologie devront correspondre à celles repérées par l'utilisation d'autres méthodes basées sur des hypothèses voisines des nôtres.

### 1.5.2 Données manquantes, visualisation

La partie descriptive de ce document reprendra chacune de ces phases, en insistant au fur et à mesure sur les choix réalisés, ainsi que sur deux éléments fondamentaux présents dans le domaine d'application :

1. La question de la prise en compte des données manquantes dans chaque traitement ;
2. Les possibilités de visualisation que génère chaque étape de la chaîne de traitement.

En effet, les algorithmes disponibles dans les outils couramment utilisés (*ex.* matlab) ou dans les bibliothèques méthodologiques construites par diverses recherches (*ex.* netlib, statlib) ne sont pas, la plupart du temps, prévus pour opérer sur des données manquantes. Des outils statistiques puissants comme sas offrent ces techniques, mais il s'agit de systèmes fermés, ne pouvant être intégrés en ligne à des applications autres.

Nous insisterons particulièrement sur les modes de représentation des diverses quantités calculées au fur et à mesure. La visualisation est en effet au degré zéro de l'aide à la décision [Tuf97]. Elle devra proposer des représentations synthétiques et intelligibles. Moles [Mol91] développe ces principes comme moyens d'analyse sur des phénomènes imprécis ; Tufte [PT94] les applique à la conception de la feuille de soins hospitaliers.

Nous chercherons à implémenter ces techniques de visualisation pour l'aide à l'interprétation en ligne dans la plate-forme d'aide au diagnostic Aiddiag [RCJ+94].

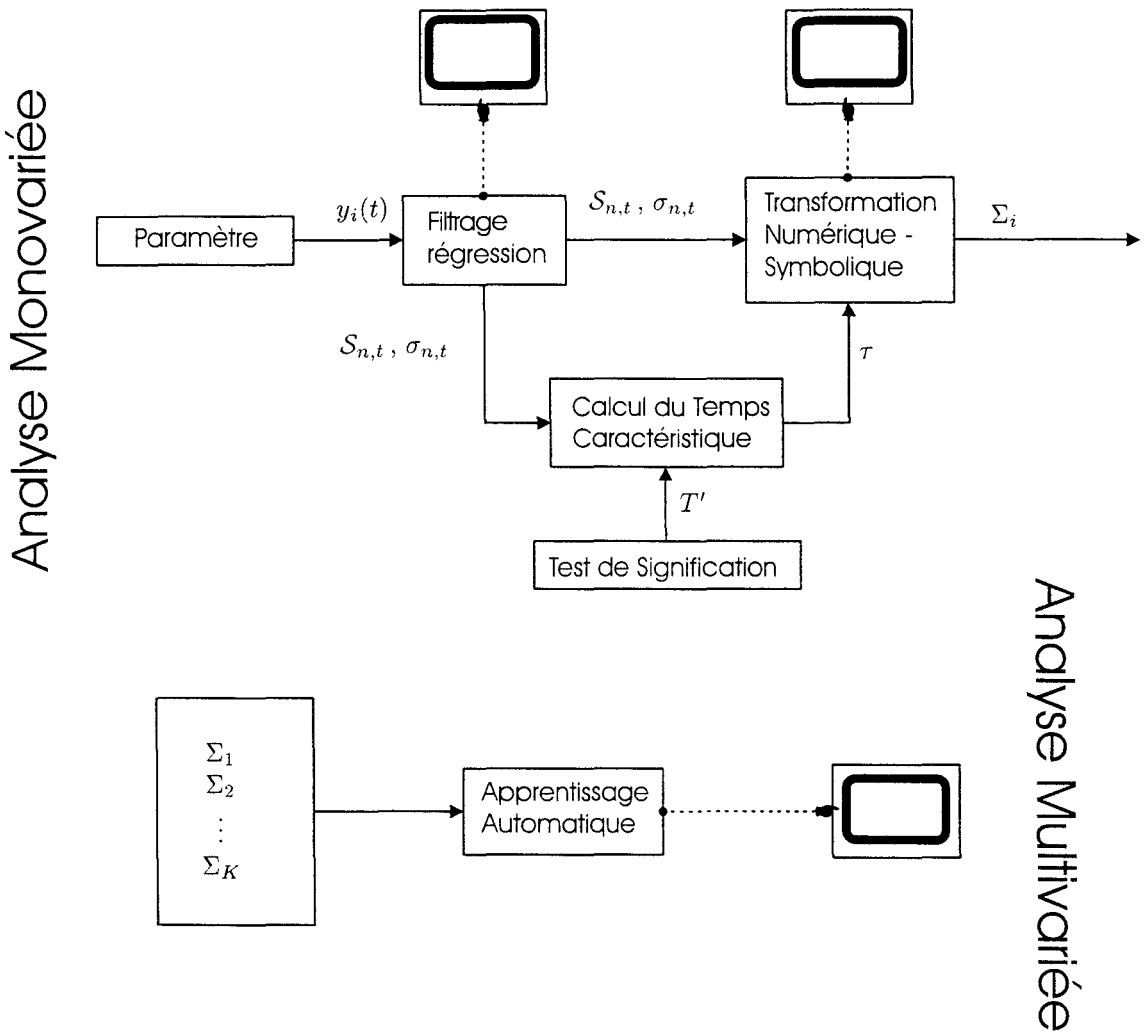



FIG. 1.7 – La chaîne de traitement proposée

En haut, la partie univariée du traitement. Elle part des paramètres physiologiques disponibles sous forme de données de monitorage. Elle passe par un filtrage permettant l'extraction de tendances, ce qui nécessite la détermination d'un temps (ou échelle) caractéristique. Les données filtrées font l'objet d'une conversion numérique-symbolique.

Ces symboles sont injectés dans un système d'apprentissage automatique, qui en retire une représentation abstraite : un *modèle*. La succession de ces modèles dans le temps permet de suivre l'évolution du sous-système sous-jacent.

Les figures signalées par ▭ sont les points où la visualisation intervient de façon directe vers une aide au diagnostic.

Les traitements réalisés aux étapes de filtrage régression, calcul du temps caractéristique, transformation numérique-symbolique ne sont pas spécifiques à la méthodologie exposée, et peuvent être potentiellement transposées dans d'autres domaines.

Les étiquettes des liens correspondent aux notations que nous utiliserons tout au long du document, et reprises en Annexe F.

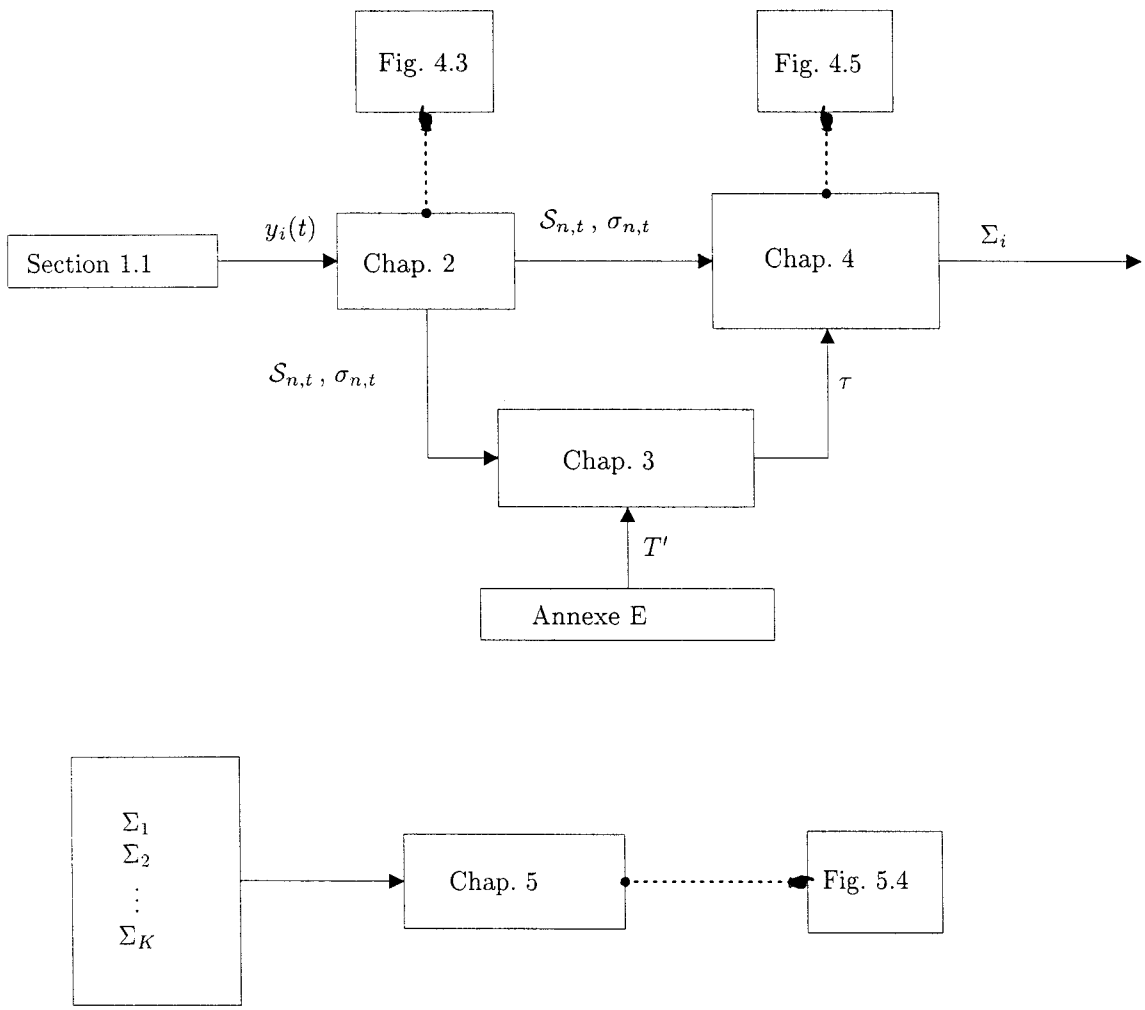



FIG. 1.8 – Plan du document : Correspondance entre les éléments de la chaîne de traitement Fig. 1.7 et les sections du document

## 1.6 Plan

Nous allons exposer notre méthodologie suivant le schéma Fig. 1.8 : nous aborderons les questions relatives à l'extraction de la tendance au Chap.2 ; à partir de celle-ci nous expliciterons les moyens (Chap.3) d'effectuer une transformation numérique-symbolique des données qui s'y prêtent au Chap.4 ; nous passerons en revue la famille de méthodes d'apprentissage que nous utiliserons avant de les appliquer au Chap.5. Nous discuterons des choix et des alternatives par la suite au Chap.6, avant de conclure sur les perspectives du travail. Les hypothèses d'application, ajustements effectués et moyens de validation et de visualisation seront abordés à chacune des étapes. Certains détails techniques sont rapportés en annexe, afin d'éviter de rompre la fluidité du discours.

...
*El tiempo, ya que al tiempo y al destino*
*se parecen los dos : la imponderable*
*sombra diurna y el curso irrevocable*
*del agua que prosigue su camino.*
...

Jorge Luis Borges, *El Reloj de Arena*

# 2

# Extraction de la tendance locale

Où des méthodes efficaces de calcul de la tendance sont proposées, qui dépendent de l'échelle à laquelle on veut observer les données.

## Sommaire

Le problème d'estimation de la tendance d'un paramètre physiologique en vue de l'intégration de notions temporelles dans les systèmes de raisonnement automatique a été envisagé suivant plusieurs points de vue dans la littérature.

Le problème général de la détection de tendance a été décliné suivant deux approches générales [Ste96] :

1. Considérer la tendance comme une qualité de forme, en la caractérisant par rapport à des patrons de référence [HK96]. La connaissance du domaine est introduite dès la phase de détection de tendance.
2. Définir la tendance comme une caractéristique extraite du signal. Cette caractéristique est soit d'ordre global, comme la forme générale des signaux par rapport à une famille de courbes [BBJ+79], soit d'ordre local, en tant que rapprochement du comportement à court terme du signal [AC90].

Nous nous plaçons ici dans la deuxième perspective, en définissant comme tendance une approximation du comportement local de chaque signal, assimilable à une dérivée. Le comportement global sera considéré comme l'évolution des valeurs et des tendances dans le temps.

Il ressort en effet des recherches appliquées aux domaines biomédicaux deux tendances essentielles : le traitement de l'information temporelle par des moyens cognitifs et l'abstraction de ces informations à partir des données. Il s'agit de problèmes séparés et pour lesquels les outils diffèrent [MHPP97].

Dans le cas de l'exploitation des données de soins intensifs, les données sont caractérisées par leur fréquence élevée [HME+97]. Des méthodes de validation et de filtrage doivent être mises au point, qui exploitent la haute fréquence.

Nous rentrons dans cette problématique en reportant les problèmes liés à l'introduction de la connaissance, à la méthode d'apprentissage utilisée par la suite. En effet, nous essaierons d'introduire la connaissance d'expert uniquement après conversion en symboles de nos données. De plus, nous nous plaçons dans une perspective basée sur les données, de sorte que nous essayons de produire des modèles à partir des données. L'intégration de ces modèles du comportement des données avec la connaissance du domaine ne rentre pas dans l'immédiat dans nos préoccupations, et ne sera abordée qu'en terme de travail futur.

Les approches basées sur les abstractions temporelles (*ex.* [Ker96]) essaient de générer dès la phase d'extraction de la tendance des épisodes temporels (représentables sous forme d'intervalles [HM99]). Dans l'approche présente, la notion de tendance permet de définir des valeurs *locales* qui pourront subir un traitement d'inférence temporelle horizontale par la suite, après traitement multivarié. C'est-à-dire que l'agrégation d'événements temporels successifs en périodes où le même comportement est observé se fera non pas à partir de chaque flux de données comme pour [SC98], mais à partir des évolutions conjointes des données.

De même, nous nous éloignons des propositions d'utilisation de patrons de tendance (*trend templates* [HK96]), au sens où ces approches intègrent la connaissance d'expert dès la phase d'extraction de tendance. Nous nous plaçons dans le cadre de modèles de tendance *isocybernétiques* [AC90], *c-*

*à-d.* sans modèle de référence pour la structure du processus : on modélise uniquement les mesures.

Dans la suite de ce chapitre, les traitements sont effectués *variable par variable*, aucune notion de lien entre variables n'est évoquée, et, lorsque cela n'est pas explicité, le terme "donnée" implique un seul flot de mesures (comportant ou non des absences) pour un paramètre physiologique.

## 2.1 Des procédures efficaces pour le calcul de la tendance

On cherche à estimer la droite qui passe au mieux par un ensemble de points espacés dans le temps.

Un critère raisonnable est celui de minimisation de l'erreur quadratique moyenne d'approximation. Raisonnable car :

- menant à des solutions analytiques sur bon nombre de cas,
- compatible avec la notion intuitive de "meilleure approximation",
- ne nécessitant pas en général de temps de calcul exorbitants, même dans le cas d'applications itératives.

Ces arguments sont à contraster avec ceux qu'on peut avancer pour les critères de moindre médiane quadratique des erreurs [Zha97] : meilleure robustesse aux points extrêmes, mais pas d'expression analytique en général et temps de calcul plus importants.

Pour d'autres critères généraux comme la longueur de description minimale (MDL [BO95]), c'est l'aspect intuitif et les difficultés d'implémentation d'un codage universel pour chaque application qui rendent son utilisation problématique.

Traduire ce critère dans notre cas d'estimation d'une tendance locale, revient à résoudre un problème de régression linéaire.

## 2.2 Régression Linéaire

La régression linéaire est la résolution au sens des moindres carrés (*c-à-d.* la minimisation du terme gauche) d'un système d'équations linéaire par rapport aux inconnues [Sch96, pp.208].

Par exemple, une écriture très générale d'un tel système est :

$$\begin{aligned}
\forall t \in [1 \ldots T] \qquad & \lambda_{11}\, f_1^1(x_t^{(1)}) + \cdots + \lambda_{1j}\, f_j^1(x_t^{(1)}) + \cdots + \lambda_{1n_1}\, f_{n_1}^1(x_t^{(1)}) + \cdots \\
+ \quad & \lambda_{i1}\, f_1^i(x_t^{(i)}) + \cdots + \lambda_{ij}\, f_j^i(x_t^{(i)}) + \cdots + \lambda_{in_i}\, f_{n_i}^i(x_t^{(i)}) + \cdots \\
+ \quad & \lambda_{N1}\, f_1^N(x_t^{(N)}) + \cdots + \lambda_{Nj}\, f_j^N(x_t^{(N)}) + \cdots + \lambda_{Nn_N}\, f_{n_N}^N(x_t^{(N)}) \\
= \quad & 0
\end{aligned}$$

Les $f_j^i$ sont des fonctions connues à valeurs dans $\mathbb{R}$ ; il existe un $i$ pour lequel les $\lambda_{ij}$ sont connus pour tout $j$ ; les autres $\lambda_{ij}$ sont les inconnues du problème ; les $x_t^{(i)}$ sont des valeurs connues, typiquement des mesures. L'indice $t \in [1 \ldots T]$ représente le numéro de la mesure parmi les $T$

mesures, les indices $i \in [1 \dots N]$ repèrent les $N$ variables, les indices $j \in [1 \dots n_i]$ marquent chacun des termes faisant intervenir la même variable.

Le problème posé sous cette forme est celui de la recherche d'une combinaison linéaire de fonctions (éventuellement non-linéaires) des variables, qui reproduisent au mieux les mesures.

Dans le cas classique multivarié des statistiques (*ex.* [Sap90, pp. 375–380]) $n_N = 1$, $\lambda_{N1} = -1$, les $f_j^i$ sont des fonctions identité, et on note $y_t = \lambda_{N1}\, f_1^N(x_t^{(N)})$. Cette dernière variable est l'*observation* (ou variable *dépendante*, ou *critère* ou encore variable *à expliquer*) ; les autres sont les variables *explicatives* (ou *indépendantes*, ou *prédicteurs* ou encore *régresseurs*).

Dans le cas univarié, nous avons $N = 2$, $n_N = 1$, $\lambda_{N1}$ est connu et non-nul — par conséquent on peut le considérer égal à un. On note $\forall t \quad y_t = -\lambda_{N1}\, f_1^N(x_t^{(N)})$. Alors le système devient (avec l'allègement d'un indice dans la notation) :

$$\forall i \in [1 \dots T] \quad \lambda_1 f_1(x_i) + \dots + \lambda_j f_j(x_i) + \dots + \lambda_N f_N(x_i) = y_i$$

Typiquement, les $y_i$ sont des sorties d'un système, les $x_i$ sont des entrées et les $f_j$ sont des fonctions d'une variable, indépendantes des $\lambda_j$ (*ex.* des polynômes, des fractions rationnelles, des fonctions trigonométriques, ...). Ce système d'équations peut être écrit sous forme matricielle :

$$\begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_N(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_N(x_2) \\ \vdots & \vdots & & \vdots \\ f_1(x_T) & f_2(x_T) & \cdots & f_N(x_T) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}$$

$$X\Lambda = Y$$

Cette équation est linéaire par raport au vecteur de paramètres $\Lambda = [\lambda_1 \dots \lambda_N]'$.

Lorsque la matrice $X'X$ est inversible (si elle ne l'était pas, une formulation employant une inverse généralisée pourrait être utilisée), la solution peut s'écrire :

$$\Lambda = (X'X)^{-1}X'Y$$

## 2.3 Régression en tant que filtrage

Pour des données où les $x_t$ sont régulièrement espacés, la ligne de tendance moyenne estimée par solution de l'équation matricielle de régression peut être obtenue par filtrage linéaire. Nous proposons ici le développement de ce résultat pour le cas particulier des données régulièrement échantillonnées.

De façon générale, lorsque la matrice des régresseurs $X$ est constante, il est possible de précal-

culer $X^* = (X'X)^{-1}X'$, ramenant la résolution de la régression au calcul d'un produit matriciel. Pour des données qui sont des séries chronologiques, cela revient à réaliser un filtrage linéaire.

Si les données ont été échantillonnées aux instants $\{1, 2, \ldots, n\}$[1] la matrice des variables explicatives $X$ a cette forme :

$$X = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ \vdots & \vdots \\ n & 1 \end{bmatrix}$$

donc $X^* = (X'X)^{-1}X'$ est constante et peut être précalculée pour tout $n$ :

$$X^* = \frac{6}{n^2 - n} \begin{bmatrix} \frac{2}{n+1} - 1 & \cdots & \frac{2\,i}{n+1} - 1 & \cdots & \frac{2\,n}{n+1} - 1 \\ \\ \frac{2\,n+1}{3} - 1 & \cdots & \frac{2\,n+1}{3} - i & \cdots & \frac{2\,n+1}{3} - n \end{bmatrix}$$

Alors :

$$\Lambda = \begin{bmatrix} \mathcal{S}_{n,t} \\ \mathcal{P}_{n,t} \end{bmatrix} = \begin{bmatrix} s_{n,1} & \cdots & s_{n,j} & \cdots & s_{n,n} \\ p_{n,1} & \cdots & p_{n,j} & \cdots & p_{n,n} \end{bmatrix} \begin{bmatrix} y_t \\ \vdots \\ y_{t+n-1} \end{bmatrix}$$

avec

$$s_{n,j} = 6\,\frac{2j - (n+1)}{n^3 - n}$$

$$p_{n,j} = 2\,\frac{-3j + (2n+1)}{n^2 - n}$$

Donc, les paramètres de la meilleure droite d'approximation de la tendance peuvent être obtenus — hors ligne — par convolution. La pente $\mathcal{S}_{n,t}$ (*resp.* l'ordonnée à l'origine $\mathcal{P}_{n,t}$) au temps $t$ et à l'échelle $n$ est la valeur en $t$ du produit de convolution du vecteur des données avec le vecteur $[s_{n,n} \ldots s_{n,1}]'$ (*resp.* $[p_{n,n} \ldots p_{n,1}]'$).

En ligne, la pente (*resp.* ordonnée à l'origine) peut être obtenue par produit scalaire des vecteurs $[s_{n,1} \ldots s_{n,n}]'$ (*resp.* $[p_{n,1} \ldots p_{n,n}]'$) et un tampon des $n$ derniers points de mesure $[y_{t-n+1} \ldots y_t]'$.

Il est donc possible de ramener le problème de recherche de la meilleure droite d'approximation depuis un problème de régression vers un filtrage linéaire. En ligne ou hors ligne, nous avons un moyen efficace de calculer les régressions[2] pour des données régulièrement espacées. La limitation

[1] Il est facile de généraliser ce cas à un échantillonnage de période $\rho$, avec des échantillons à $\{t, t+\rho, \ldots, t+(n-1)\rho\}$.

[2] Le temps de calcul est fonction directe de $n$, plus précisément il est "de complexité $O(n)$" dans la notation de Landau, voir [Bro89] par exemple.

majeure par rapport à un calcul de régression par inversion matricielle concerne la prise en compte de valeurs manquantes.

Avec une procédure de régression générale, la matrice $X$ peut être construite en ne tenant compte que des valeurs non manquantes et de leurs instants.[3]

La procédure d'estimation par filtrage ne permet pas de traiter des valeurs manquantes, si ce n'est à travers l'utilisation de valeurs spéciales comme *not a number* (NaN), dont la valeur est propagée au résultat de toute opération dans les systèmes implémentant les standards correspondants (*ex.* IEEE-754). Si c'est le cas, marquer les valeurs d'entrée comme des NaN résultera en des valeurs NaN pour les pentes, les erreurs d'approximation, les ordonnées à l'origine,...pour les $n$ points concernés.

Outre ce recours à un marquage des données manquantes, il est difficile d'introduire, sans pénalité en espace mémoire et en temps de calcul — *cf.* Annexe D — la prise en compte des données manquantes.

La forme caractéristique des coefficients de filtrage — essentiellement des rampes, *cf.* Fig.2.1 — permet des optimisations supplémentaires.

## 2.4 Implémentation incrémentale

Dans le but d'une implémentation de la procédure de filtrage en ligne, une forme très rapide peut être mise au point.

Le calcul de $\mathcal{S}_{n,t+1}$ peut être réalisé à partir des valeurs de $\mathcal{S}_{n,t}$, $y_{t+1}$, $y_{t-n+1}$ et $\mathbb{S}_{n,t}$ (la somme des données à l'intérieur de la fenêtre de $n$ points précédant $y_t$). Ce calcul se fait en temps *constant* par rapport à $n$, contrairement au filtrage linéaire proposé précédemment, qui est calculé en temps linéaire par rapport à $n$.

Soit

$$\mathcal{S}_{n,t} = \sum_{j=1}^{n} s_{n,j}\, y_{t-j+1} = \sum_{j=1}^{n-1} s_{n,j}\, y_{t-j+1} + s_{n,n} y_{t-n+1}$$

la valeur de la pente des droites de régression à l'instant $t$ à l'échelle $n$, et

$$\mathbb{S}_{n,t} = \sum_{j=t-n+1}^{t} y_j$$

les sommes partielles à l'intérieur de la fenêtre considérée ; on a :

[3] Qu'il s'agisse de la bonne façon d'opérer en présence de valeurs manquantes reste discutable, mais aucun ajustement de la méthode n'est nécessaire.

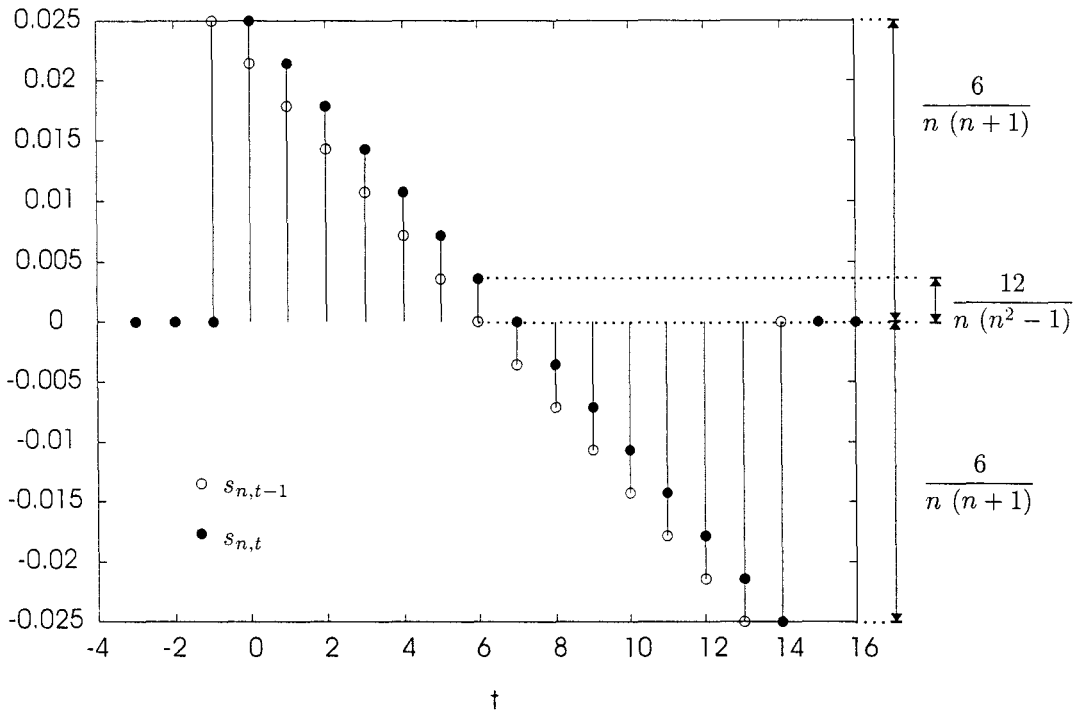



FIG. 2.1 – Coefficients de filtrage pour la tendance

En abscisse, le temps, en nombre de points de mesure, $n = 15$. La différence temporelle entre les coefficients de filtrage utilisés à l'instant $t$ et ceux utilisés à l'instant $t-1$ contient :

- Deux termes extrêmes correspondant à $y_{t+1}$ (à $t = 14$ sur la figure) et à $y_{t-n+1}$ (à $t = -1$ sur la figure), correspondant aux coefficients de valeur $\pm\frac{6}{n\,(n+1)}$.
- Des termes constants, de valeur $\frac{12}{n\,(n^2-1)}$.

Ces termes constants seront convolués avec les termes $y_{t-n+2}$ jusqu'à $y_t$. On peut donc calculer leur apport au résultat du filtrage en connaissant uniquement $\sum_{i=t-n+2}^{t} y_i$.

$$\mathcal{S}_{n,t+1} = s_{n,1} y_{t+1} + \sum_{j=2}^{n} s_{n,j} y_{t-j+2}$$

Or,

$$s_{n,j+1} = s_{n,j} + \frac{12}{n^3 - n} \qquad \forall j \in [1, n-1]$$

donc, en remplaçant ci-dessus :

$$\mathcal{S}_{n,t+1} = s_{n,1} y_{t+1} + \mathcal{S}_{n,t} - s_{n,n} y_{t-n+1} + \frac{12}{n^3 - n} \sum_{j=1}^{n-1} y_{t-j+1}$$

En développant $s_{n,n}$ et introduisant $\mathbb{S}_{n,t}$ :

$$\mathcal{S}_{n,t+1} = \mathcal{S}_{n,t} + \frac{12}{(n^2 - 1)\, n} (\mathbb{S}_{n,t} - y_{t-n+1}) - \frac{6}{n^2 + n} (y_{t+1} + y_{t-n+1})$$

Une formule équivalente peut être posée pour $\mathcal{P}_{n,t+1}$ :

$$\mathcal{P}_{n,t+1} = \mathcal{P}_{n,t} - \frac{6}{n^2 - n} (\mathbb{S}_{n,t} - y_{t-n+1}) + \frac{4}{n} y_{t+1} + \frac{2}{n} y_{t-n+1}$$

La mise à jour des sommes partielles est triviale :

$$\mathbb{S}_{n,t+1} = \mathbb{S}_{n,t} + y_{t+1} - y_{t-n+1}$$

Dans ces écritures, tous les coefficients fonctions de $n$ peuvent être pré-calculés. La procédure finale n'a besoin, pour un $n$ donné, que de 6 (*resp.* 7, 3) opérations arithmétiques pour obtenir $\mathcal{S}_{n,t}$ (*resp.* $\mathcal{P}_{n,t}$, $\mathbb{S}_{n,t}$). L'occupation en mémoire est de cinq jeux de coefficients, et un tampon des $n$ dernières mesures. C'est assez efficient en temps de calcul et en utilisation de mémoire pour être utilisé en ligne, quelque soit la plate-forme d'implémentation.

## 2.5 Erreur et écart-type

La variance locale $\sigma^2_{n,t+1}$ peut être calculée en temps constant à partir de $\sigma^2_{n,t}$ :

$$\begin{aligned}(n-1)\,\sigma^2_{n,t+1} = (n-1)\,\sigma^2_{n,t} \quad &+ \quad y^2_{t+1} - y^2_{t-n+1} \\ &+ \quad \frac{2}{n} (y_{t-n+1} \mathbb{S}_{n,t} - y_{t+1} \mathbb{S}_{n,t+1}) \\ &+ \quad \frac{2}{n} (\mathbb{S}_{n,t} - \mathbb{S}_{n,t+1}) (\mathbb{S}_{n,t+1} - y_{t+1})\end{aligned}$$

$$+ \quad \frac{1}{n}\left(\mathbb{S}^2_{n,t+1} - \mathbb{S}^2_{n,t}\right)$$

L'erreur quadratique moyenne d'approximation vaut alors $\sigma_{n,t}\sqrt{1-\mathcal{R}^2_{n,t}}$ pour un coefficient de régression $\mathcal{R}_{n,t}$ [Sap90, pp.366] :

$$\mathcal{R}_{n,t} = \mathcal{S}_{n,t}\frac{\sqrt{\frac{n^2-1}{12}}}{\sigma_{n,t}}$$

Le terme $\sqrt{\frac{n^2-1}{12}}$ correspond à l'écart-type du vecteur $[1\ 2 \ldots n]'$.

Il nous sera parfois utile de calculer $\sigma^2_{n,t}$ à partir de $\sigma^2_{n-1,t}$. On sait

$$\sigma^2_{n,t} = \frac{1}{n}\sum_{i=1}^{t}\left(y_i - \frac{1}{n}\mathbb{S}_{n,t}\right)^2$$

et

$$\mathbb{S}_{n,t} = \mathbb{S}_{n-1,t} + y_t$$

On a par ailleurs l'expression de mise à jour incrémentale de la variance pour $j+1$ échantillons à partir de la variance de $j$ échantillons :

$$\sigma^2_{(j+1)} = \left(1 - \frac{1}{j}\right)\sigma^2_{(j)} + (j+1)\left(\mu_{(j+1)} - \mu_{(j)}\right)^2$$

où $\mu_{(j)}$ est la moyenne pour $j$ échantillons. Nous appliquons cette formule, sachant que l'on rajoute l'échantillon $y_t$ dans le calcul de la variance $\sigma^2_{n,t}$ à partir de $\sigma^2_{n-1,t}$. Alors :

$$\sigma^2_{n,t} = \frac{n-2}{n-1}\sigma^2_{n-1,t} + \frac{1}{n\,(n-1)^2}(n\,y_t - \mathbb{S}_{n,t})^2$$

sachant

$$\sigma^2_{2,t+1} = \frac{1}{2}(y_{t+1} - y_t)^2$$

Nous nous servirons pratiquement de ces expressions pour calculer *toutes* les régressions (*c-à-d.* pour tous les $n$ jusqu'à un $n_{max}$) au Chap.3.

## 2.6 Conclusion : une approximation linéaire efficace de la tendance

Nous avons défini et développé des méthodes efficaces de régression par filtrage linéaire (en temps de calcul linéaire en $n$) et par approche incrémentale (en temps de calcul constant par rapport à $n$). Nous avons détaillé exclusivement le cas de la régression linéaire d'une variable par rapport au temps. L'extension à des approximations du comportement temporel d'ordre plus élevé est abordée en Annexe C.

Ces approximations de degré supérieur de la tendance sont, pour les données dont nous nous occupons, sensibles aux anomalies et ont tendance à suivre de près les erreurs et les valeurs aberrantes.

D'un autre côté, du fait des usages courants et de la culture statistico-mathématique du milieu médical, une droite de meilleure approximation est plus facile à inclure dans un système à base de connaissances : elle est bien plus intuitive pour le personnel médical qu'une parabole de meilleure approximation.

D'autre part, le problème de la recherche d'un ordre d'approximation polynomiale vient à compliquer les choix dans la volonté de rester dans une approche basée sur les données.

L'arbitraire du choix du premier ordre est donc justifié par sa simplicité et sa facilité d'interprétation.

De plus, une approximation de la dérivée comme celle que nous proposons permet de reconstruire le signal par intégration (en fait par sommation cumulée) ; le résultat retient les composantes de basse fréquence, en éliminant le bruit considéré comme haute fréquence.

La fréquence de coupure du filtre ainsi construit (qui est un filtre moyenneur, avec des coefficients pour les points extrêmes qui accentuent les points en avant et réduisent ceux en arrière) est fonction directe de $n$.

Un déphasage est introduit en même temps. Il peut être corrigé hors-ligne par une avance de $\lfloor \frac{n}{2} \rfloor$ unités de temps, mais il introduit un retard inévitable en ligne.

# 3

# Une échelle caractéristique séparant le long terme du court terme

Où l'on cherche une échelle caractéristique pour le calcul des tendances de chaque donnée.

Sommaire

Pouvant calculer une tendance locale, assimilable à une dérivée, il est désormais possible de décomposer chaque flot de données en valeur, tendance et stabilité. Cette dernière sera assimilée à l'écart-type des données dans la fenêtre considérée. Elle peut être interprétée comme un indicateur de la variabilité locale des données. Ainsi, on a un moyen de définir le comportement à court terme ; les variations de ces trois variables (valeur, tendance et stabilité) détermineront le comportement à long terme. Il reste à établir la frontière entre les deux.

Comme hypothèse de base, on suppose l'existence d'une échelle temporelle $\tau$ , caractéristique et propre à chaque donnée, qui permet de séparer le court terme du long terme.

Dans le système Respaid [CRC+89], cette échelle était définie arbitrairement, et de façon unique pour tous les paramètres physiologiques mesurés. Nous cherchons ici à éliminer cet aspect arbitraire en proposant des critères qui permettent d'établir une échelle caractéristique $\tau$ propre à chaque paramètre physiologique. Nous tiendrons ainsi compte de leur diversité et leurs dynamiques différentes (et celles des systèmes physiologiques sous-jacents). De plus, nous essaierons de cette façon de différencier les données par rapport aux traitements qu'elles ont subis depuis le point de mesure sur le patient, à travers les capteurs jusqu'à la sortie du système de monitorage.

Nos objectifs sont :

1. de pouvoir effectuer un filtrage passe-bas des données, sans l'introduction de connaissances *a priori* — autres que les critères établis et leurs paramètres d'application ;
2. l'extraction d'une variable assimilable à une dérivée, pour la traduction des notions de variabilité — employées dans les discours courants du personnel soignant — en des données numériques ;
3. l'obtention d'un indicateur local de variabilité, pour l'exploitation de la notion de stabilité et la comparaison et la caractérisation de chaque paramètre physiologique.

Nous devrons définir les ensembles de données de base sur lesquelles opérer.

Afin de déterminer leur adaptation à nos objectifs courants et généraux, les critères proposés — que nous aurons formalisé en indicateurs — subiront une évaluation et une comparaison.

Nous allons opérer de la façon suivante : après avoir formalisé les critères empiriques proposés, nous les déclinerons en un ensemble d'indicateurs quantitatifs. Pour définir ceux-ci, un certain nombre de prérequis seront abordés et résolus. Ensuite, nous appliquerons les indicateurs sur des séries chronologiques artificielles possédant des caractéristiques proches de celles de nos données. Enfin, nous les appliquerons sur nos données, de façon à les évaluer de façon intrinsèque, par leur spécificité à chaque paramètre physiologique et leur sensibilité aux conditions de calcul.

## 3.1 Critères de définition

Nous posons ici les traductions formelles des critères pour définir l'échelle caractéristique $\tau$ introduite précédemment.

Informellement, nous cherchons à définir l'échelle pour laquelle on peut établir "une bonne estimation de dérivée". Formellement, cela revient à :

1. chercher l'échelle au-delà de laquelle nous n'améliorons plus l'approximation locale de la dérivée, ou bien
2. chercher l'échelle la plus petite pour laquelle on a pour la première fois une bonne approximation de la dérivée.

Essayons de trouver des implémentations formelles précises de ces notions.

### 3.1.1 Stabilisation de la tendance

Le premier critère se voit formalisé de deux façons :

a) d'abord, on cherche l'échelle à partir de laquelle l'approximation reste stable, en terme de distribution de la tendance ;
b) ensuite, on cherche l'échelle à partir de laquelle, l'approximation reste globalement stable, en référence à un modèle général de l'évolution des données.

Concrètement, a) se traduit par la recherche d'un indicateur caractéristique des distributions de la tendance en fonction de l'échelle, et la recherche d'une échelle au-delà de laquelle cette caractéristique est stabilisée. Nous avons exploré l'utilisation d'une statistique de Kolmogorov-Smirnoff afin d'établir la normalité possible de la distribution de la tendance. Or, la forme typique de cette distribution n'est pas gaussienne.

De façon plus fine, nous avons essayé de caractériser la distribution de la différence temporelle de la tendance (qui approche donc la dérivée seconde localement). Cette quantité contient l'information sur l'accélération des paramètres ; sauf cas de données variant de façon "anormale" — *c-à-d.* avec des sauts très brusques — l'accélération doit être distribuée d'une façon spécifique, et concentrée autour de quelques modes caractéristiques. L'accélération exhibe une distribution centrée et symétrique, mais beaucoup plus concentrée vers le zéro qu'une distribution normale.

Nous avons cherché à caractériser de façon plus générale cette distribution. L'exploration montre qu'elle est en général unimodale et symétrique. L'écart-type apparaît donc comme une bonne façon de la caractériser.

Nous avons calculé $\sigma\left(\Delta_t \mathcal{S}_{n,t}\right)$, l'écart-type de la différence temporelle des tendances, à toutes les échelles temporelles — jusqu'à un maximum raisonnable en terme de pertinence physiologique et de faisabilité informatique. Cet indicateur est une mesure de la dispersion des accélérations de la donnée considérée [CCPR97]. Il est strictement décroissant, ce qui indique que plus l'échelle croît, moins l'accélération de chaque donnée varie.

On peut alors définir un seuil $\alpha$ au-dessous duquel on considérera que l'accélération de la donnée est stabilisée : sa dispersion est inférieure à $\alpha$. Au vu de la distribution typique de $\Delta_t \mathcal{S}_{n,t}$, quelque 75% de la distribution est comprise dans l'intervalle $[-\alpha, +\alpha]$. Pour un $\alpha$ donné, les divers paramètres physiologiques présentent des $\tau$ différents, proches pour le même sous-système physiologique, *cf.* Fig. 3.2.

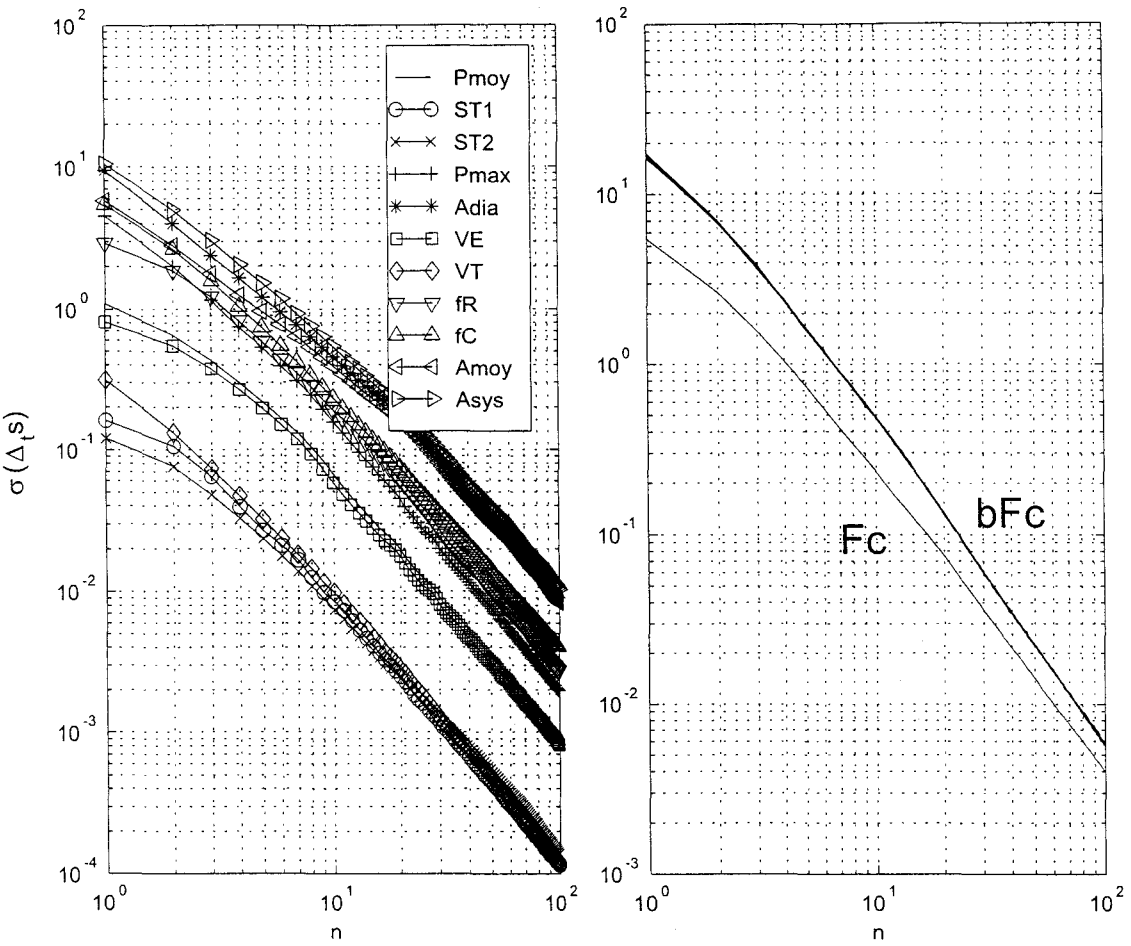



FIG. 3.1 – Application de l'indicateur $\sigma(\Delta_t \mathcal{S}_{n,t})$
A gauche, divers paramètres physiologiques, pour un même patient présentent des caractéristiques différentes. A droite, la fréquence cardiaque $F_c$ et un ensemble de trente brouillages de sa série chronologique b$F_c$. Ceci montre (et c'est le cas pour presque tous les paramètres, avec l'exception notable de $V_T$) que l'information temporelle est effectivement repérée par le traitement. Échelles logarithmiques. L'abscisse représente le nombre de points considérés pour l'estimation de la tendance.

Cette évolution de $\sigma(\Delta_t \mathcal{S}_{n,t})$ en fonction de l'échelle est grossièrement logarithmique, avec des formes et des décalages qui différencient et regroupent les divers paramètres physiologiques.

La Fig. 3.1 représente les comportements typiques de nos données par rapport à ce critère.

Pour un nombre — certes petit (six) — de patients, l'indicateur calculé sur le recueil d'une journée de $24h$ est semblable (qualitativement) pour le même paramètre entre patients, regroupant les paramètres correspondant aux mêmes sous-systèmes physiologiques.

Une comparaison par *brouillage* permet d'argumenter en faveur de la pertinence de l'approche.

La technique de brouillage (*surrogate data*) ne fait intervenir, à la manière du *bootstrap*, aucune hypothèse sur les paramètres des distributions des variables [ST95]. Il s'agit de comparer les résultats obtenus par la technique que l'on teste avec ceux qu'on obtient par la même technique

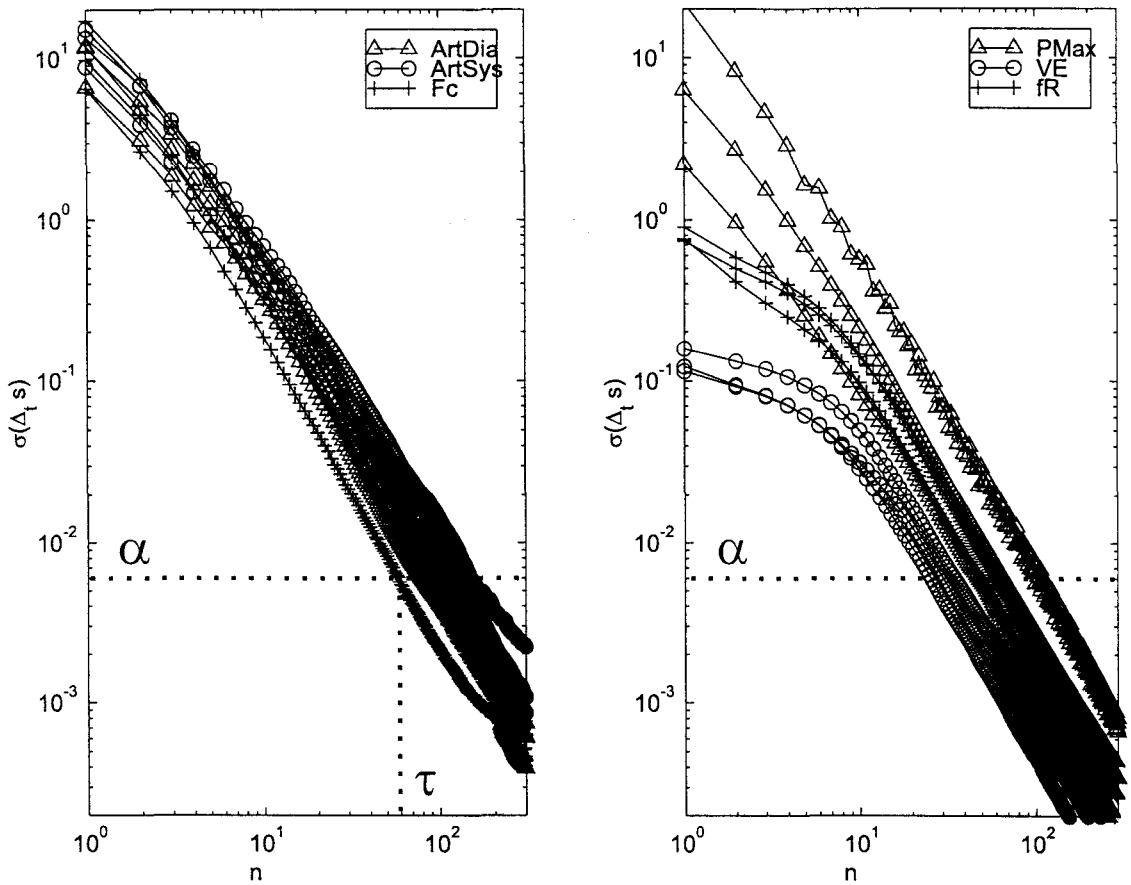



FIG. 3.2 – Indicateur $\sigma\left(\Delta_t \mathcal{S}_{n,t}\right)$ : comparaison entre patients
A gauche, les paramètres cardio-vasculaires présentent des comportements de $\sigma\left(\Delta_t \mathcal{S}_{n,t}\right)$ proches (ici trois patients, pour la lisibilité du schéma). A droite, les paramètres respiratoires. Le seuil $\alpha$ de $5 \cdot 10^{-3}$ est représenté pour mieux repérer les variations de $\tau$ d'un sous-système à l'autre. Échelles logarithmiques. En abscisse, l'échelle $n$, *c-à-d.* le nombre de points sur lequel on calcule la tendance.

sur des données brouillées. Celles-ci présentent un certain nombre de caractéristiques des données réelles, sauf celle qu'on essaye de mettre en évidence.

Dans notre cas, on cherche à savoir si la dépendance temporelle des données est à l'origine des résultats qu'on observe. On effectue un certain nombre de *permutations temporelles* des données ; par ces réalisations brouillées, on conserve la distribution originale, et on élimine toute information portant sur l'autocorrélation. Alors, si l'application du traitement produit des résultats qualitativement différents pour ces données brouillées, c'est que l'ordre temporel intervient.

Ces techniques sont proposées par exemple par [TGL+92]. Ce principe est à la base de nombreux travaux, par exemple [TT96] dans le contexte des processus stochastiques appliqués à l'étude de l'ECG, ou encore [Pal96] dans le contexte de la caractérisation du chaos et la stochasticité de l'EEG.

Le paramètre $\alpha$ doit cependant être fixé *a priori* pour pouvoir déterminer le temps caractéristique $\tau$. La dépendance de $\tau$ par rapport à $\alpha$ est importante. Même si on garde la différenciation entre variables, il n'est pas facile de donner une justification — autre que physiologique, pour

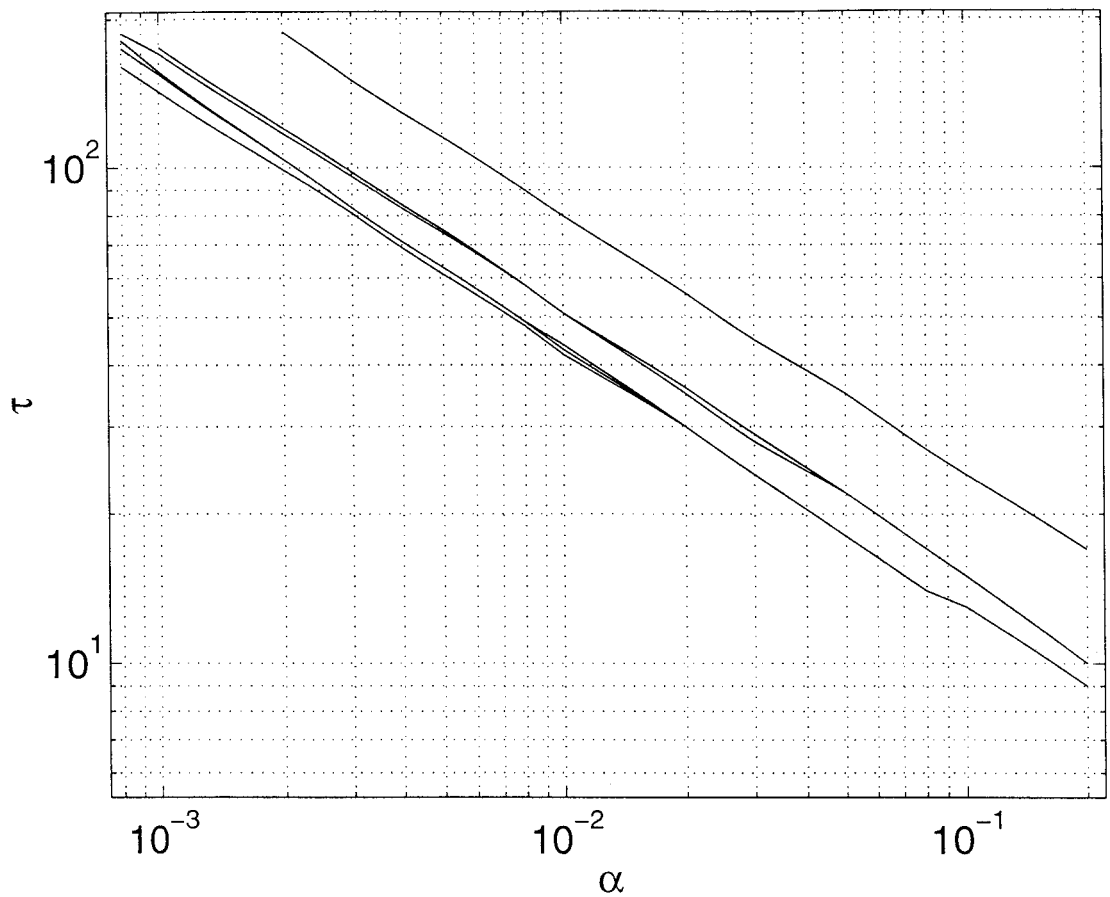



FIG. 3.3 – Indicateur $\sigma(\Delta_t \mathcal{S}_{n,t})$ : sensibilité au seuil $\alpha$ pour $F_c$
Variation de $\tau$ (en nombre de points de mesure, soit ici une toutes les 15s) en fonction de $\alpha$, pour le paramètre $F_c$ calculé sur six patients. La variation est essentiellement exponentielle. Échelles logarithmiques.

autant que la littérature traite de ces questions — du choix d'un $\alpha$ particulier. A titre d'exemple, la Fig. 3.3 représente le comportement de $\tau(\alpha)$, pour un même paramètre $F_c$, pour six patients.

Par rapport à l'étude de la distribution, on a posé un seuil pour $\sigma(\Delta_t \mathcal{S}_{n,t})$, qui correspond à une mesure globale de la variabilité de l'accélération. Le pendant non paramétrique de cette approche a été abordé de la même façon : on définit un indicateur $\Delta_t \mathcal{S}_{n,t}|_{.95}$ qui donne l'accélération correspondant à 95% de la masse de la distribution. Les résultats sont tout à fait semblables à ceux obtenus à partir de $\sigma(\Delta_t \mathcal{S}_{n,t})$.

### 3.1.2 Minimisation de l'erreur

Pour l'interprétation b) de §3, on propose d'étudier l'erreur d'approximation commise localement, en fonction de l'échelle choisie.

Sur une longueur $T$ d'enregistrement, la distribution de l'erreur suit en général un comportement dépendant peu de $n$. Seules les valeurs extrêmes présentent des extrema en fonction de $n$.

Quant à la médiane, elle présente parfois des minima, mais ce n'est souvent pas le cas. L'écart-type tend plutôt à se stabiliser.

Ce critère n'est pas exploitable. Le précédent implique la définition arbitraire d'un seuil qui, même s'il peut être interprété de façon concrète, doit être ajusté par l'expert. Or, nous voulons précisément renvoyer le plus en aval possible l'introduction de cette connaissance. Ainsi, nous laissons de côté ces critères et proposons de nous baser sur la notion de signification statistique, pour avoir un critère dont les paramètres sont les moins arbitraires possibles.

### 3.1.3 Bonne approximation globale

Afin d'intégrer les critères précédents (stabilisation globale de la tendance et qualité de l'approximation), on propose finalement d'étudier la *signification* de l'approximation en fonction de l'échelle. En effet, chaque régression locale peut être soumise à un test statistique, qui met en évidence le décalage entre les résultats obtenus et ceux qui ne seraient dûs qu'au seul hasard. Ce test permet de déterminer si le calcul de la régression est significatif ou non.

Concrètement, on compare chaque résultat à la distribution d'une statistique calculée dans le cas de deux variables indépendantes. A un risque $\alpha$ de se tromper, un seuil calculé à partir de la distribution de référence est établi. S'il est dépassé — en valeur absolue : nous traitons des distributions symétriques — alors on a $\alpha/2$ chances que la valeur qu'on teste ne provienne pas de cette distribution. On peut alors affirmer — avec $\alpha$ chances de se tromper — que cette valeur n'est pas issue de deux variables indépendantes, il y a donc corrélation effective.

Pour un $\alpha$ fixé, le test fournit pour chaque valeur de régression — donc pour tout point et pour toute échelle considérée — un indicateur booléen de signification, noté $\mathcal{K}_{n,t}$.

Afin d'établir l'échelle caractéristique, nous chercherons l'échelle pour laquelle les régressions sont "le plus souvent" significatives. C'est donc l'échelle à laquelle on a le plus souvent le droit d'approcher localement les données par leur approximation linéaire.

On peut décliner cette approche suivant les deux critères donnés en §3.1.

Pour le premier, nous chercherons l'échelle qui maximise en premier le nombre de régressions significatives. C'est donc la première échelle à laquelle nous avons le droit globalement, *c-à-d.* la plupart du temps, d'approcher la tendance par la droite de régression.

Pour le deuxième, nous chercherons l'échelle à partir de laquelle les zones (les intervalles temporels) d'approximations significatives sont stabilisées. C'est l'échelle après laquelle, on obtiendra les mêmes découpages en zones d'approximation correcte (*a priori* les zones qui présentent des caractéristiques d'évolution stables et bien approchées par des droites localement) et zones d'approximation incorrecte (*a priori* les zones de perturbation ou de transitoires).

En Annexe E nous détaillons et discutons le test statistique employé.

## 3.2 Affinement des critères

Ayant un indicateur booléen $\mathcal{K}_{n,t}$ en chaque point de mesure $t$ et pour chaque échelle $n$, indiquant la signification de l'approximation par la tendance, au risque $\alpha$ déterminé, nous pouvons procéder à l'établissement de critères formels. Ceux-ci se déclinent suivant deux orientations : à partir d'un décompte des points de signification et à partir d'un décompte des périodes de signification.

### 3.2.1 Décompte des points de signification

Définissons :

$$\mathcal{N}(n) = \sum_{t=t_1}^{t_2} \mathcal{K}_{n,t}$$

pour un ensemble de mesures comprenant les instants de $t_1$ jusqu'à $t_2$. Le choix de ces bornes — *c-à-d.* le choix de la population de référence — sera précisé en §3.4 ; pour l'instant, nous considérons une approche hors ligne, dans laquelle on peut prendre $t_1 = 0$ et $t_2 = T$ la longueur totale de mesure.

L'indicateur $\mathcal{N}(n)$ représente le nombre de points de mesure pour lequel l'approximation locale par une droite est significative (au risque $\alpha$), en fonction de l'échelle. La Fig. 3.4 schématise le traitement effectué.

L'indicateur $\mathcal{N}(n)$ suit, pour nos données, l'un des comportements caractéristiques suivants :

1. Croissant en $n$ et stabilisant, suivant un comportement qualitativement comparable à une exponentielle (Fig. 3.5).
2. Présentant des maxima locaux pour une tendance générale croissante (Fig. 3.6).
3. Brutalement croissante puis décroissante enfin suivant la tendance générale exponentielle (Fig. 3.7).
4. Croissante de façon qualitativement linéaire, en fait exponentielle de *très* longue constante (Fig. 3.8).
5. Croissante et décroissante, avec des valeurs très basses, voire plate, pour des données globalement constantes par morceaux (Fig. 3.9).

Les figures 3.5 à 3.9 en pages 51 à 55 illustrent ces comportements-type.

D'après ces comportements typiques, nous proposons de définir l'échelle caractéristique $\tau_{\mathcal{N}}$ comme un maximum en $n$ du comportement de $\mathcal{N}(n)$. Plusieurs choix sont encore possibles :

PRM Le premier maximum, avec un lissage médian afin d'éviter des artefacts ;

LNG Le maximum "le plus long", *c-à-d.* celui qui est dépassé le plus tard, suivant les $n$ croissants ;

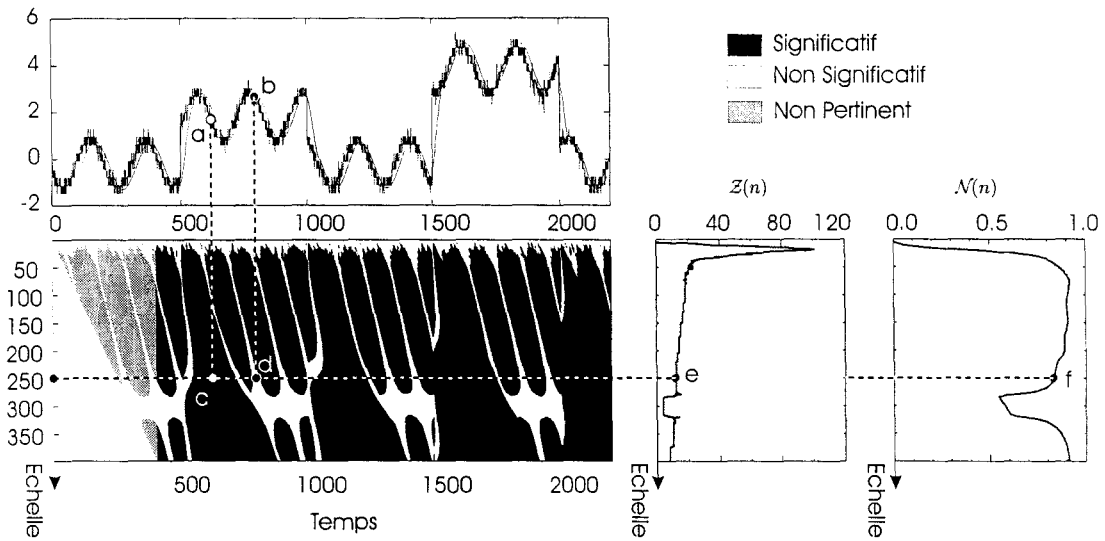



FIG. 3.4 – Illustration du calcul de $\mathcal{N}(n)$

En haut à gauche, une série chronologique artificielle, un sinus décale bruité puis quentifié. La ligne fine continue représente la reconstruction de cette série par intégration de $\mathcal{S}_{n,t}$ à l'échelle 100. Notez le retard introduit et le filtrage effectif du bruit et la quantification. L'abscisse est donnée en unité arbitraire de temps, l'ordonnée en unité arbitraire de valeur.

En bas à gauche, la carte de signification de la série. L'abcsisse représente le temps et correspond au point *final* de chaque ensemble de $n$ mesures passées. L'ordonnée représente l'échelle, *croissante vers le bas*. Les régressions significatives sont représentées en noir, les non significatives en blanc. Les points en gris correspondent aux points qui n'ont pas été pris en compte pour le calcul de $\mathcal{N}(n)$ et $\mathcal{Z}(n)$, d'où non pertient.

En bas à droite, de gauche à droite : le décompte du nombre de zones de signification contiguës $\mathcal{Z}(n)$ et la part de régressions significatives $\mathcal{N}(n)$ comme fraction de la taille totale. Pour le calcul des deux, les point non pertinents ont été exclus. *L'échelle est croissante vers le bas.*

Par exemple : A l'instant 610 (point a), la régression sur 250 points précédents est significative (point c). A l'instant 805 (point b), la régression sur 250 points précédents ne l'est pas (point d). A l'echelle 250, douze zones de régression significative contiguë sont présentes (point e), hors points non pertinents. A la même échelle, 82% des points pertinents mènent à une régression significative (point f).

Le risque est de $\alpha = 10^{-2}$.

ABS Le maximum absolu.

Le premier critère répond au principe de la recherche de la meilleure approximation de la dérivée : on prendra comme échelle caractéristique celle à laquelle on a pour la première fois une bonne approximation locale ; après, on perdra de la qualité d'approximation. Ce critère détecte correctement les comportements-type 1 et 3 ; le comportement-type 2 étant plus complexe.

Le deuxième répond mieux à la situation où des échelles multiples coexistent. Dans ce cas, le temps caractéristique ne sera plus le premier maximum, mais le maximum qui correspond à l'échelle dominante. Par là nous voulons dire qu'au-delà de cette échelle, nous perdrons de la qualité d'approximation jusqu'à des échelles bien plus grandes ; et en deçà nous aurons d'autres échelles, plus proches de la dérivée mais moins pertinentes pour un traitement global. Nous faisons en sorte de ne considérer le maximum asymptotique que dans le cas où le comportement est strictement monotone. Cette approche s'insère bien dans le cas des comportements-type 2, 3 et 4.

Le troisième critère englobe les comportements-type 1, 3 et 4. Cependant, vu le comportement globalement croissant en $n$ de $\mathcal{N}(n)$, il faudra étudier des $n < n_{max}$ très grands pour avoir un maximum absolu qui ne dépende pas directement du $n_{max}$ utilisé.

Dans les trois cas, on prendra comme $\tau_{\mathcal{N}}$ l'échelle située à 5% au-dessous et avant le maximum en question. Ceci afin d'absorber les asymptotes horizontales.

### 3.2.2 Décompte du nombre de zones significatives

Les critères basés sur $\mathcal{N}(n)$ sont appropriés pour une classe de données bruitées, quantifiées, localement approximables par des droites. La classe plus restreinte des données au comportement temporel approximativement linéaire par morceaux est aussi incluse dans celle-là.[1]

Souvent, nos données présentent un aspect proche de la classe de séries chronologiques linéaires par morceaux, BQTP. Il semble alors adéquat de proposer une définition d'échelle caractéristique prenant directement compte de cet aspect. Définissons :

$$\mathcal{Z}(n) = \left| \{t | \mathcal{K}_{n,t} \neq \mathcal{K}_{n,t-1} \quad t \in [t_1, t_2]\} \right|$$

Il s'agit du décompte des passages d'une zone de mesures significatives à une zone de mesures non significatives. Alors $\lfloor \mathcal{Z}(n)/2 \rfloor$ est le nombre de zones contiguës significatives délimitées à l'échelle $n$. Lorsque les données sont effectivement linéaires par morceaux BQTP, on définit $\tau_{\mathcal{Z}}$ tel que à partir de $n = \tau_{\mathcal{Z}}$, on aura repéré toutes les zones linéaires, séparées entre elles par des épisodes non significatifs, correspondant aux cassures et aux perturbations.

Il s'agit ici d'une perspective pessimiste de définition de l'échelle caractéristique : $\tau_{\mathcal{Z}}$ est l'échelle au-delà de laquelle il n'est plus utile d'opérer, alors que les $\tau_{\mathcal{N}}$ définis précédemment essayent tous de saisir le plus petit $\tau$ pour lequel il est possible d'opérer.

Le comportement typique sur nos données de $\mathcal{Z}(n)$ est illustré en Fig. 3.10 : il s'agit d'une

[1] Au bruit, la quantification et les absences près, désormais noté "BQTP".

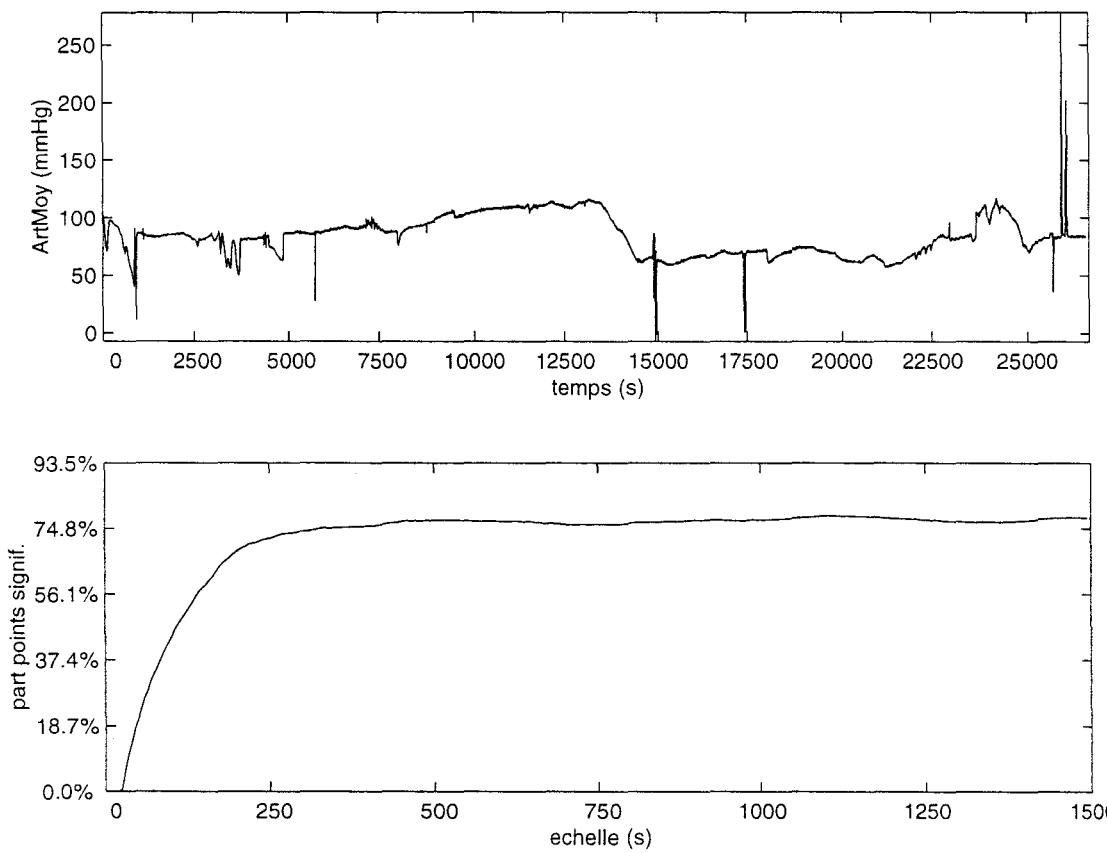



FIG. 3.5 – Comportement-type de $\mathcal{N}(n)$ : croissant exponentiel
Le paramètre physiologique représenté en haut est la pression artérielle moyenne. L'indicateur $\mathcal{N}(n)$ correspondant est représenté dans la figure du bas : il y a stabilisation asymptotique autour de 75% de points de régression significative.

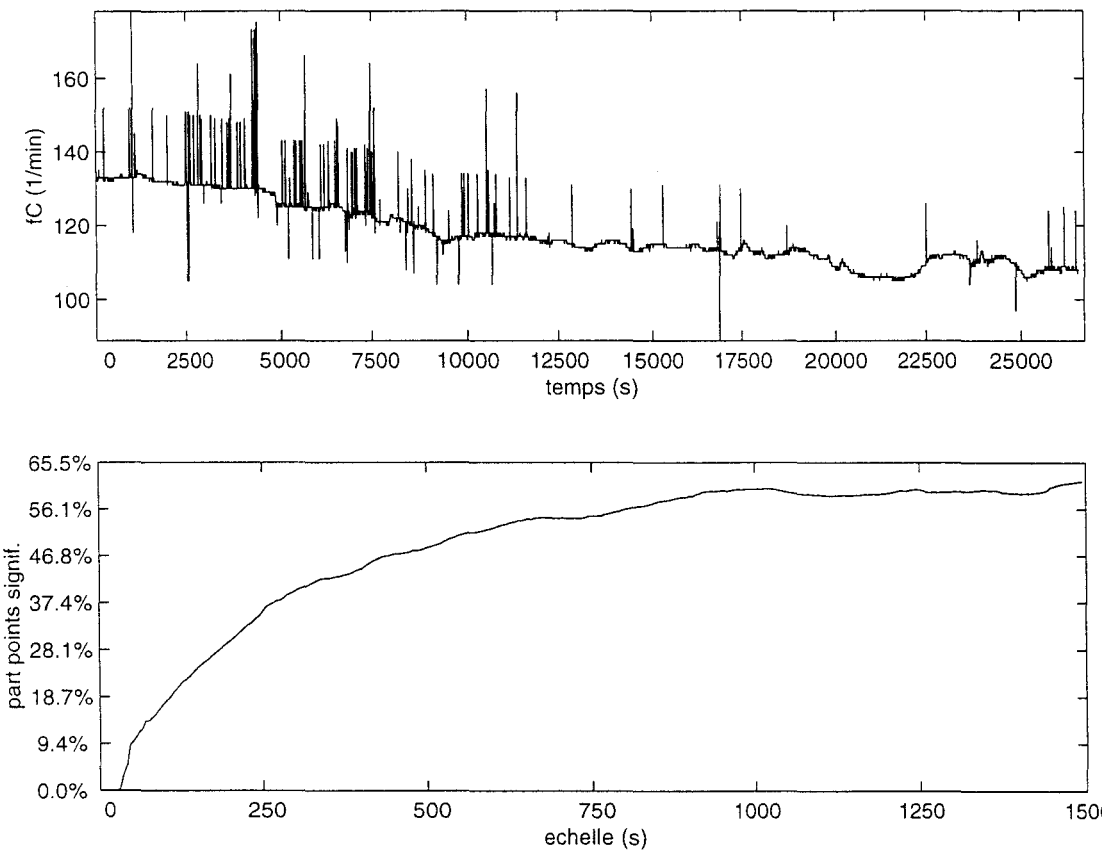



FIG. 3.6 – Comportement-type de $\mathcal{N}(n)$ : maxima locaux sur croissance exponentielle
La fréquence cardiaque, en haut, présente des échelles caractéristiques multiples : les maxima et plateaux locaux observés sur les échelles $(n)$ croissantes, avant stabilisation vers 60% de points significatifs.

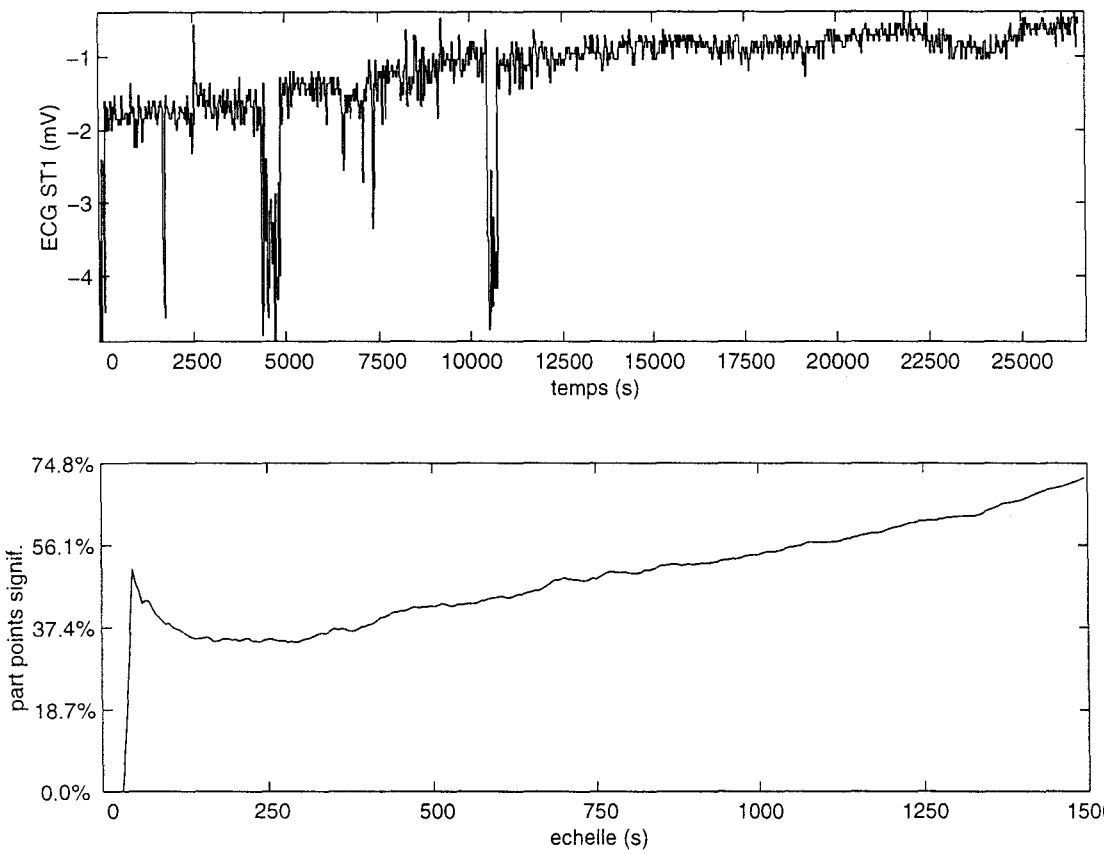



FIG. 3.7 – Comportement-type de $\mathcal{N}(n)$ : croissance brutale, décroissance, croissance exponentielle
Le paramètre $ST_1$ présente un $\mathcal{N}(n)$ qui part d'une valeur assez élevée pour atteindre un plateau puis suivre une croissance progressive, vers des valeurs fortes. Les perturbations fortes qu'on peut observer sont qualitativement synchrones des perturbations observées en Fig. 3.8, et sont probablement dues à des manipulations effectuées sur le patient. Ce genre de connaissance n'est appliquable que par analyse multivariée ; elle n'entre pas dans la méthodologie à ce stade.

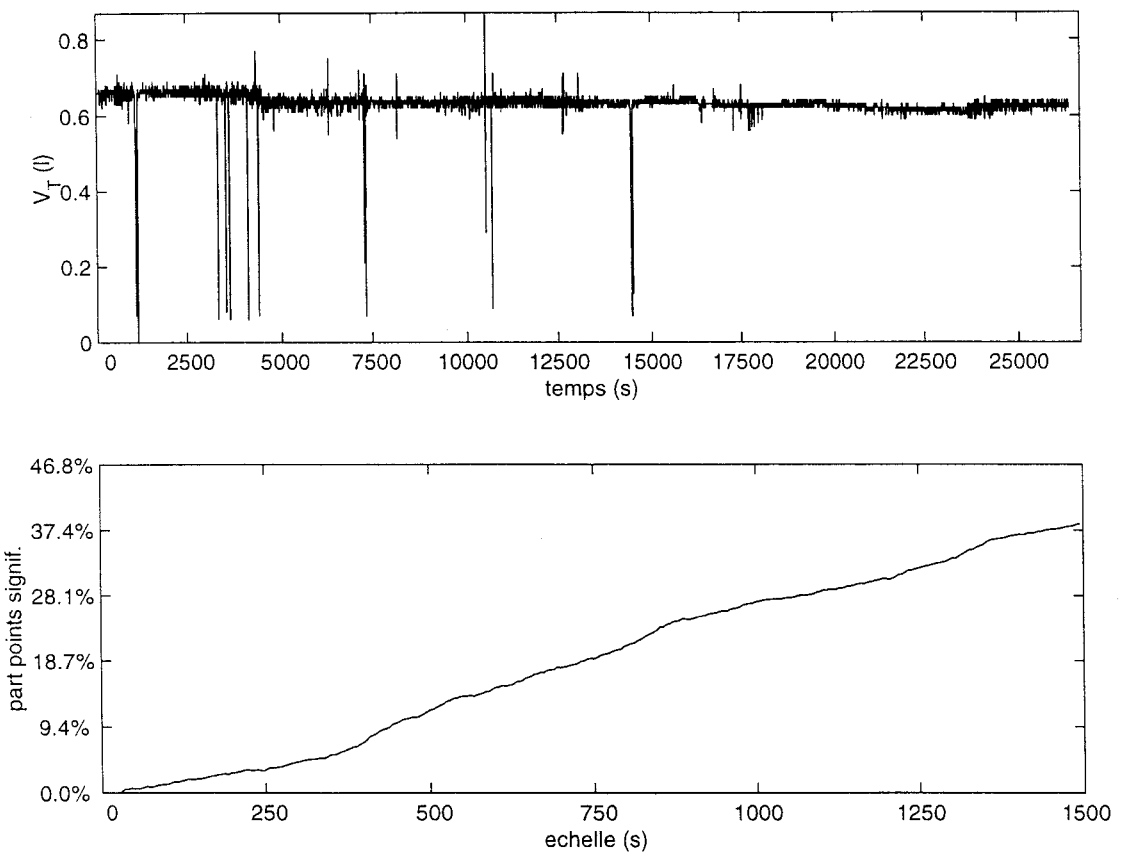



FIG. 3.8 – Comportement-type de $\mathcal{N}(n)$ : linéaire-exponentielle

Le volume courant présente, malgré des perturbations (probablement dues à des aspirations), une caractéristique $\mathcal{N}(n)$ essentiellement linéaire, dans les valeur faibles et qui est en fait une exponentielle de très grande constante de temps. Par ailleurs ce paramètre présente des caractéristiques qui le rapprochent d'un bruit. Cet aspect d'indépendance temporelle est rendu par la difficulté à définir une échelle caractéristique pour le calcul de la tendance.

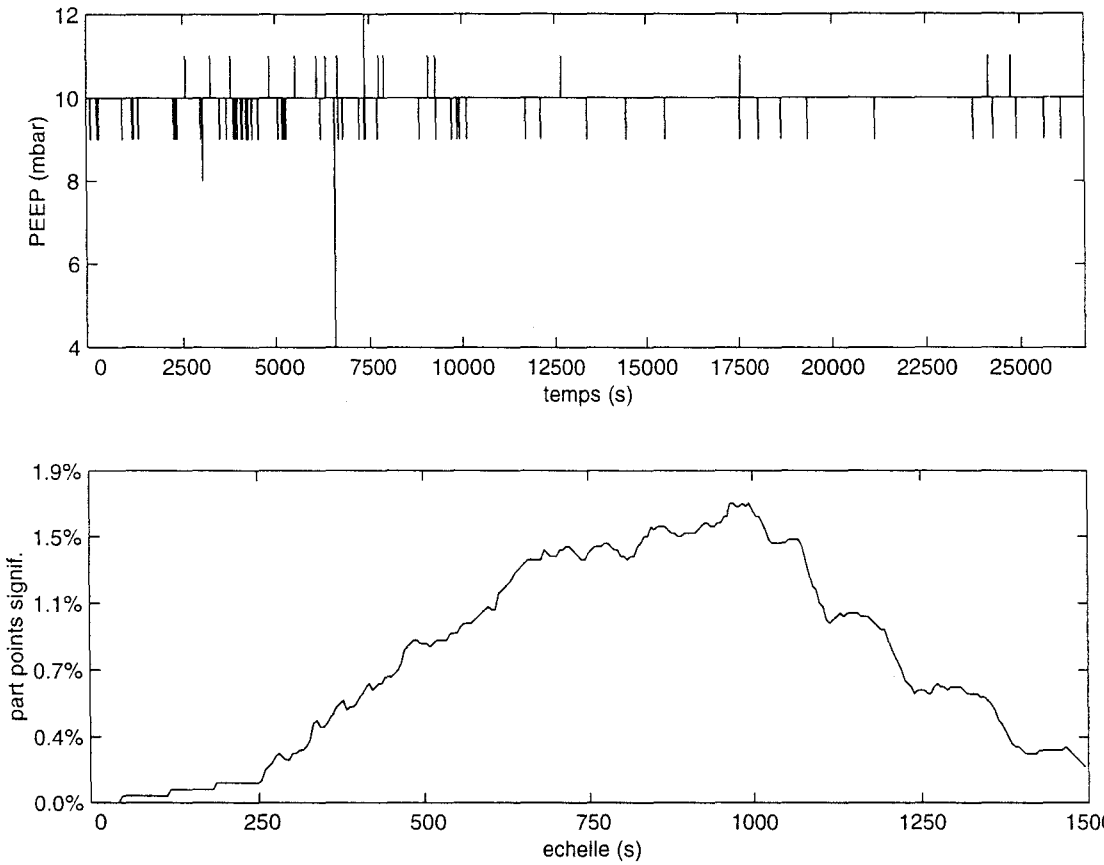



FIG. 3.9 – Comportement-type de $\mathcal{N}(n)$ : cas particulier

La pression positive en fin d'expiration est un réglage du système de ventilation. Il s'agit en principe d'une consigne, d'où la valeur essentiellement constante. Le nombre infime de régressions significatives montre bien qu'il n'est pas adéquat de parler de tendance pour de tels paramètres. De la même façon, des paramètres comme $FIO_2$, la fraction d'oxygène inspiré (non représenté), qui a souvent un comportement constant par morceaux présente une caractéristique $\mathcal{N}(n)$ strictement croissante, mais tout aussi inadéquate.

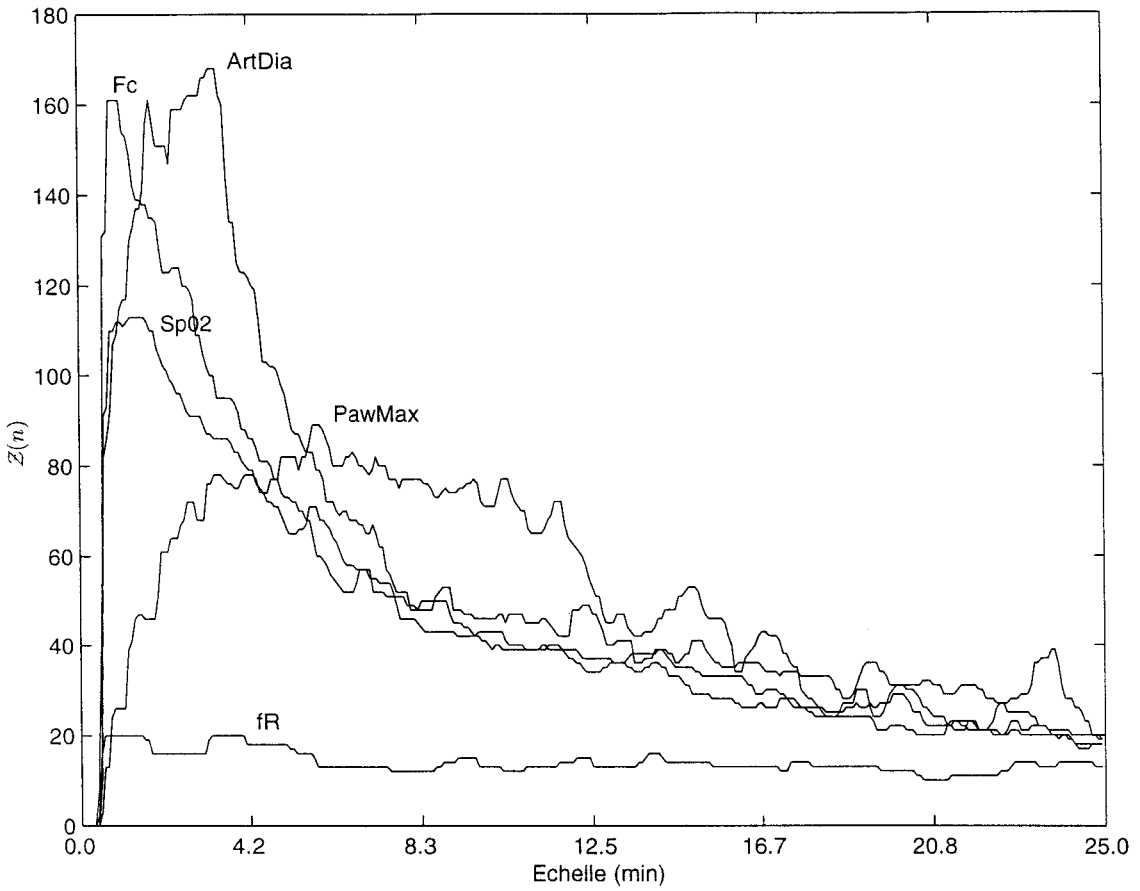



FIG. 3.10 – Comportement typique de l'indicateur $\mathcal{Z}(n)$
Nous avons représenté le comportement de $\mathcal{Z}(n)$ sur cinq paramètres physiologiques, d'un recueil de $8h$. Un lissage médian sur cinq points a été introduit pour aider à la clarté du schéma. Le comportement typique croissant subitement puis décroissant est évident et différencié pour les données $F_c$, $SpO_2$, $P_{aw}max$ et $Art_{Dia}$. La fréquence respiratoire $f_R$, par contre, présente un comportement différent, dû au nombre faible de modalités de cette donnée. La série chronologique (non représentée) est en effet proche d'un signal constant par morceaux.

croissance brutale suivie d'une décroissance beaucoup plus douce. Des artefacts sont présents. Pour définir $\tau_{\mathcal{Z}}$ on peut proposer, à côté des critères PRM, LNG et ABS :

NZM on prendra la dernière échelle située à 5% au-dessus de la valeur finale, sur un lissage médian de $\mathcal{Z}(n)$. C'est l'échelle à laquelle on a trouvé pour la première fois un modèle global linéaire par morceaux.

## 3.3 Application à des données artificielles

Pour explorer les propositions pour $\tau_{\mathcal{Z}}$ et $\tau_{\mathcal{N}}$ suivant PRM, LNG ABS et NZM nous avons recours à un ensemble de données artificielles qui miment certaines propriétés des données physiologiques.

Spécifiquement, nous avons utilisé :

$\mathcal{E}_{\text{cm}}$ Un signal constant par morceaux, une des formes-types que nos signaux peuvent présenter. Le fait que, pour un signal strictement constant, la régression suivant le temps n'est pas significative (il n'y a pas de dépendance temporelle) peut induire des artefacts d'interprétation ; nous testons ici jusqu'à quel point cela est vrai.

$\mathcal{E}_{\text{cp}}$ Une constante avec des impulsions sporadiques. Est-ce que les impulsions sont surexposées par la méthode, comme toute méthode basée sur des mesures de distance quadratique ?

$\mathcal{E}_{\text{ss}}$ Un signal avec deux composantes sinusoïdales de fréquence relative et amplitude relative variable. Ceci pour repérer les extensions possibles et les adéquations et/ou contradictions avec le genre de signaux typiques que les méthodes classiques de traitement du signal peuvent caractériser et traiter.

$\mathcal{E}_{\text{sd}}$ Un signal sinusoïdal avec une composante continue intermittente irrégulière, constante sur l'ordre de grandeur de la période de la sinusoïde. Si les oscillations sont rares dans nos données, est-ce parce qu'elles sont noyées dans le bruit ?

A ces séries chronologiques de base, nous avons rajouté du bruit gaussien avec divers rapports signal-bruit, et nous les avons requantifiées, à divers niveaux de précision.

Les résultats sont illustrés en Figs. 3.12 à 3.15, en pages 61 à 64. Ces figures étant denses, la Fig. 3.11, page 60 fournit une grille de lecture.

Globalement, sur ces exemples démonstratifs et d'autres essais non repris ici, nous pouvons remarquer :

- L'indicateur $\mathcal{N}(n)$ est moins sensible à l'amplitude du bruit que $\mathcal{Z}(n)$. Ceci rend difficile l'application du critère de définition de $\tau_{\mathcal{Z}}$.
- Les deux indicateurs sont peu sensibles aux réalisations particulières du bruit.
- Sur $\mathcal{N}(n)$, il est possible de repérer les oscillations, comme des séquences de décroissance et croissance, voir $\mathcal{E}_{\text{ss}}$ (Fig. 3.13).
- Les perturbations peuvent introduire des régressions significatives, mais celles-ci sont en nombre négligeable, voir $\mathcal{E}_{\text{cp}}$ (Fig. 3.15).
- Parmi les critères de définition de $\tau_{\mathcal{N}}$, en général PRM prendra une valeur qui correspondra à la plus petite période observable sur les données. Il coïncide la plupart du temps avec LNG. Le critère ABS donnera systématiquement des $\tau_{\mathcal{N}}$ plus grands que les deux autres critères.
- L'indicateur $\mathcal{Z}(n)$ atteint rarement la stabilisation en $n$, ce qui rend l'application de NZM problématique.

Par ailleurs, les deux indicateurs sont peu sensibles à la quantification, pour des valeurs raisonnables de celle-ci (quatre modalités est un cas périlleux).

De plus, ces indicateurs supposent la donnée d'un paramètre, qui est le risque $\alpha$. Ils sont qualitativement stables par rapport à des valeurs courantes de $\alpha < .1$. Voir annexe refanex :tprime pour plus de détails.

D'après ces résultats non exhaustifs, la détermination des temps caractéristiques $\tau_{\mathcal{N}}$ utilisant les critères PRM ou LNG est insensible au bruit et à la quantification modérée. De plus, les comportements oscillatoires sont repérés, et les échelles multiples d'évolution sont dégagées par l'indicateur $\mathcal{N}(n)$, qui est plus robuste que $\mathcal{Z}(n)$ par rapport au bruit.

## 3.4 Comparaison et Évaluation des critères

Ayant relevé les limites et les moyens d'interprétation des indicateurs définissant l'échelle caractéristique, nous sommes passés à l'application sur les données réelles.

Nous avons évalué les critères retenus par rapport à l'ensemble de données sur lequel ils ont été calculés (la "population de référence"), afin de déterminer leur spécificité aux caractéristiques de chaque paramètre physiologique. Concrètement, nous avons étudié leur dépendance par rapport à :

1. la taille de la population de référence;
2. la position dans le temps de cette population, sachant que les paramètres physiologiques présentent *a priori* un comportement dépendant des phases du séjour, de la différence entre le jour et la nuit, des modes de prise en charge thérapeutique (le mode de ventilation et la sédation en particulier);
3. la différence entre patients.

Pour ce faire, nous avons procédé à deux types d'expériences.

D'un côté nous avons comparé les échelles caractéristiques calculées sur des périodes de quatre heures pour tous les paramètres disponibles sur un ensemble d'enregistrements provenant de divers protocoles d'investigation clinique.

D'un autre côté, de façon beaucoup plus fine, nous avons calculé les échelles caractéristiques en fonction du nombre de mesures prises en compte et de leur position au sein d'un enregistrement.

Les moyens techniques dont nous disposons ne nous ont pas permis de mener des expériences exhaustives, qui considéreraient, pour tous les enregistrements, tous les paramètres à toutes les échelles depuis toutes les positions. En effet, même par les méthodes efficaces du Chap.2, ce genre de traitement nécessite des temps de calcul qui sont cubiques en la longueur des enregistrements — en fait, linéaires en le nombre de paramètres, linéaires en le nombre d'échelles considérées, linéaires en le nombre de positions, linéaires en le nombre de points d'enregistrement considéré pour l'évaluation de $\mathcal{N}(n)$ et $\mathcal{Z}(n)$.

Nous avons donc, dans un premier temps, calculé les échelles caractéristiques suivant les critères LNG, PRM et NZM pour des enregistrements de douze heures. Les Figs. 3.16 à 3.19 représentent les résultats pour quatre de ces essais, représentatifs d'un certain nombre de comportements typiques. Les échelles caractéristiques ont été recherchées pour des $n < n_{max} = 45min$, pour des raisons de capacité de calcul. Nous avons balayé *toutes* les populations de référence possibles, jusqu'à l'ensemble d'un enregistrement de douze heures. Les résultats sont représentés en niveaux de gris en fonction des deux paramètres, position et taille.

Sur la Fig 3.16, on peut remarquer l'uniformisation progressive des critères LNG et PRM au fur et à mesure des échelles croissantes, indépendamment des points de début de fenêtre considérés. Ce comportement est visible sur la Fig. 3.17, à la différence près qu'ici nous n'avons pas, en général, de maxima locaux pour $\mathcal{N}(n)$, et donc les critères LNG et PRM ne coïncident pas.

Sur la Fig. 3.20 nous voyons plus précisément comment cette uniformisation s'installe.

Les Figs. 3.19 et 3.18 mettent en relief les déficiences des données par rapport aux calculs.

En effet, pour la Fig. 3.19, la donnée présente un nombre important d'absences. Ceux-ci diminuent de façon artificielle le nombre de points de mesure sur lesquels nous pouvons opérer. Alors, même si l'uniformisation a lieu, le décompte est biaisé par les zones de données manquantes. Voir la Fig. 3.21 pour les détails de ce comportement.

Pour la donnée $P_{aw}min$, les résultats sont représentés en Fig. 3.18. Cette donnée ne présente pas seulement des absences — elle n'est présente que pendant les premières six heures de l'enregistrement — mais aussi une quantification extrême : seulement quatre valeurs sont présentes.

La recherche des échelles caractéristiques a été menée de façon exhaustive sur douze jours-patient. Nous avons remarqué précédemment que l'échelle caractéristique se stabilise quelle que soit la position de calcul, pourvu que la population d'apprentissage comprenne plus de deux heures d'enregistrement.

Nous avons donc cherché à déterminer la variabilité de l'échelle caractéristique d'après :
- la donnée physiologique sous-jacente,
- le patient,
- le moment de la journée.

Nous avons découpé douze jours-patient en parts de quatre heures, puis avons calculé $\tau$ pour chaque variable, suivant les critères LNG (Fig. 3.22), PRM (Fig. 3.23) et NZM (Fig. 3.24).

Les résultats apparaissent en pages 71 à 73.

Nous avons éliminé les paramètres physiologiques dont la présence est sporadique et les variables non physiologiques. A partir de ces résultats nous pouvons caractériser chaque donnée en terme de variabilité de $\tau$.

- Les paramètres physiologiques issus des moniteurs cardio-vasculaires présentent une faible variabilité vis-à-vis des positions dans le temps des enregistrements. Lorsque variabilité il y a, elle est contemporaine des périodes de jour et de nuit. Ceci est vrai, même en calculant l'échelle caractéristique à partir de NZM, ce qui est hasardeux.
- Les mesures de capnographie présentent de même des temps caractéristiques proches la plupart du temps, mieux groupés par LNG que par PRM.
- Les échelles caractéristiques des paramètres respiratoires présentent une variabilité plus importante, avec des $\tau$ supérieurs à ceux concernant le système cardio-vasculaire. Pour ces paramètres, le critère NZM n'est pas appliquable.
- Les paramètres issus de l'analyse des gaz du sang en continu présentent des temps caractéristiques plus importants que les paramètres cardio-vasculaires, avec une dispersion plus grande,*c-à-d.* une plus grande variabilité entre patients.

## 3.5 Limites du principe

Le principe de séparation court terme *vs.* long terme suppose qu'il soit "raisonnable" d'agir de la sorte, *c-à-d.* cela est lisible et apporte une interprétation fructueuse par la suite. Par exemple,

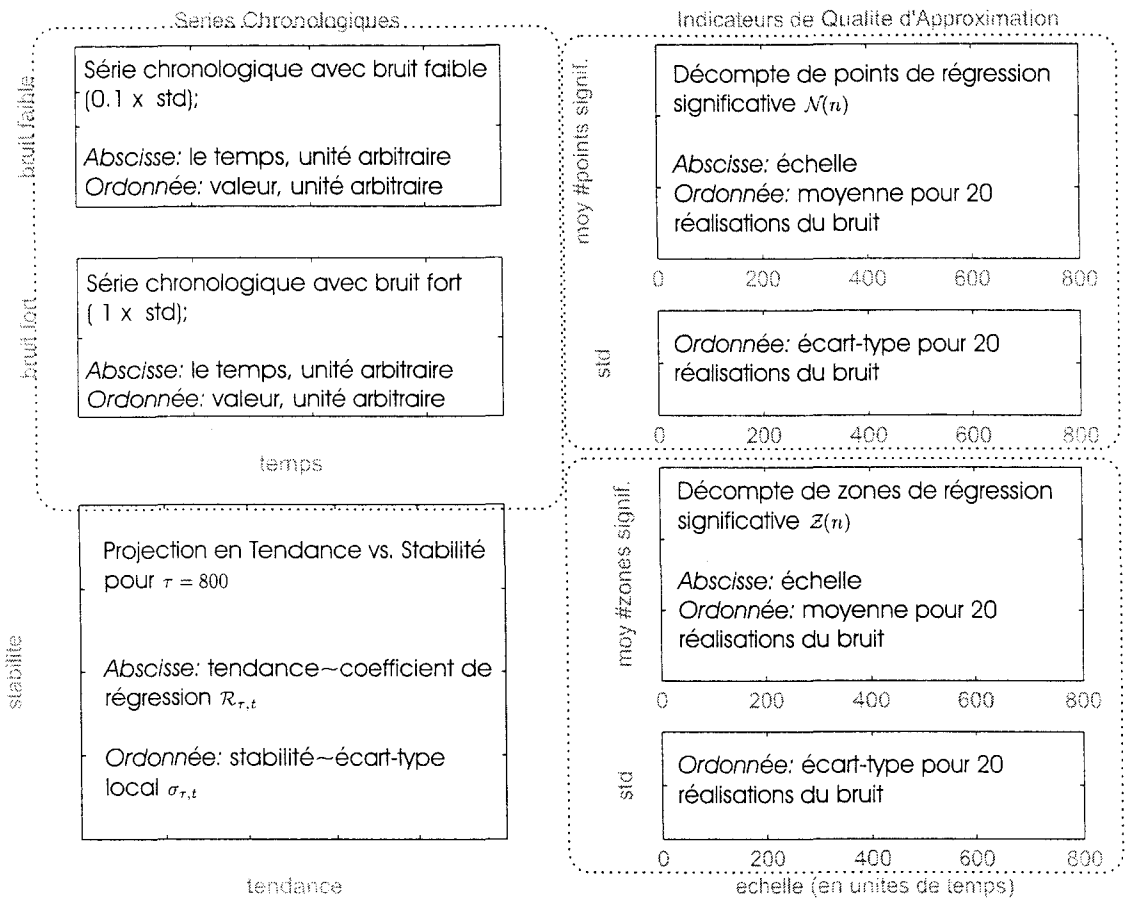



FIG. 3.11 – Recherche d'échelles caractéristiques : légende

En haut à gauche, les séries chronologiques peu et fortement bruitées.

Le bruit représenté est additif gaussien, de moyenne nulle et d'écart-type $0.1\sigma$ (faible) et $\sigma$ (fort), avec $\sigma$ l'écart-type des données. Les pointillés correspondront aux données les plus bruitées. La quantification pour tous ces exemples produit une précision de 12.5%.

En bas à gauche la projection en tendance *vs.* stabilité des deux séries, pour les échelles les plus grandes.

A droite, à partir du haut : moyenne de $\mathcal{N}(n)$, écart-type de $\mathcal{N}(n)$, moyenne de $\mathcal{Z}(n)$, écart-type de $\mathcal{Z}(n)$. Ces moyennes et écarts-types correspondent à vingt réalisations différentes du bruit.

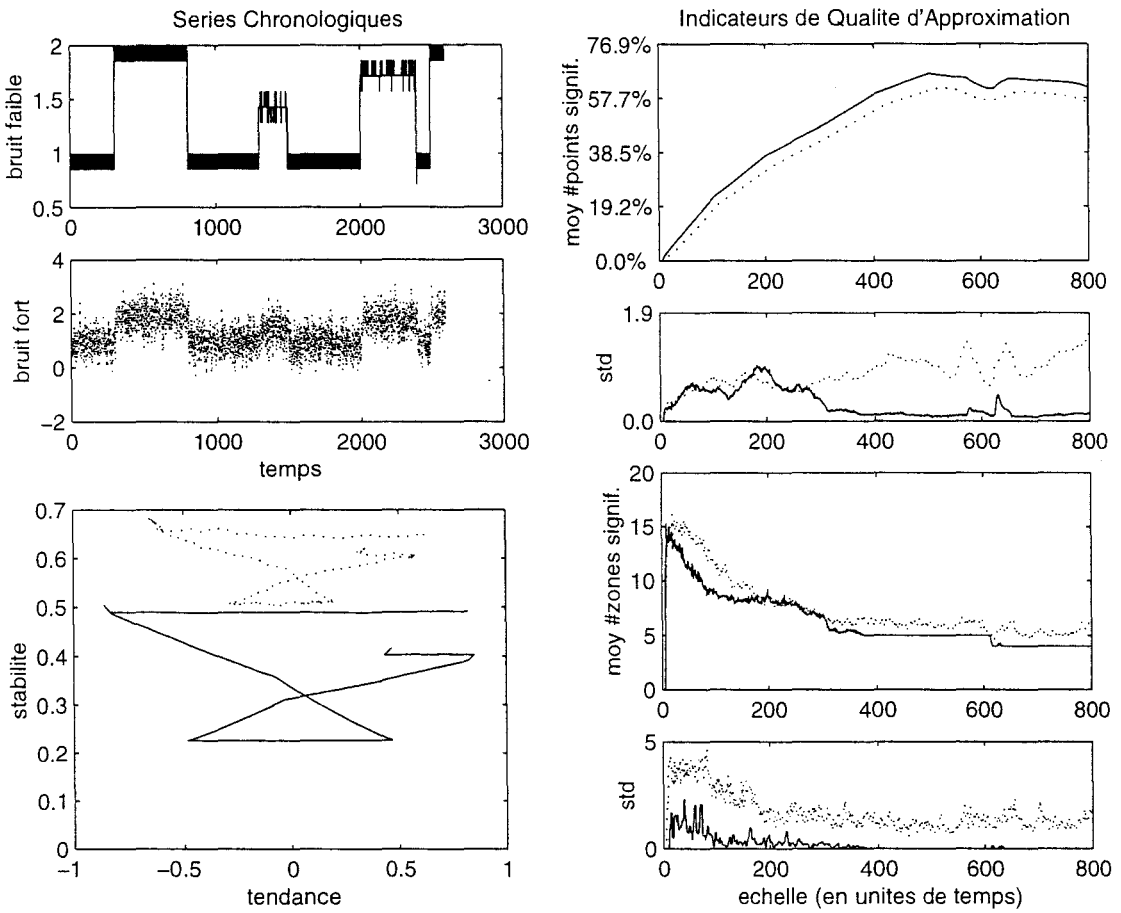



FIG. 3.12 – Recherche d'échelles caractéristiques pour $\mathcal{E}_{\mathrm{cm}}$

Voir Fig. 3.11 pour une légende.

Les projections en tendance *vs.* stabilité (en bas à gauche) montrent le comportement typique en présence de bruit : décalage vers les écarts-types croissants, diminution de la plage de variation en tendance. Les tracés on été faits, pour la clarté de la figure, avec des lignes continues. En fait les points pour lesquels le calcul de $\mathcal{R}_{n,t}$ n'est pas possible ne sont pas représentés. C'est par exemple le cas lorsque $\sigma_{n,t}$ est nul.

L'indicateur $\mathcal{N}(n)$ est peu sensible au bruit, contrairement à $\mathcal{Z}(n)$.

La décroissance de $\mathcal{N}(n)$ vers $n \sim 500$ correspond à une période approximative d'apparition des décalages sur les séries chronologiques.

Les diverses réalisations de bruit additif ne perturbent pas énormément le comportement de $\mathcal{N}(n)$, comme il apparaît des tracés des écarts-types (droite, deuxième figure à partir du haut).

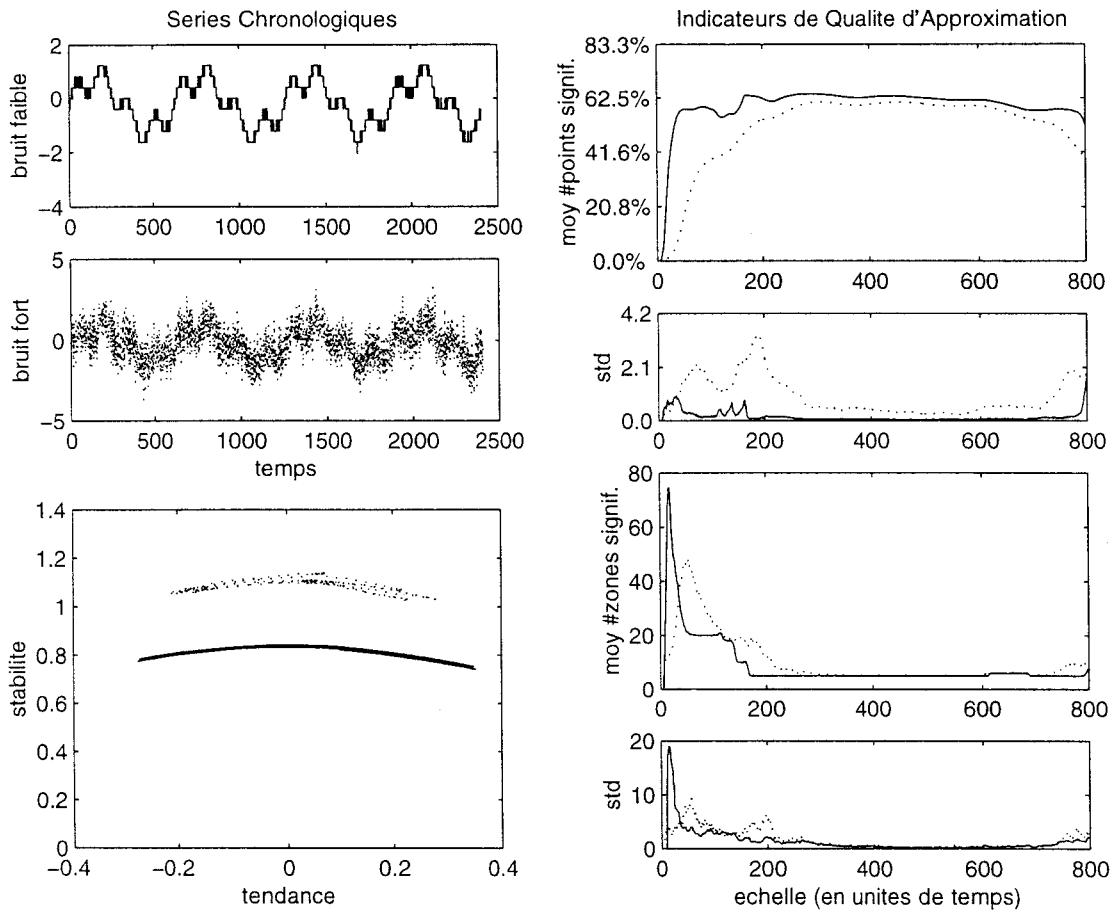



FIG. 3.13 – Recherche d'échelles caractéristiques pour $\mathcal{E}_{ss}$

Voir Fig. 3.11 pour une légende.

Les oscillations se présentent sur le plan tendance *vs.* stabilité comme des allers-retours de tendance, à un niveau de stabilité sensiblement constant (mais variant directement avec la tendance).

Sur le tracé de $\mathcal{N}(n)$ et $\mathcal{Z}(n)$, on peut repérer l'oscillation de plus faible amplitude, qui se retrouve noyée dans le bruit fort.

La décroissance de $\mathcal{N}(n)$ vers les grandes échelles correspond à la longueur des cycles de l'oscillation de plus basse fréquence.

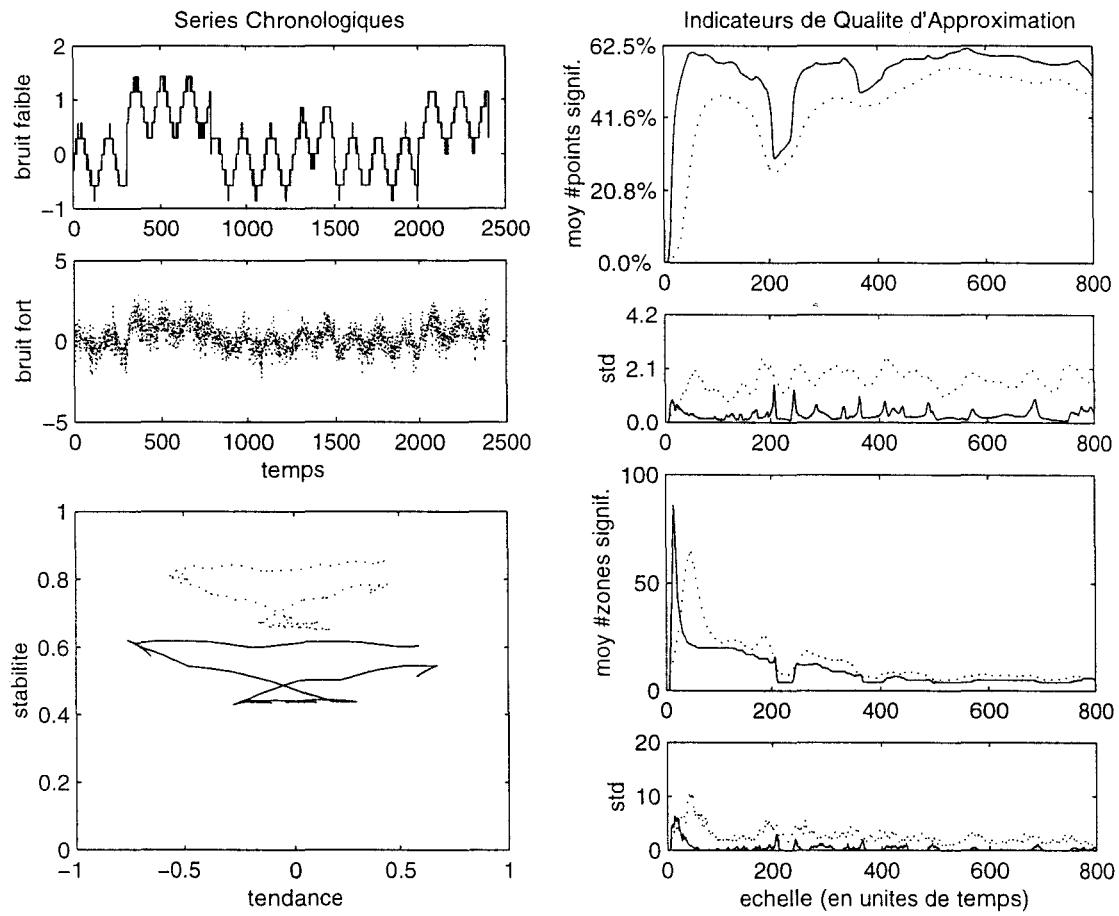



FIG. 3.14 – Recherche d'échelles caractéristiques pour $\mathcal{E}_{\mathrm{sd}}$

Voir Fig. 3.11 pour une légende.

Les parties "plates" du tracé de la projection tendance *vs.* stabilité correspondent aux oscillations ; les extrêmes aux changements de ligne de base.

Les décroissances de $\mathcal{N}(n)$ et $\mathcal{Z}(n)$ vers $n \sim 200$ correspondent à la longueur d'un cycle de l'oscillation.

La sensibilité au bruit est faible pour les deux indicateurs.

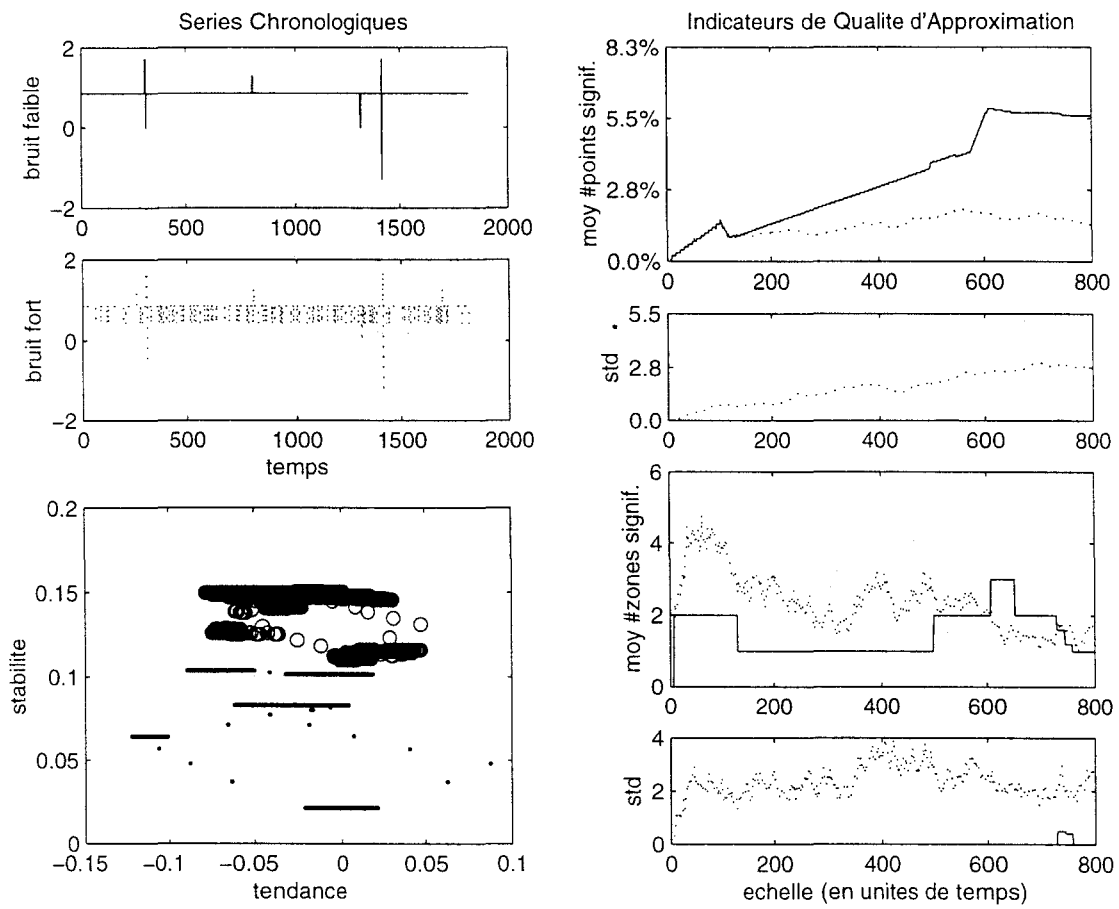



FIG. 3.15 – Recherche d'échelles caractéristiques pour $\mathcal{E}_{\text{cp}}$

Voir Fig. 3.11 pour une légende.

Nous avons tracé sur la projection tendance *vs.* stabilité uniquement les points (cercles pour bruit fort, points pour bruit faible) où le calcul de $\mathcal{R}_{\tau,t}$ est possible : il n'y en a pas beaucoup. Ce sont les points autour desquels il y a effectivement variation de la donnée.

On retrouve cette anomalie sur le tracé de $\mathcal{N}(n)$, où la part de régressions significatives est minime. En présence de bruit, elle devient en moyenne de l'ordre de grandeur de son écart-type. C'est aussi le cas de $\mathcal{Z}(n)$.

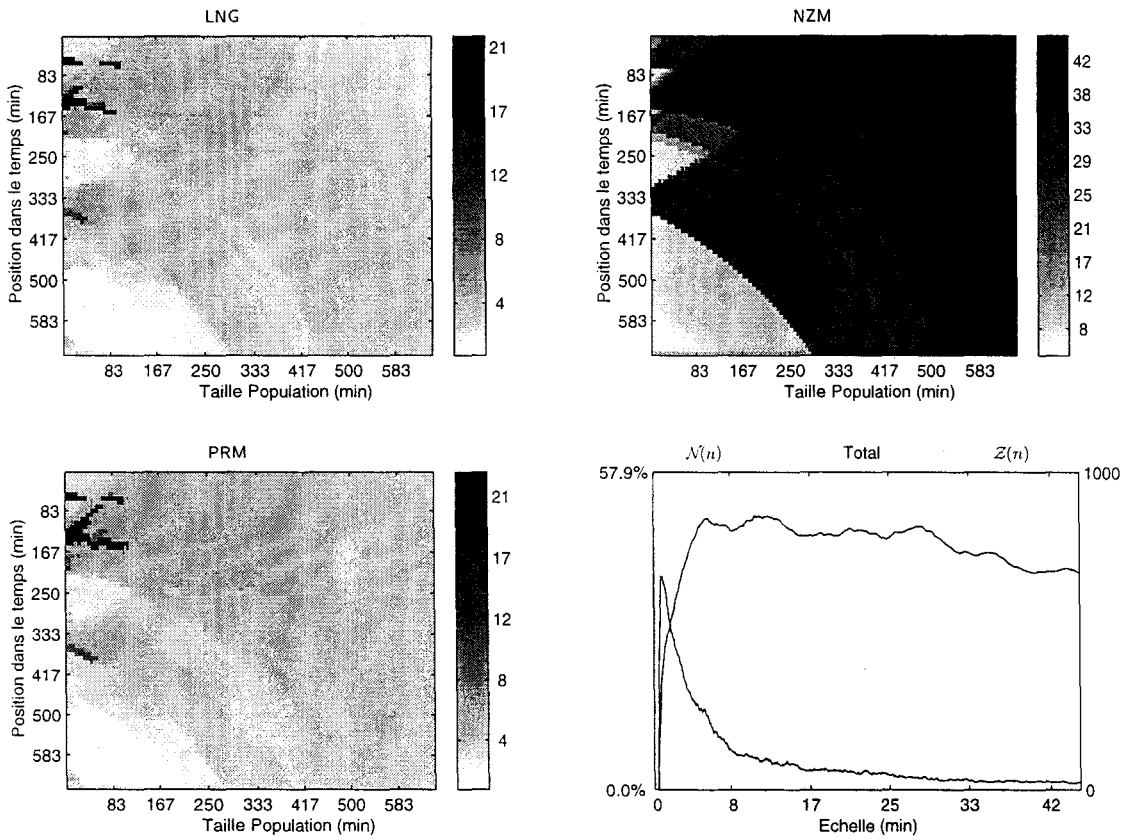



FIG. 3.16 – Évaluation des critères PRM, LNG et NZM pour $F_c$

En haut à gauche, les échelles caractéristiques obtenues par application du critère LNG. En abscisse, la taille de la population de référence ; en ordonnée, la position dans le temps du début de cette population. Nous avons représenté, à la manière de la décomposition en ondelettes (*cf.* 6.2 pour plus de détails et des références), les valeurs de $\tau$ comme comprenant tous les points de la population de référence. Ceci introduit visuellement les notions d'échelle et de granularité temporelle.

En bas à gauche, l'application du critère PRM. La similitude entre les résultats fournis par ce critère et LNG sont remarquables.

En haut à droite, l'application du critère NZM. Les échelles considérées sont systématiquement plus grandes que pour les deux autres, mettant en relief l'aspect pessimiste de cette approche.

En bas à droite, pour référence, les caractéristiques $\mathcal{N}(n)$ et $\mathcal{Z}(n)$ calculées sur les douze premières heures d'enregistrement.

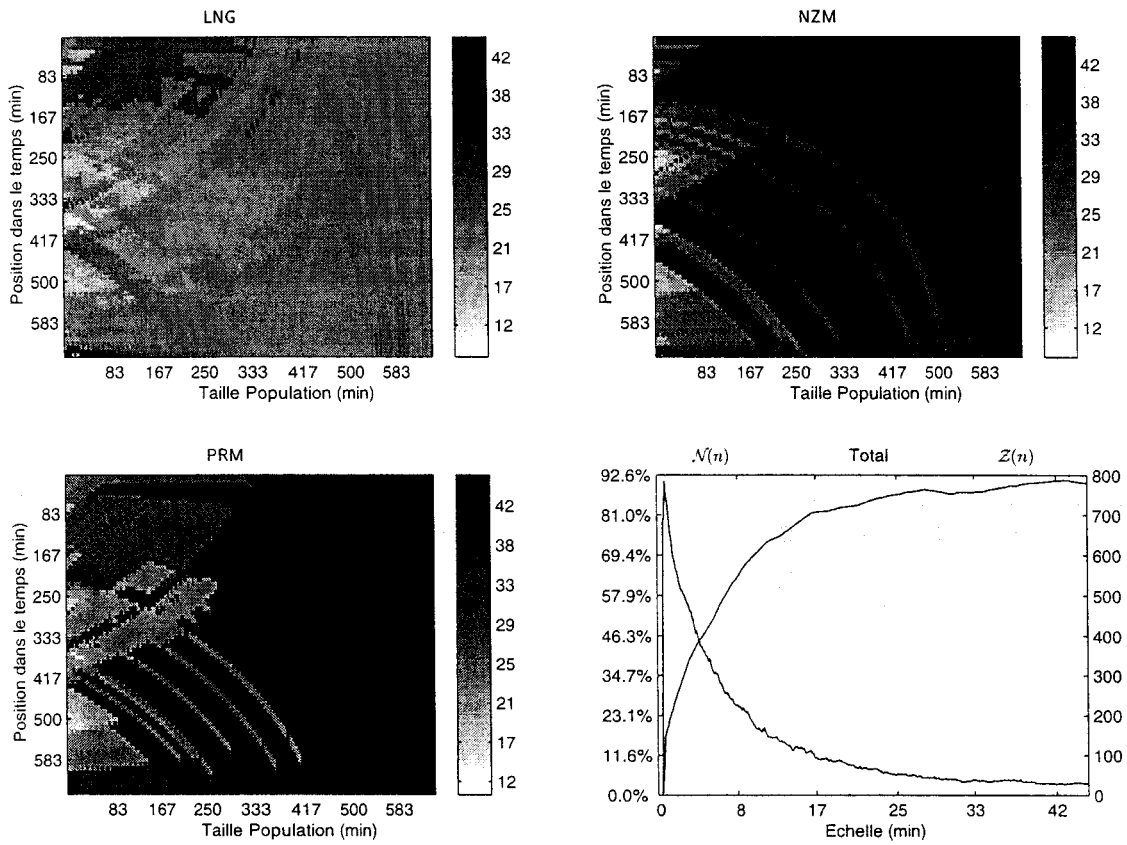



FIG. 3.17 – Évaluation des critères PRM, LNG et NZM pour $T°$

Pour le paramètre $T°$ de cet enregistrement, les échelles caractéristiques définies par PRM sont plus pessimistes que pour LNG. Ceci est dû à la courbe de $\mathcal{N}(n)$, qui ne présente pas de maxima locaux. Le critère LNG choisira alors une échelle caractéristique par rapport au maximum absolu, alors que PRM ne retrouvera pas cette échelle.

En bas à droite, pour référence, les caractéristiques $\mathcal{N}(n)$ et $\mathcal{Z}(n)$ calculées sur les douze premières heures d'enregistrement.

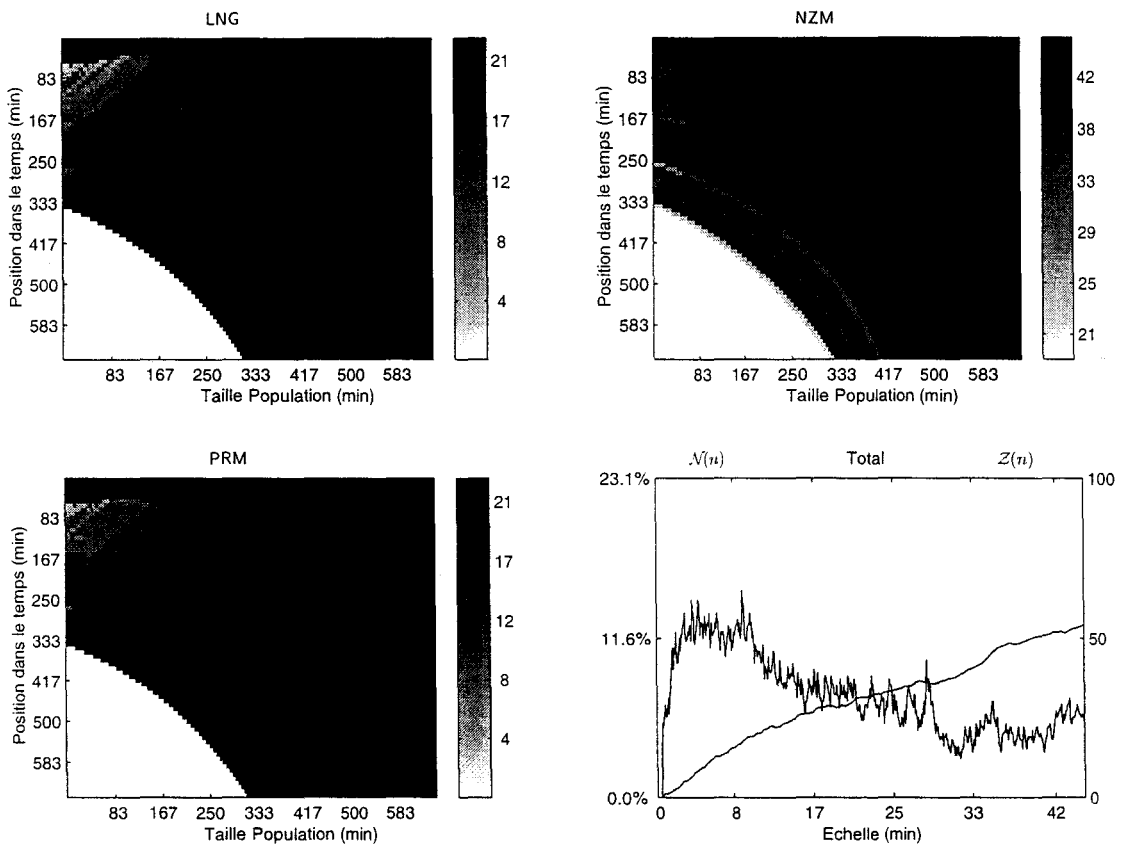



FIG. 3.18 – Évaluation des critères PRM, LNG et NZM pour $P_{aw}min$

Même si les résultats pour LNG et PRM se rejoignent, pour des échelles suffisamment grandes, le nombre important de données manquantes (les zones blanches en bas à gauche des figures correspondant aux trois critères) ne permet de définir aisément une échelle unique.

En bas à droite, pour référence, les caractéristiques $\mathcal{N}(n)$ et $\mathcal{Z}(n)$ calculées sur les douze premières heures d'enregistrement. La caractéristique essentiellement croissante dans les faibles valeurs est typique de données au comportement aberrant : très forte quantification en l'occurrence.

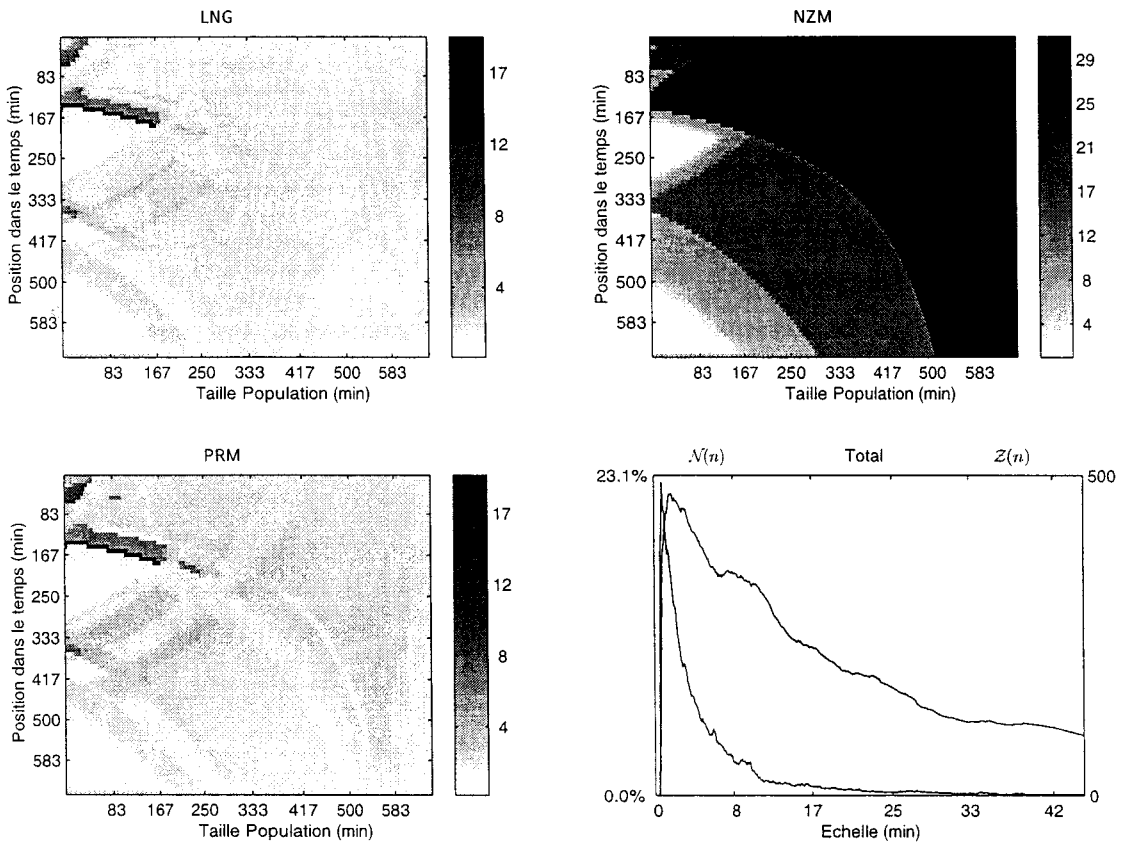



FIG. 3.19 – Évaluation des critères PRM, LNG et NZM pour $SpO_2$

En bas à droite, pour référence, les caractéristiques $\mathcal{N}(n)$ et $\mathcal{Z}(n)$ calculées sur les douze premières heures d'enregistrement. La décroissance atypique est un artefact de la présence de nombreuses données manquantes.

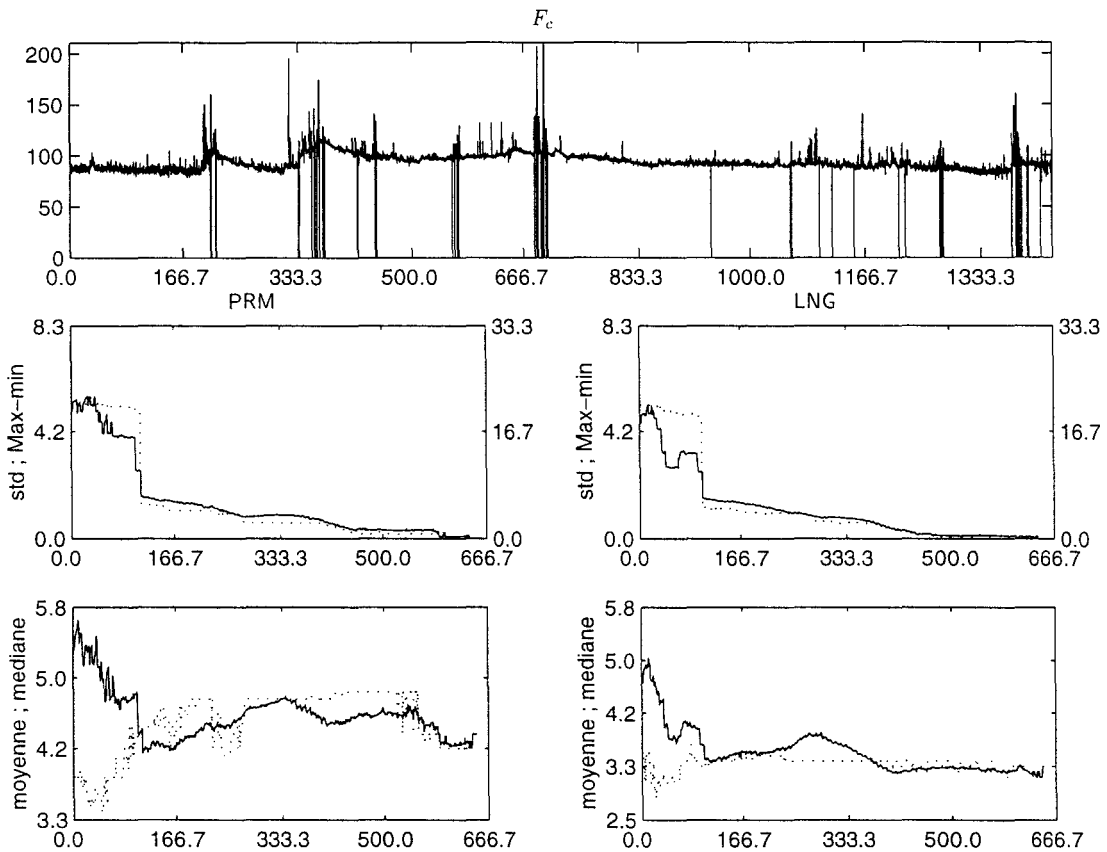



FIG. 3.20 – Uniformisation des échelles caractéristiques par PRM et LNG pour $F_c$

En haut, la série chronologique de $F_c$. Seules les douze premières heures ($720min$) sont prises en compte, le reste étant représenté pour relativiser visuellement la continuité des données. L'unité en ordonnée est $min^{-1}$, l'abscisse est donnée en minutes. L'enregistrement débute à zéro heures GMT.

A gauche (*resp.* droite), les échelles caractéristiques définies par PRM (*resp.* LNG), en minutes, en fonction de la taille de la population de référence utilisée pour le calcul. En haut, les écarts-types (trait plein, échelle de droite) et les rangs de variation (pointillés, échelle de gauche), en bas la moyenne (trait plein) et la médiane (pointillés) calculées sur tous les points de début de la population de référence.

L'uniformisation des valeurs pour les deux critères est visible par la chute brutale des écarts-types et rangs de variation. C'est à cette taille que les médianes et moyennes coïncident.

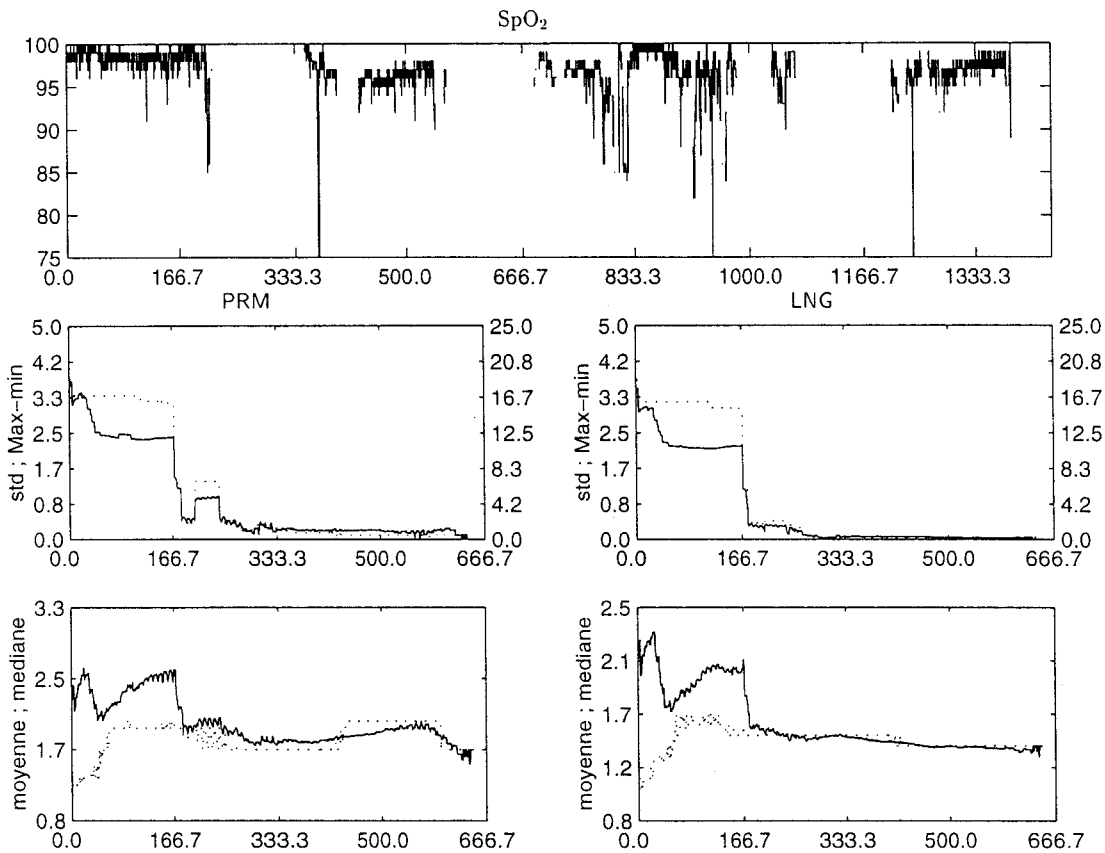



FIG. 3.21 – Uniformisation des échelles caractéristiques par PRM et LNG pour $SpO_2$

Voir Fig. 3.20 pour une légende détaillée.

Les données manquantes de l'enregistrement sont mis en évidence en haut. Unité : pourcent, abscisse : temps en minutes depuis 00:00GMT. Il y a des épisodes d'absence d'une centaine de minutes (*peut-être* des débranchements de capteur), mais aussi des intermittences (vers $670min$ par exemple, *peut-être* des agitations). L'origine de ces deux formes d'absence n'est pas documentée. La chute des écarts-types et des rangs de variation montre l'uniformisation des échelles caractéristiques pour des tailles de population de référence supérieures à deux ou trois heures.

Ce comportement est exacerbé par des artefacts générées par les données manquantes : la décroissance vers $170min$ d'échelle correspond aussi à la taille moyenne des épisodes de présence et d'absence des données.

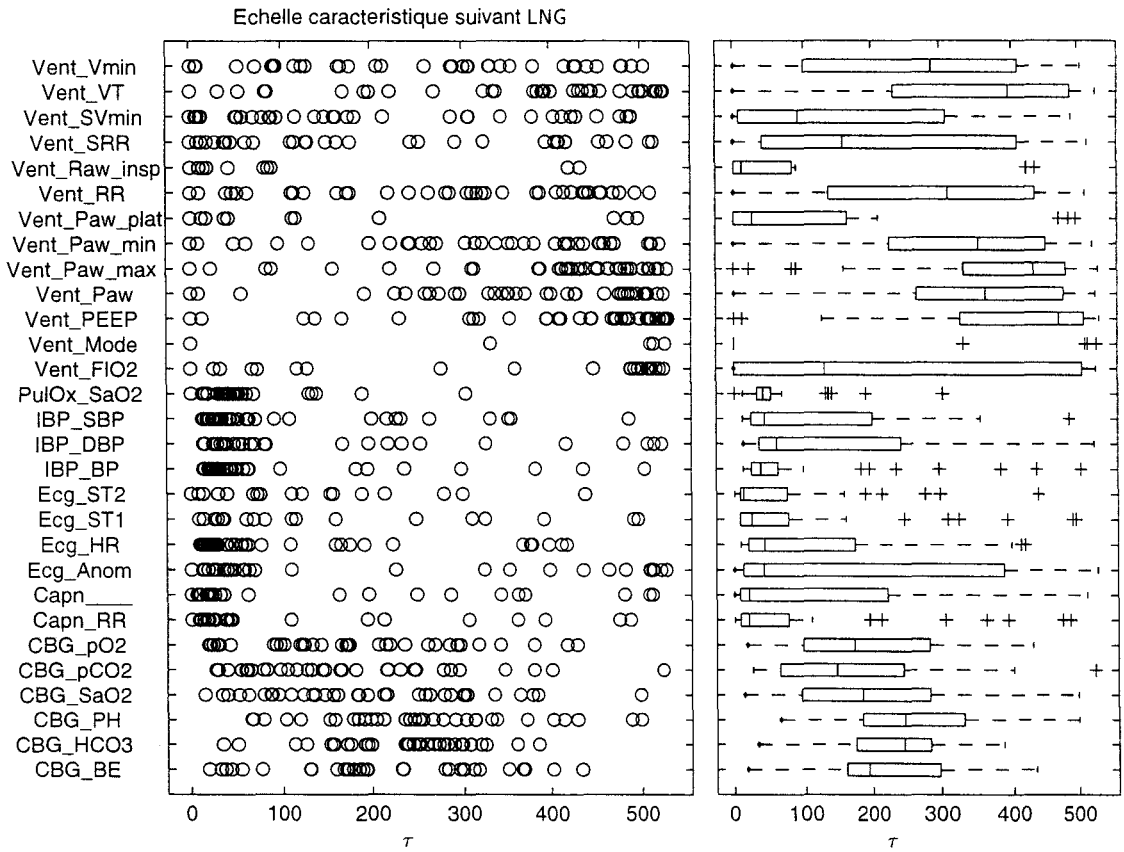



FIG. 3.22 – Échelles caractéristiques suivant LNG

Ligne à ligne, le nom du paramètre physiologique. En abscisse, les échelles caractéristiques trouvées pour chaque tranche de quatre heures sur douze jours-patient. À gauche, les cercles représentent chaque échelle caractéristique ; à droite, le diagramme en boîtes (*boxplot* [Tuk77]) résumant ces distributions. Les paramètres respiratoires sont préfixés par Vent_ ; les mesures de gaz du sang en continu sont préfixées par CBG_. Les autres mesures sont réalisées par le moniteur cardio-vasculaire.

Les variabilités sont plus grandes pour les variables respiratoires. Les valeurs médianes sont plus faibles pour l'ensemble des paramètres cardio-vasculaires. On retrouve les données au comportement aberrant : $FIO_2$, $PEEP$, et le mode ventilatoire, qui est un codage numérique du mode ventilatoire — il aurait pu être enlevé, en tout état de cause. Les échelles caractéristiques de valeur 1 ou proche de 540 (qui est la plus grande échelle que nous nous sommes fixés, soit $45min$) représentent les cas de calcul impossible et/ou d'inadaptation des critères, et du principe de l'échelle caractéristique, aux données.

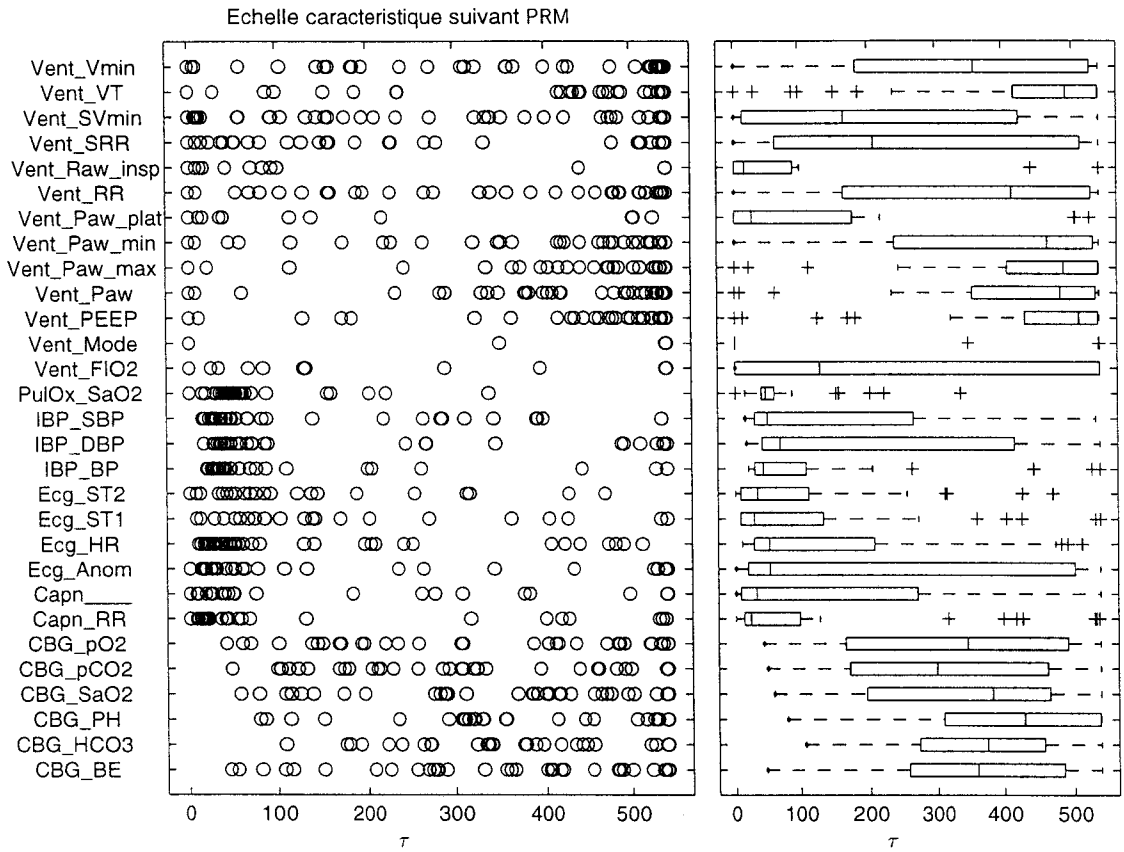



FIG. 3.23 – Échelles caractéristiques suivant PRM

Les échelles caractéristiques repérées par PRM coïncident en général avec celles repérées par LNG. Les $\tau$ calculés sur les mesures de gaz du sang en continu (les données préfixées par CBG_) sont plus dispersés que pour LNG. Les paramètres physiologiques présentant la variabilité de $\tau$ la plus faible sont ceux issus des moniteurs cardio-vasculaires. Les paramètres ventilatoires sont assez bien concentrés vers les grands $\tau$, mais il s'agit davantage d'un artefact lié au critère. Les points à $\tau = 1$ et $\tau \sim 540$ ne sont pas représentatifs du comportement réel de $\mathcal{N}(n)$ : le critère est ici mis en défaut.

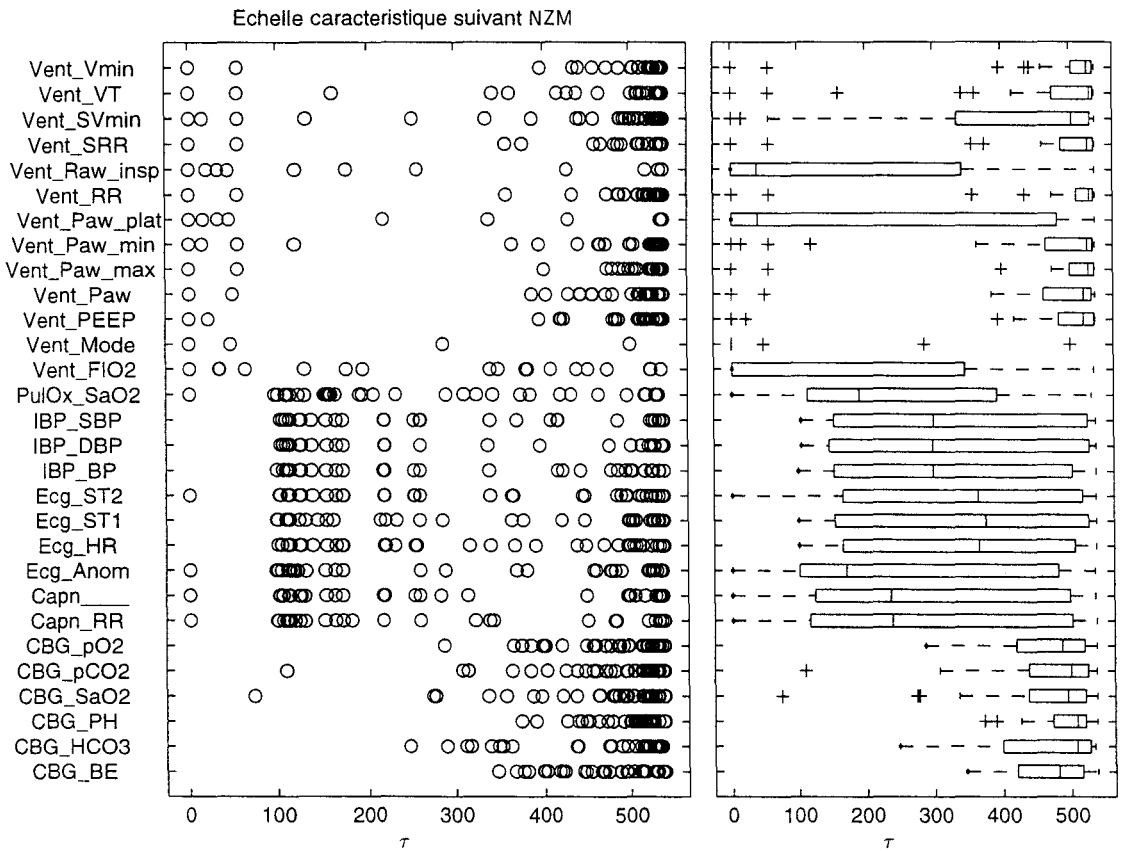



FIG. 3.24 – Échelles caractéristiques suivant NZM

Les échelles caractéristiques sont concentrées vers les $\tau$ les plus grands, ce qui montre la dépendance très forte de ce critère par rapport à l'échelle maximale de calcul. Seuls les paramètres cardio-vasculaires présentent des groupements vers des $\tau \sim 100$ qui peuvent être représentatifs. Le nombre de points à $\tau = 1$ et cette dépendance par rapport à l'échelle maximale montrent les difficultés d'application de ce critère, pourtant bien fondé. Il faudrait procéder à des lissages supplémentaires sur $\mathcal{Z}(n)$ pour obtenir un $\tau_{\mathcal{Z}}$ représentatif.

c'est une opération utile pour le filtrage de données bruitées, même si elle introduit un déphasage — "bruit" voulant dire composante additive stochastique indépendante. Elle permet aussi de retrouver l'information de dérivée pour des données bruitées et quantifiées. La classe de séries chronologiques pour lesquelles cette opération a du sens inclut les séries linéaires par morceaux, bruitées et quantifiées. Pourtant, toutes les données que l'on traite n'appartiennent pas à cette classe.

Par exemple, les données constantes par morceaux, ou trop sévèrement quantifiées par rapport à leur composante bruitée, posent problème : peut-on vraiment parler de tendance pour de telles données ? On peut argumenter qu'à des fins d'explication, le fait d'avoir changé de niveau constitue une variation qui possède un sens,*c-à-d.* une tendance. Cette interprétation se fait en considérant les instants passés comme lointains ; c'est ce que nous essayons de formaliser dans la notion de long terme. Cette variation doit donc être interprétée en termes de changement de valeur, et non pas en terme de tendance, notion qu'on gardera dans le sens de tendance *locale.*

D'autres cas particuliers peuvent aussi venir enrayer le principe de la méthodologie. Lorsque la quantification est extrême (quelques modalités), $\tau$ n'est pas un indicateur fiable de la variabilité de la donnée. Dans ce cas il faut proposer d'autres indicateurs basés sur, par exemple, la distribution des différences et l'histogramme de celle-ci.

Un cas plus rare mais réel est celui des données qui sont actualisées à des fréquences plus basses que celle d'échantillonnage. Dans ce cas (qu'on peut repérer facilement), il faut opérer un sous-échantillonnage pour que $\tau$ reflète la dynamique réelle de la donnée *à son échelle d'évolution.*

Malgré ces cas particuliers, qui doivent être repérés avant de procéder au calcul de $\tau$, l'indicateur semble effectivement être caractéristique de chaque paramètre physiologique, pourvu qu'il soit calculé sur une période supérieure à deux heures.

## 3.6 Conclusion : une échelle effectivement caractéristique

Nous avons proposé de traiter le problème du choix de l'échelle de filtrage en cherchant *une* échelle, caractéristique de chaque paramètre, à laquelle le filtrage puisse être opéré.

Nous avons proposé divers indicateurs et nous avons comparé les résultats obtenus par un certain nombre de critères sur ces indicateurs. Il résulte de nos expérimentations que le critère LNG défini à partir du décompte de points de régression significative $\mathcal{N}(n)$ est différencié suivant les sous-systèmes physiologiques considérés. Il est stable entre patients pour les paramètres hémodynamiques en particulier, pourvu qu'il soit calibré sur une durée de l'ordre de deux à trois heures.

En choisissant le critère LNG, nous avons une échelle caractéristique $\tau_{\mathcal{N}}$ pour chaque paramètre qui permet d'opérer un filtrage linéaire, équivalent de la recherche d'une approximation de la dérivée de la série chronologique en tout point. A l'échelle caractéristique ainsi définie, il est tolérable d'approcher localement la tendance de $y_t$ par $\mathcal{S}_{\tau_{\mathcal{N}},t}$. La variance locale $\sigma^2_{\tau_{\mathcal{N}},t}$ fournit en même temps un indicateur de la qualité d'approximation et condense la déviation de la tendance linéaire en un seul indicateur d'erreur, qui peut être interprété comme un indicateur de stabilité.

Par la suite nous détaillons cette décomposition et ses propriétés, dans le but d'établir une transformation symbolique monovariée des données filtrées à l'échelle caractéristique.

*Penser c'est être expert en catégories.*

Ludwig Wittgenstein

# 4

# Transformation numérique-symbolique

Où le calcul de tendance à l'échelle caractéristique permet de proposer des transformations des données numériques en symboles représentant la dynamique.

## Sommaire

Nous avons défini et adopté une façon de filtrer nos données après avoir déterminé une échelle propre à chaque flux de données. Ce filtrage produit un indicateur numérique de la *tendance*, assimilable à la dérivée du signal. En plus, l'erreur d'approximation commise fournit un indicateur de l'exactitude de cette assimilation. Cet indicateur peut être interprété en termes de *stabilité*.

Nous avons ainsi, pour chaque donnée en tout point du temps et en plus de la valeur, deux indicateurs représentant deux notions centrales dans l'expression commune de l'évolution d'un paramètre physiologique : la tendance et la stabilité.

A partir de ces indicateurs numériques, nous essayons d'opérer une quantification adéquate, qui permettra d'en dériver des *symboles*. Ceux-ci sont les données de base sur lesquelles nous construirons par la suite des modèles symboliques par apprentissage automatique.

Nous allons proposer une forme de visualisation à partir de ces indicateurs, puis nous allons détailler un certain nombre de façons d'opérer la transformation numérique-symbolique. Les critères sur lesquels elle se base sont empiriques, et nous allons retrouver ici les limites de la méthodologie basée sur les données : nous n'aurons pas de moyens immédiats de valider ou de comparer les différentes transformations possibles.

Au fur et à mesure, nous illustrerons les approches sur un signal artificiel ("SBQD") qui est un signal sinusoïdal avec des décalages de sa composante continue. Il a été, à l'instar des traitements subis au sein des moniteurs, bruité puis quantifié. Nous utiliserons de même à titre illustratif, un signal réel, "72040:$SpO_2$", qui est la donnée de la saturation en oxygène par oxymétrie de pouls mesurée sur un patient pris au hasard (le même qu'au chapitre précédent), pendant $7h25mn$. Ces deux séries chronologiques sont représentées en Fig. 4.1.

## 4.1 Projection en tendance *vs.* stabilité

Comme produit direct de la décomposition en valeur, tendance et stabilité, nous pouvons proposer une représentation visuelle des séries chronologiques. Représenter une série chronologique $y_t$ directement dans l'espace bidimensionnel $(\mathcal{S}_{\tau_N,t}\,,\,\sigma_{\tau_N,t})$ permet d'avoir une représentation condensée dans le temps du comportement des paramètres physiologiques pris un par un.

L'interprétation de $\sigma_{\tau_N,t}$ comme stabilité du paramètre et de $\mathcal{S}_{\tau_N,t}$ comme tendance permet de repérer des séquences de continuité de ces états, que l'on peut déterminer visuellement dans une analyse exploratoire, prérequis utile à toute analyse ultérieure [WG94, Tuk77]. En effet, un certain nombre de comportements-type (comme l'oscillation, les impulsions, les changements de plateau) donnent sur cette représentation des formes typiques, faciles à identifier.

A titre d'illustration, les représentations du signal SBQD et du signal 72040:$SpO_2$ sont données en Fig.4.2 et Fig.4.3. Le lissage introduit par le filtre linéaire permet de mieux repérer visuellement les comportements [Tuk77, SSN+93].

Il est en plus possible d'exploiter pour la représentation les dimensions supplémentaires que fournit la couleur. Nous pouvons par la couleur réintégrer les deux dimensions qui sont absentes :

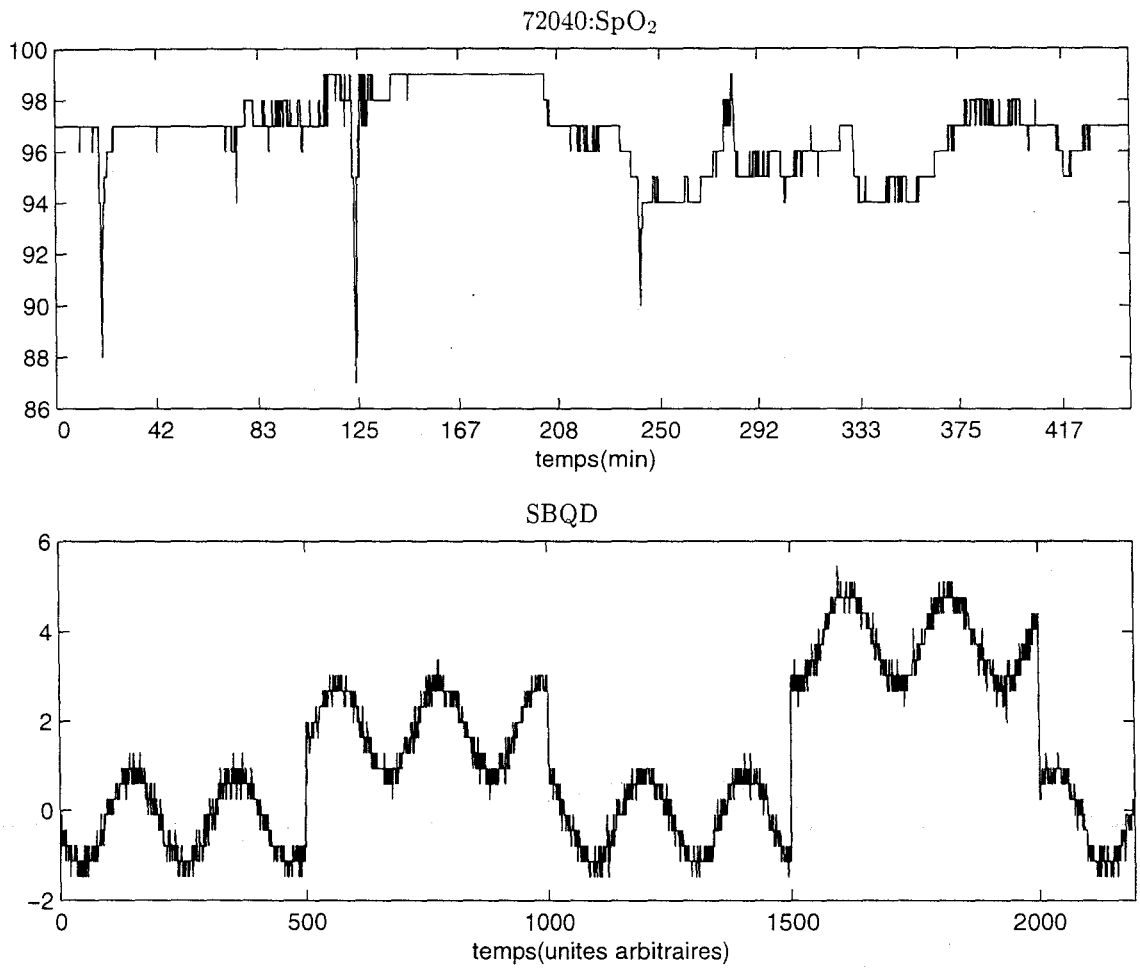



FIG. 4.1 – Les signaux d'illustration :72040:$SpO_2$ et SBQD
En haut, une mesure d'oxymétrie de pouls, 72040:$SpO_2$. L'abscisse est donnée en points de mesure; la période d'échantillonnage est de $5s$, mais cela n'intervient pas de façon déterminante dans la suite. En bas, un signal artificiel : un sinus décalé, bruité et quantifié SBQD.

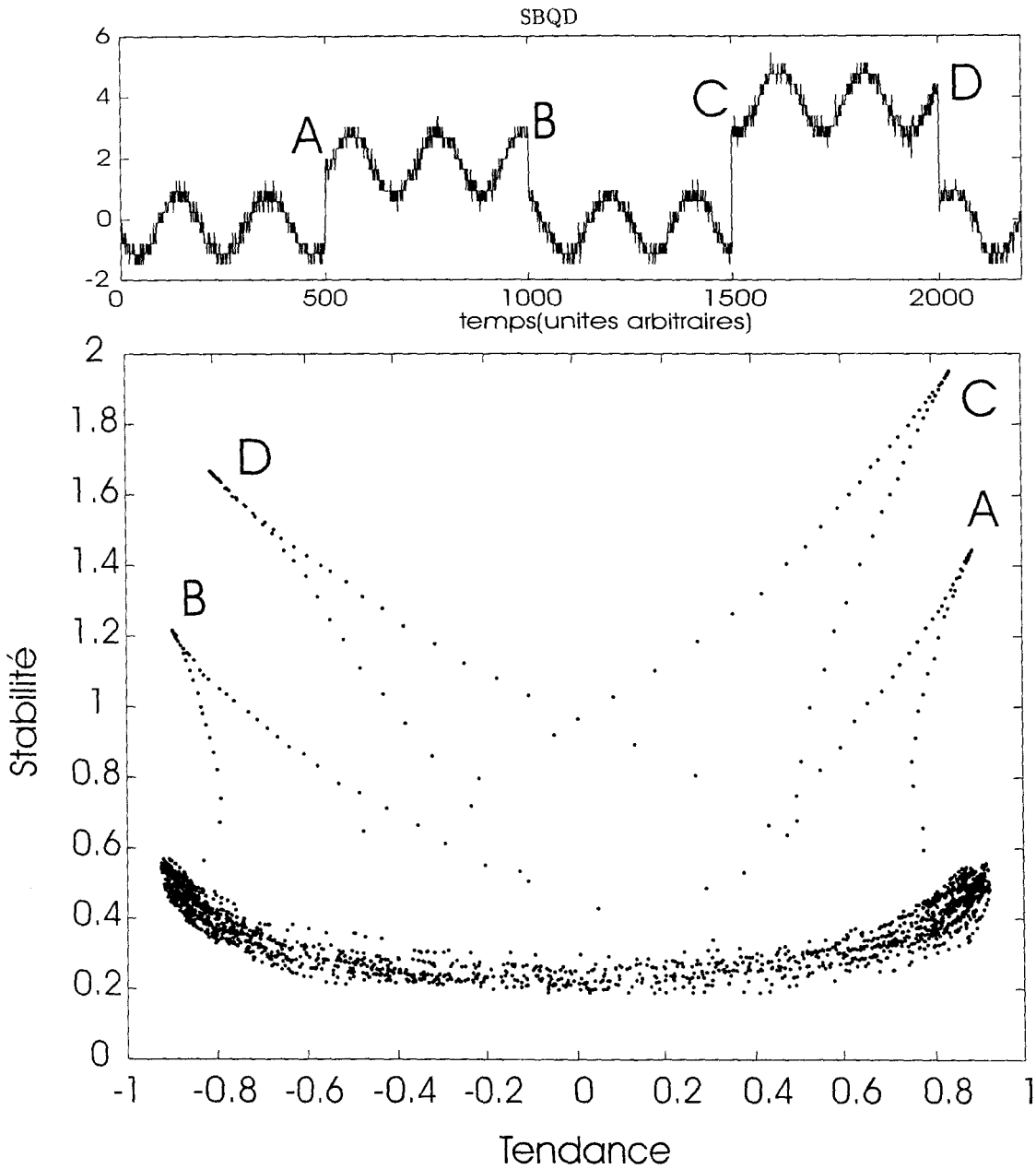



FIG. 4.2 – Exemple de projection en tendance *vs.* stabilité du signal SBQD

En haut, pour référence, la série chronologique de SBQD.

En bas, la projection tendance *vs.* stabilité de SBQD. En abscisse, le coefficient de corrélation $\mathcal{R}_{n,t}$, en ordonnée l'écart-type local $\sigma_{n,t}$.

Les quatre "pics" correspondent aux décalages de la ligne de base du signal, identifiés par A, B, C et D sur les séries chronologiques ; le comportement oscillatoire est concentré dans les basses variances. Les points de régression non significative ne sont pas représentés.

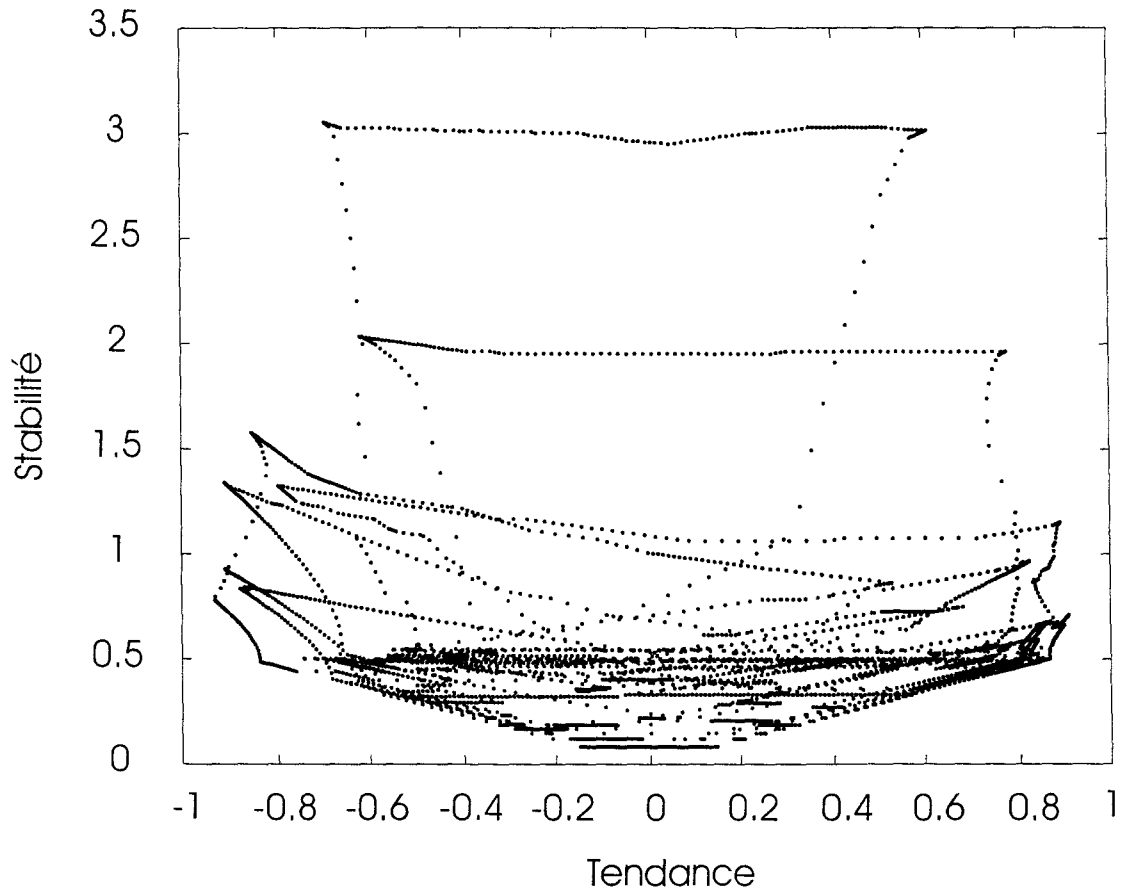



FIG. 4.3 – Exemple de projection en tendance *vs.* stabilité du signal 72040:$SpO_2$
Le coefficient de corrélation (en abscisse) permet, par la normalisation par rapport au domaine de variation, de comparer les tendances de plusieurs paramètres. Dans cet enregistrement on peut remarquer des comportements de changement de ligne de base, correspondant aux "pics"en forme de $\Lambda$, et une perturbation de type impulsionel, correspondant à la trajectoire dans les forts écarts-types en forme de $\Pi$.

la valeur et le temps. Une échelle de tonalités permet de représenter les valeurs du paramètre indépendamment de la position de celui-ci dans le plan tendance *vs.* stabilité. Une échelle d'intensités ou de saturations permet de représenter, de façon indépendante, le temps — au moins en termes de distance vers le passé. Ceci permet en même temps de repérer les valeurs extrêmes des paramètres, et d'introduire une notion d'horizon de mémoire.

Nous représenterons alors quatre dimensions sur le plan tendance *vs.* stabilité. Pour chaque variable $Y$, à l'instant $t$ et ayant défini l'échelle caractéristique $\tau$, nous représentons :

Abscisse Le coefficient de régression $\mathcal{R}_{\tau,t}$. Il a, par définition, fait l'objet d'une normalisation par rapport à sa plage de variation et à son unité.

Ordonnée L'écart-type $\sigma_{\tau,t}$ des mesures $y_t$ à $y_{t-\tau+1}$. Celle-ci est indépendante de $\tau$ et représente la stabilité locale de la variable. On peut utiliser aussi l'erreur d'approximation $\sigma_{\tau,t}\sqrt{1-\mathcal{R}^2_{\tau,t}}$, mais celle-ci fournit des formes caractéristiques moins aisément discernables, à cause du coefficient $\sqrt{1-\mathcal{R}^2_{\tau,t}}$ qui applatit les courbes vers les $\mathcal{R}_{\tau,t}$ extrêmes.

Tonalité La valeur de $y_t$, par rapport à une échelle dans laquelle les valeurs de consigne et les seuils soient mis en évidence. Par tonalité nous voulons parler uniquement de la position dans le spectre des couleurs.

Saturation La distance vers le passé. Concrètement, nous utiliserons une saturation qui est une fonction décroissante de $v$ pour les points $y_{t-v}$. Sur un fond de saturation nulle et de couleur neutre, les points les plus anciens se verront fondus, et les points récents mis en valeur.

Cette représentation permet de résumer de façon synthétique un enregistrement long, indépendamment de la granularité emplyée pour la visualisation. Ceci est à contraster avec les représentations suivant le temps des séries chronologiques, dans lesquelles la granularité des phénomènes et les échelles de visualisation jouent énormément sur la perception.

L'exploitation des formes caractéristiques que la représentation fournit dépasse le cadre que nous nous sommes fixé, au sens où elle implique l'identification de patrons de référence, donc de la connaissance — *a priori* ou par apprentissage. Elle ne sera pas abordée.

C'est à partir de la projection dans l'espace tendance *vs.* stabilité que nous allons procéder à la quantification des données afin d'obtenir des symboles [CCPR99].

## 4.2 Partitionnement tendance *vs.* stabilité

Le partitionnement de l'espace tendance *vs.* stabilité nécessite des hypothèses portant sur le type, le nombre des symboles et leur interprétation. Même si nous pouvons définir un certain nombre de façons *a priori* intéressantes d'établir ces hypothèses, il est difficile, à ce stade, de fournir des moyens de comparaison et de validation. Celles-ci ne pourront être établies que par la validation de l'approche dans sa globalité, par l'évaluation des sorties de la chaîne complète de traitement (*cf.* Fig.1.7).

### 4.2.1 Catégories

La seule base certaine de laquelle nous partons est l'intégration des notions de tendance qualitative (augmentation, diminution, constance) et de stabilité (stable, instable). Par-dessus, nous pouvons intégrer les valeurs en termes de valeur normale, anormale, aberrante par rapport aux seuils de normalité issus de la connaissance médicale.

Les catégories peuvent êtres définies par partitionnement classique net, ou bien par *granularisation* (*c-à-d.* quantification floue [Zad96]). Sur ce point, c'est la méthode d'apprentissage utilisée ultérieurement qui déterminera si le flou peut être intégré ou non.

Des qualificateurs linguistiques peuvent ou non être appliqués à ces catégories.

### 4.2.2 Modalités de partitionnement

Nous pouvons proposer deux familles de modalités de partitionnement :

$d_{ist}$ Par découpage des distributions suivant des critères fixes. Typiquement, un découpage en quantiles.

$v_{al}$ Par découpage des plages de variation. Typiquement, en fixant des seuils sur les valeurs possibles de la tendance et la stabilité.

Ces deux modalités peuvent être implémentées :

$p_{rio}$ En suivant des critères fixés *a priori*.

$a_{just}$ Par ajustement aux données (de façon incrémentale ou par rapport à une calibration).

De plus, les partitionnements peuvent être réalisés :

$m_{arg}$ A partir des distributions et valeurs marginales de la tendance et la stabilité.

$c_{onj}$ Par la prise en compte de la distribution conjointe.

Parmi ces différentes formes d'attaque du problème, les paramètres à définir dépendront de la connaissance d'un expert du domaine ou d'un expert de l'analyse. Il interviendra de façon plus ou moins directe : très directement pour la fixation de seuils dans $v_{al}$ ci-dessus ; plutôt indirectement, par le choix de critères et de méthodes pour $a_{just}$.

Les différents choix et les ajustements ultérieurs devront se faire dans un souci de lisibilité et d'interprétation aisée. Les critères de précision ou de qualité de modélisation ne peuvent être évalués *a priori* et ne sont applicables que si l'on considère la symbolisation comme une étape de prétraitement pour une analyse postérieure. On pourra alors utiliser des critères sur les résultats de l'analyse pour ajuster les paramètres que nous avons signalés.

## 4.3 Exemples de partitionnement

Il n'est pas immédiat, à ce stade, de déterminer des critères objectifs de comparaison et de validation des diverses approches. Nous proposons dans la suite trois modalités de partitionnement. Nous nous limitons à la détermination des catégories $\{\nearrow, \searrow, \rightarrow, \rightsquigarrow\}$.

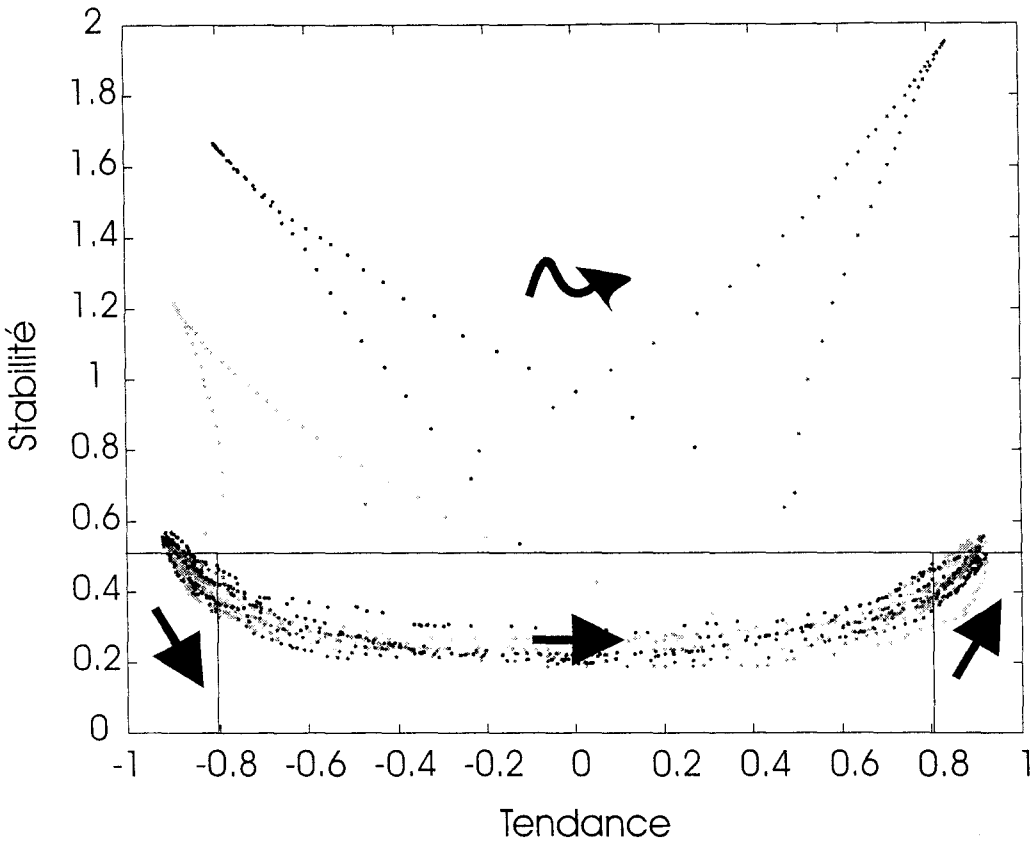



FIG. 4.4 – Exemple de symbolisation en fonction des distributions pour SBQD
Les points correspondant aux changements de ligne de base rentrent bien dans la zone définie comme instable $\rightsquigarrow$, *c-à-d.* au-delà du percentile 80%. Néanmoins, les frontières entre stable $\rightarrow$ et augmentation $\nearrow$ et diminution $\searrow$ restent arbitraires, même si elles dépendent de la distribution de référence et que les quartiles sont une représentation courante de la dispersion [Tuk77]. Le niveau de gris correspond à l'"âge" des données : foncé vers la fin de l'enregistrement, clair vers le passé, afin de mettre en évidence l'écoulement du temps *cf.*§4.1.

### 4.3.1 En fonction des distributions

A titre d'exemple illustratif, nous avons déterminé des catégories $\{\nearrow, \searrow, \rightarrow, \rightsquigarrow\}$ indépendamment sur la tendance et la stabilité. Une quantification $\mathsf{d_{ist}m_{arg}p_{rio}}$ peut être définie par :

Tendance En divisant la distribution en quartiles, la diminution correspond au percentile 25%, l'augmentation au percentile 75%.

Stabilité On considérera l'instabilité pour les écarts-types locaux supérieurs au percentile 80% de la distribution.

De cette manière, nous prenons en compte les différences de distribution suivant les paramètres. L'illustration est faite en Fig.4.4 pour SBQD et Fig.4.5 pour 72040:$SpO_2$. Nous avons aussi inclus la variable temps pour illustrer le propos du §4.1. La clarté attribuée à chaque point correspond à l'âge de la mesure en question. C'est l'application au noir et blanc de la saturation pour les couleurs. Les valeurs n'ont pas été représentées par leur couleur.

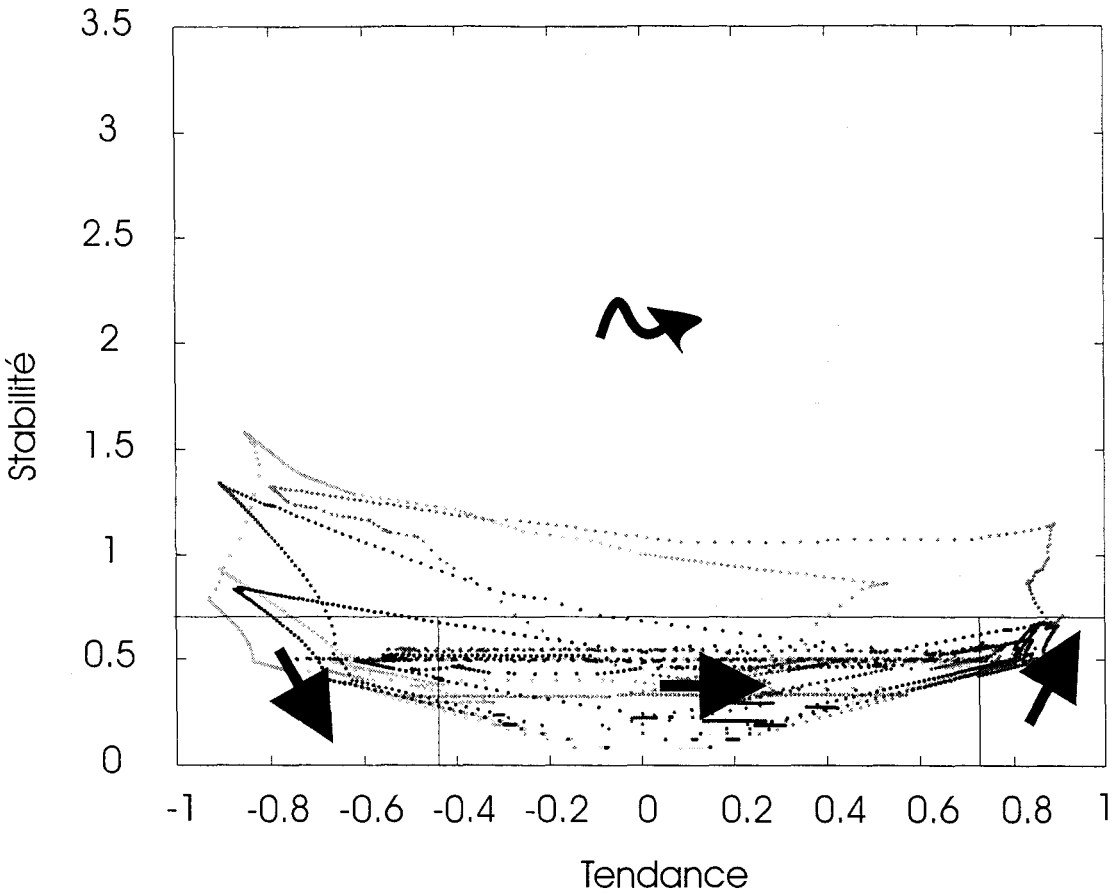



FIG. 4.5 – Exemple de symbolisation en fonction des distributions pour 72040:$SpO_2$
Les séquences de changement de ligne de base, et d'impulsion sont bien repérés dans la classe instable $\rightsquigarrow$. Par contre, le comportement hors stabilité $\rightarrow$ vers les tendances positives ou négatives s'accompagne d'une augmentation de l'instabilité. La clarté du tracé correspond à la distance des points vers le passé.

Il ressort de cet essai que la définition de frontières de façon indépendante pour la tendance et la stabilité n'est pas adaptée à l'interprétation visuelle qu'on peut faire : la déviation positive ou négative s'accompagne toujours d'une augmentation de l'écart-type local.

Ayant défini les frontières à partir de la distribution, nous avons déterminé que le comportement de tendance est donné par la majorité des occurrences $\rightarrow$, ce qui définit augmentation $\nearrow$ et diminution $\searrow$. Les situations $\nearrow$, $\searrow$ se voient ainsi attribuer un caractère relatif.

Par contre, le caractère *a priori* exceptionnel de $\rightsquigarrow$ n'est pas pris en compte par la frontière au percentile 80% : nous avons déterminé *a priori* que l'instabilité est définie par l'appartenance au 20% des valeurs les plus élevées de variance locale.

### 4.3.2 En fonction des valeurs

Alternativement à l'approche précédente, basée sur les distributions, nous proposons un partitionnement absolu. De cette façon, nous éliminons le caractère relatif des frontières, par l'introduction d'un partitionnement *a priori*.

A titre d'exemple, nous avons opéré une quantification $\mathsf{val}_{\mathsf{arg}}\mathsf{prio}$ :

Tendance Les catégories $\nearrow$ et $\searrow$ sont déterminées pour des valeurs dépassant 0.8 en valeur absolue.

Stabilité L'instabilité sera décrétée pour tout écart-type local supérieur à deux fois la précision des données.

Les partitionnements de type $\mathsf{m}_{\mathsf{arg}}$ comme celui-ci se placent dans une optique basée sur la connaissance. Ceci présente les inconvénients propres à de telles approches (*cf.* §6.3), notamment la possibilité d'appliquer le même partitionnement à des données issues de conditions différentes (thérapeutiques, démographiques...).

Ces inconvénients sont à contraster avec ceux inhérents aux approches $\mathsf{d}_{\mathsf{ist}}$, basées sur les données, à savoir :

- la détermination des populations de référence : taille, origine pour un même patient ;
- la détermination automatique de la base de cas à laquelle faire référence : caractéristiques de plusieurs populations sur plusieurs patients ;
- la latitude dans l'ajustement des paramètres : qu'est-ce qui dépend des données courantes, qu'est-ce qui est établi d'après la référence, quels critères *a priori* sont préférés par rapport à quels ajustements...
- les périodes de validité des paramètres extraits : coefficients d'"oubli"...

D'un autre côté, les deux partitionnements $\mathsf{d}_{\mathsf{ist}}\mathsf{m}_{\mathsf{arg}}\mathsf{prio}$ et $\mathsf{val}_{\mathsf{arg}}\mathsf{prio}$ sont établis de façon marginale. Or, les deux dimensions tendance et stabilité sont couplées.

### 4.3.3 Par classification automatique

Pour exploiter les deux dimensions du plan tendance *vs.* stabilité, nous explorons un partitionnement de type $\mathsf{d}_{\mathsf{ist}}\mathsf{c}_{\mathsf{onj}}\mathsf{a}_{\mathsf{just}}$.

L'algorithme utilisé est itératif, par centroïdes mobiles. On part de quatre prototypes pour les catégories $\nearrow$ (le point $(+1,0)$ du plan tendance *vs.* stabilité), $\rightarrow$ $(0,0)$, $\searrow$ $(-1,0)$ et $\rightsquigarrow(0,2)$. Les points sont étiquetés suivant leur distance aux prototypes, puis ceux-ci sont recalculés comme moyenne des points portant l'étiquette correspondante. Une petite dizaine d'itérations est suffisante pour stabiliser les centroïdes. La sensibilité de l'algorithme à la position initiale des prototypes a été diminuée par bruitage de ces points et sélection des centroïdes finaux comme modes des distributions finales, parmi celles qui conservent le nombre de centroïdes.

La Fig. 4.6 (*resp.* Fig. 4.7) illustre l'évolution des symboles repérés par les trois méthodes pour 72040:$SpO_2$ (*resp.* SBQD).

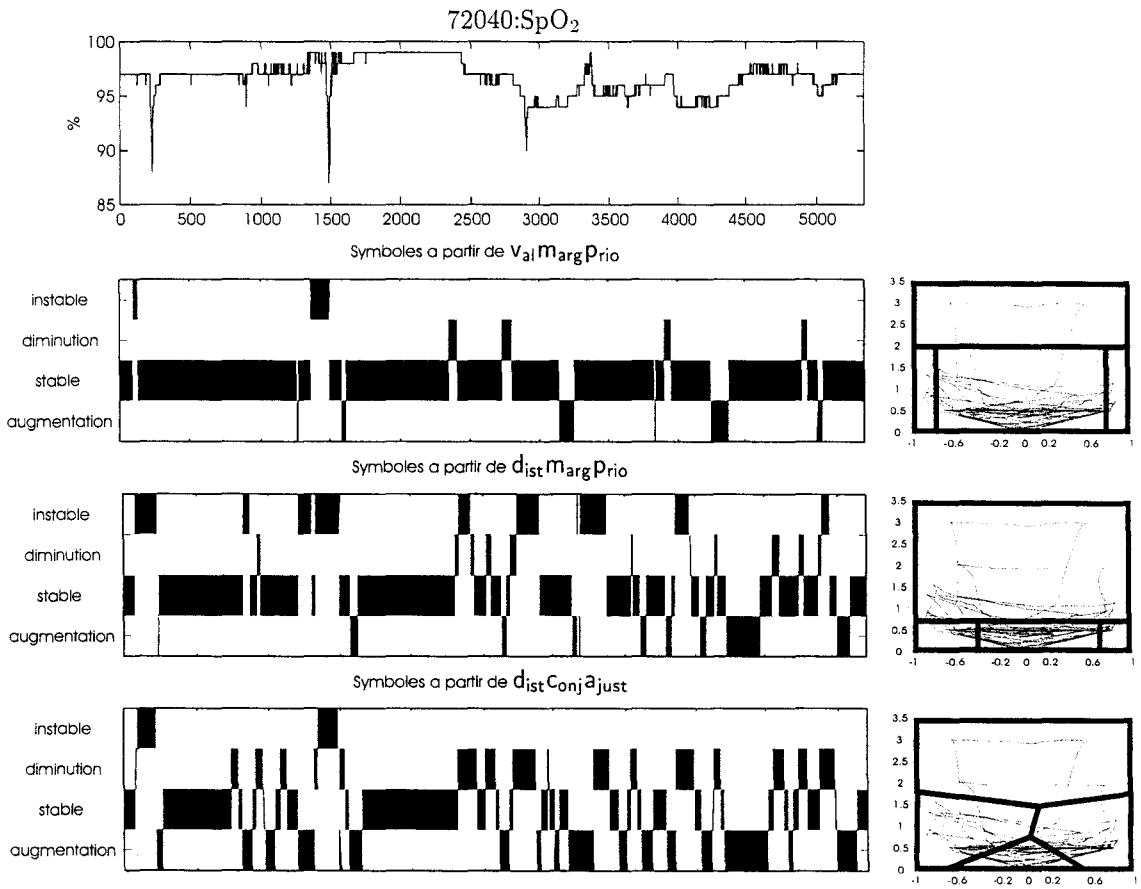



FIG. 4.6 – Application de la symbolisation à 72040:$SpO_2$

A gauche, de haut en bas : la série chronologique de 72040:$SpO_2$, les symboles repérés par $Val\,Marg\,Prio$, ceux issus de $Dist\,Marg\,Prio$ et ceux issus de $Dist\,Conj\,Ajust$. A droite des symboles sont représentés les partitionnements du plan tendance *vs.* stabilité qui leur correspondent.

Les zones contiguës de même symbole sont nettement plus longues pour $Val\,Marg\,Prio$. L'aspect exceptionnel de $\rightsquigarrow$ est aussi plus évident. Le partitionnement induit par $Dist\,Conj\,Ajust$ suit de plus près les variations de la donnée.

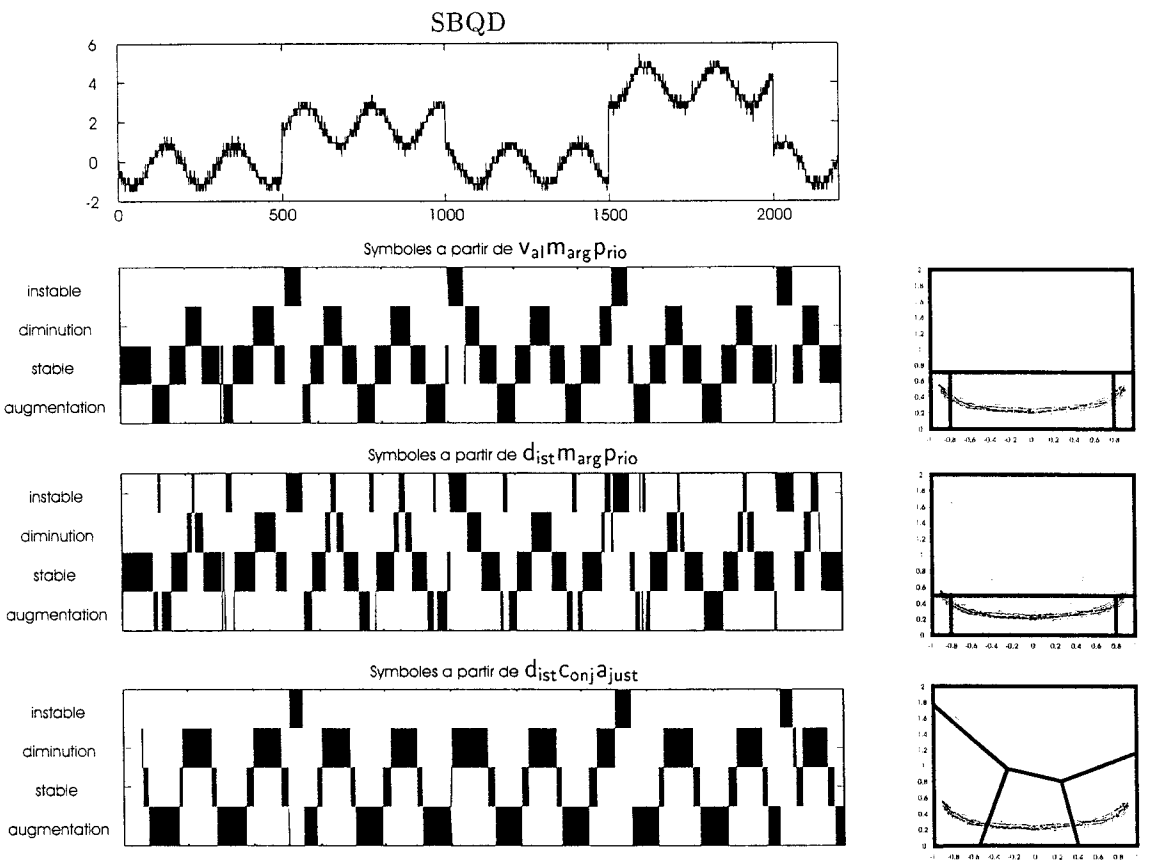



FIG. 4.7 – Application de la symbolisation à SBQD

A gauche, de haut en bas : la série chronologique de SBQD, les symboles repérés par $v_{al}m_{arg}p_{rio}$, ceux issus de $d_{ist}m_{arg}p_{rio}$ et ceux issus de $d_{ist}c_{onj}a_{just}$. A droite, les partitionnements correspondants du plan tendance *vs.* stabilité.

Les patrons d'oscillation $\nearrow\rightarrow\searrow\rightarrow$ sont mieux repérés par $v_{al}m_{arg}p_{rio}$ et $d_{ist}c_{onj}a_{just}$. Néanmoins, il a fallu fixer un seuil arbitraire pour la détermination de $\rightsquigarrow$ suivant $v_{al}m_{arg}p_{rio}$, et une population de référence arbitraire pour $d_{ist}c_{onj}a_{just}$. La régularité du signal SBQD est retrouvée par les trois approches.

## 4.4 Évaluation ?

D'après les comparaisons empiriques que nous avons effectuées, avec des paramètres établis arbitrairement et une exploration partielle des possibilités, nous pouvons dégager un certain nombre d'éléments :

- La normalisation opérée pour le calcul de $\mathcal{R}_{\tau,t}$ autorise à définir des seuils sur les valeurs.
- Le caractère exceptionnel des épisodes d'instabilité peut être repéré davantage par une approche seuillée. Néanmoins, nous n'avons pas de critère de normalisation pour $\sigma_{\tau,t}$ qui permette de définir des frontières indépendamment de la distribution de $\sigma_{\tau,t}$ pour chaque variable.
- Les seuillages indépendants sur $\mathcal{R}_{\tau,t}$ et $\sigma_{\tau,t}$ ne rendent pas bien compte de certains comportements. L'augmentation systématique de $\sigma_{\tau,t}$ pour des $|\mathcal{R}_{\tau,t}| \sim 1$ par exemple est prise en compte par le calcul de l'erreur d'approximation (voir ci-dessus). Mais cet indicateur de stabilité présente des formes typiques plus difficiles à distinguer visuellement. Il faut donc *a priori* préférer des approches de type $\mathsf{c_{onj}}$ à des approches de type $\mathsf{m_{arg}}$, malgré la simplicité de mise en œuvre de celles-ci (dans le cas $\mathsf{d_{ist}}$, notamment).
- Pour des approches de type $\mathsf{c_{onj}}$, il est encore nécessaire de définir le nombre de points qui seront utilisés pour la classification. Ces approches étant basées sur les données, les interprétations précises des catégories dépendront des données effectivement présentes, ce qui rend le partitionement relatif à une période. Il s'agit d'un point de vue complémentaire de celui des zones catégorielles *a priori* qui cherchent des correspondances entre les valeurs réelles et des références.
- Il faut définir précisément les conditions dans lesquelles les seuils seront définis : si on considère des partitionnements à partir de certaines caractéristiques des données (distributions, distributions conjointes, densité d'occupation du plan...), celles-ci devront être estimées. La question de l'échelle d'observation des phénomènes se repose alors. Que faire, par exemple, de données monotones en moyenne sur de longues périodes ?

Mais l'évaluation et la comparaison des diverses approches que l'on peut proposer est difficile.

En effet, elles pourraient être faites à partir de la connaissance d'expert, mais encore faut-il que l'expert puisse dialoguer et interpréter dans le cadre de cette décomposition, ou que l'on puisse transposer dans celle-ci les propos sur les séries chronologiques et les séquences repérées sur les enregistrements. Dans cette deuxième voie, [HM99] propose de coupler le système de traitement automatique avec un système d'acquisition de la connaissance d'expert dans un même logiciel d'annotation temporelle d'événements en vue de leur caractérisation.

D'un autre côté, pour une approche plutôt basée sur les données, le seul critère objectif de comparaison et d'évaluation que l'on peut proposer est en relation avec l'utilisation ultérieure des symboles. Ainsi, c'est en validant l'approche complète (Fig.1.7) que l'on pourra ajuster les paramètres correspondant à la transformation numérique-symbolique. Les critères exposés alors seront ceux de simplicité et intelligibilité du modèle, de contraste, de précision par rapport aux événements effectivement survenus...

Il resterait à définir concrètement les techniques d'ajustement (centroïdes, partitionnement flou, reconnaissance de patrons...) et leurs paramètres.

## 4.5 Conclusion

Nous avons proposé et illustré la projection de chaque variable dans le plan tendance *vs.* stabilité.

A partir de cette projection nous avons suggéré la mise en place d'une procédure de quantification pour la symbolisation. Les possibilités sont nombreuses à ce stade ; nous en avons donné une classification.

Nous avons illustré trois approches parmi les plus simples.

Faute de pouvoir comparer et évaluer chacune des possibilités, ce qui fait éminemment intervenir la connaissance d'expert, nous sommes obligés pour poursuivre, de choisir arbitrairement un mode de transformation numérique-symbolique à partir de la projection tendance *vs.* stabilité. Nous emploierons en l'occurrence $\mathsf{val}_{\mathsf{arg}\,\mathsf{prio}}$.

...
*Gradualmente se vio (como nosotros)*
*aprisionado en esta red sonora*
*de Antes, Después, Ayer, Mientras, Ahora,*
*Derecha, Izquierda, Yo, Tú, Aquellos, Otros.*
...

Jorge Luís Borges,*El Golem*

# 5

# Induction par Arbres de Décision

Où les symboles définis servent à déterminer des modèles symboliques locaux et où ceux-ci mettent en évidence des changements d'état.

## Sommaire

Ayant proposé des moyens de conversion numérique-symbolique pour les données, nous allons employer des méthodes d'apprentissage pour abstraire ces symboles en représentations de niveau d'abstraction plus élevé.

Pour cela, nous allons utiliser des méthodes d'Induction par Arbres de Décision. Le choix de cette famille de techniques au détriment de systèmes de classification non supervisée et de systèmes de classification supervisée utilisant d'autres systèmes de représentation est justifiée *a priori* par :

- Le fait que le formalisme de représentation, l'arbre de décision, soit une représentation connue du personnel médical.
- La capacité de ces approches à intégrer des données symboliques *et* numériques dans le même modèle [Qui96].
- La possibilité de retranscrire les arbres de décision en ensembles de règles pour insertion ultérieure dans d'autres systèmes de raisonnement [RVBC92].
- La présence, parmi nos données, de paramètres physiologiques caractérisant l'état global du patient : $SpO_2$, $EtCO_2$,...qui peuvent servir à l'interprétation dans le rôle de variables à expliquer.

Le choix s'est fait *a priori* sur ces critères. Il n'a pas, à ce stade, été validé par rapport à d'autres approches.

Nous allons ici présenter et passer en revue les diverses méthodes et variantes de l'induction par arbres de décision. Ensuite, nous verrons les limitations majeures dans les hypothèses d'application de ces méthodes. Nous proposerons la méthodologie développée dans les chapitres précédents comme moyen de prétraitement pour diminuer l'inadéquation. De même, nous introduirons des moyens de post-traitement pour la caractérisation des modèles construits et, partant, la détermination de zones de stabilité. Nous introduirons au passage des moyens de visualisation supplémentaires.

Développées en parallèle par les communautés de l'Intelligence Artificielle et des Statistiques vers le début des années quatre-vingt, les méthodes d'*Induction par Arbres de Décision* [Mur98] font partie de la classe d'algorithmes d'*Apprentissage Automatique* (Machine Learning [KM90]). Cette classe regroupe aussi les algorithmes de classification, et de façon générale les moyens d'abstraire des représentations [BC90] à partir de données brutes.

Les applications sont nombreuses, et des termes nouveaux comme le *data mining* — coïncidant avec la vogue du concept de *data warehouse* — ou *knowledge discovery in databases* (KDD [Kod97] par exemple) sont apparus, en même temps que nombre de publications et de congrès, comme ponts entre la recherche sur le sujet et les applications. Les firmes les plus importantes dans les secteurs du traitement de données, comme Oracle, SGI ou SAS proposent toutes des solutions de data mining. A titre d'anecdote, les principaux systèmes développés dans la recherche académique ces dernières années (mais aussi les classiques comme CART) sont aujourd'hui à la base de produits commerciaux. Le projet européen ESPRIT-II Machine Learning Toolbox (1989–1993) a été établi afin de faire l'état de l'art en la matière et proposer des lignes protocolaires d'application et de sélection de méthodes.

La philosophie derrière ce boom est essentiellement celle du recours à des méthodes automa-

tiques pour le traitement de recueils effectués automatiquement ou de bases de données cumulant des grandes quantités d'information. Les méthodes d'apprentissage dans ce cadre apparaissent comme des aides à la navigation et à la compréhension de masses importantes de données.

C'est exactement en ce sens-là que nous les abordons. Le choix de travailler avec des méthodes d'apprentissage automatique tient autant à la nature des questions posées qu'à la faisabilité technique de moyens d'aide à l'interprétation de volumes de données imposants.

## 5.1 Définitions

Les méthodes d'Induction par Arbre de Décision [Mur98] ne se basent pas sur l'espace de représentation vectoriel de l'Analyse de Données, mais sur un espace des distributions. Les outils de mesure dans cet espace sont des indicateurs statistiques d'indépendance et de proximité entre distributions, telles que l'entropie et le $\chi^2$.

Les données sont vues comme une population d'apprentissage formée d'individus $O_1, O_2, \ldots, O_T$ pour lesquels des attributs $A_1, A_2, \ldots, A_n$ ont des valeurs connues parmi un ensemble de modalités possibles $a_{kj}$ avec $A_k(O_i) \in \{a_{k1} \ldots a_{k|A_k|}\}$. $|A_k|$ désigne le nombre de modalités de $A_k$.

Les attributs peuvent être qualitatifs ou quantitatifs. Dans ce deuxième cas, les valeurs supposées continues ont un statut particulier, puisque le nombre de partitions de leur continuum est infini. Une classe est donnée pour chaque individu d'après la variable à expliquer. Le classificateur essaiera de trouver un modèle de définition des classes de la variable à expliquer à partir des valeurs des attributs explicatifs.

Un arbre de décision est un arbre avec une classe à chaque feuille, un test à chaque nœud et un résultat du test à chaque branche qui en est issue. C'est une représentation d'un classificateur hiérarchique qui pourrait être écrit sous forme de règles. En partant de la racine et en partitionnant l'espace des distributions suivant les tests, on arrive à déterminer la classe d'appartenance de l'objet qu'on teste. Un exemple artificiel permet d'illustrer ces notions, voir Fig.5.1.

Les arbres de décision essaient de résoudre un problème de classification *supervisée*, *c-à-d.* que les classes sont prédéterminées, et le système doit trouver un moyen d'abstraire le partitionnement.

Il existe des méthodes proches mais pas identiques, les *arbres de régression*, qui essayent de déterminer un partitionnement de l'espace des variables explicatives en fonction des valeurs d'une variable à expliquer qui n'est plus qualitative mais quantitative. Le système CART [BFSO84] a inauguré ces approches. Nous ne nous occuperons pas ici de cette famille de méthodes. Signalons toutefois que depuis peu des essais sont faits pour rejoindre les deux approches, en ramenant le problème de la régression à un problème de classification [TG97] ; de même, les techniques d'ajustement par le flou des arbres de décision ont été appliquées aux arbres de régression [Jan94].

Par ailleurs, il existe des méthodes d'apprentissage non-supervisé, aussi appelées méthodes de classification automatique. Ces méthodes sont basées sur le choix d'une métrique pour évaluer la distance entre les individus $O_1, O_2, \ldots, O_T$ et, en fonction des proximités, établir des classes.

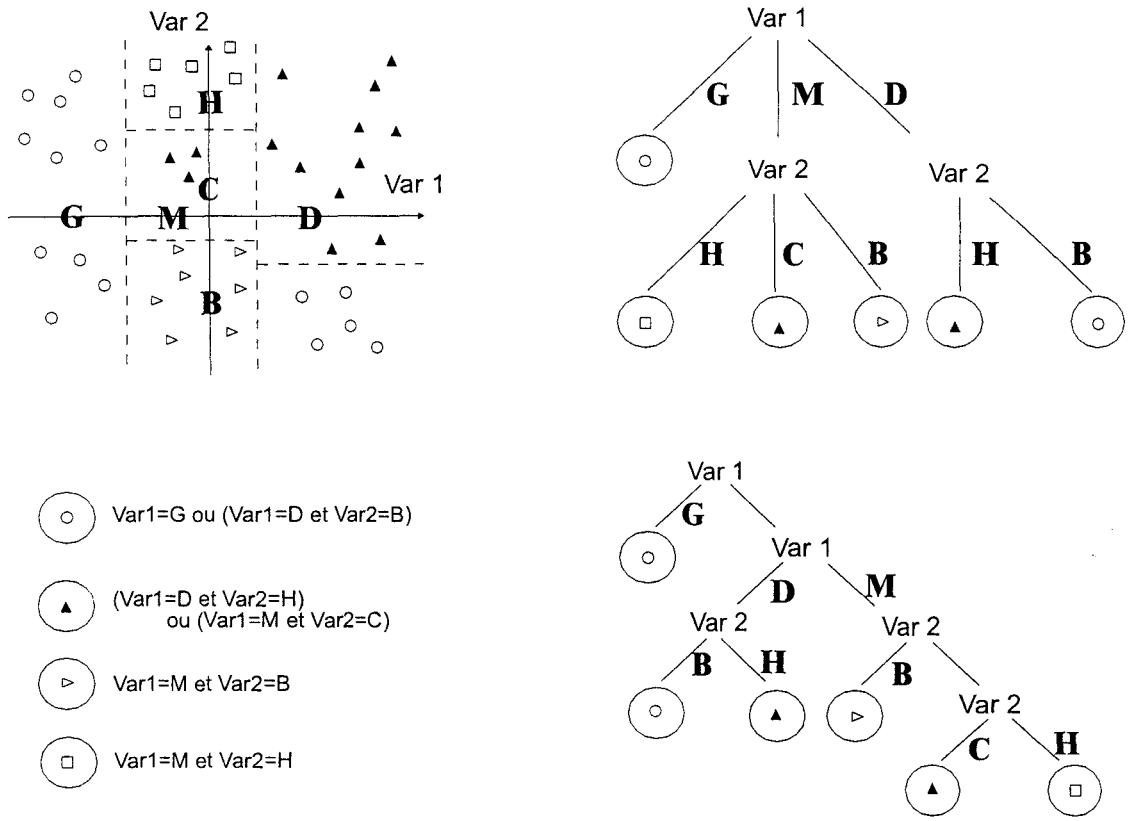



FIG. 5.1 – Induction par arbres de décision sur un exemple artificiel
L'arbre représente exactement la distribution issue des mesures rapportées suivant Var1 et Var2. Les modalités respectives sont G,M,D et H,C,B. Ces modalités peuvent avoir été déterminées dans une phase de pré-traitement, ou bien au vol au moment de la construction de l'arbre en exploitant le caractère continu de Var1 et Var2 (ce qui explique les partitions différentes suivant Var2 d'après les valeurs de Var1).
En bas à gauche, l'ensemble de règles équivalent au classificateur établi par l'arbre de décision. En bas à droite, un arbre binaire équivalent.

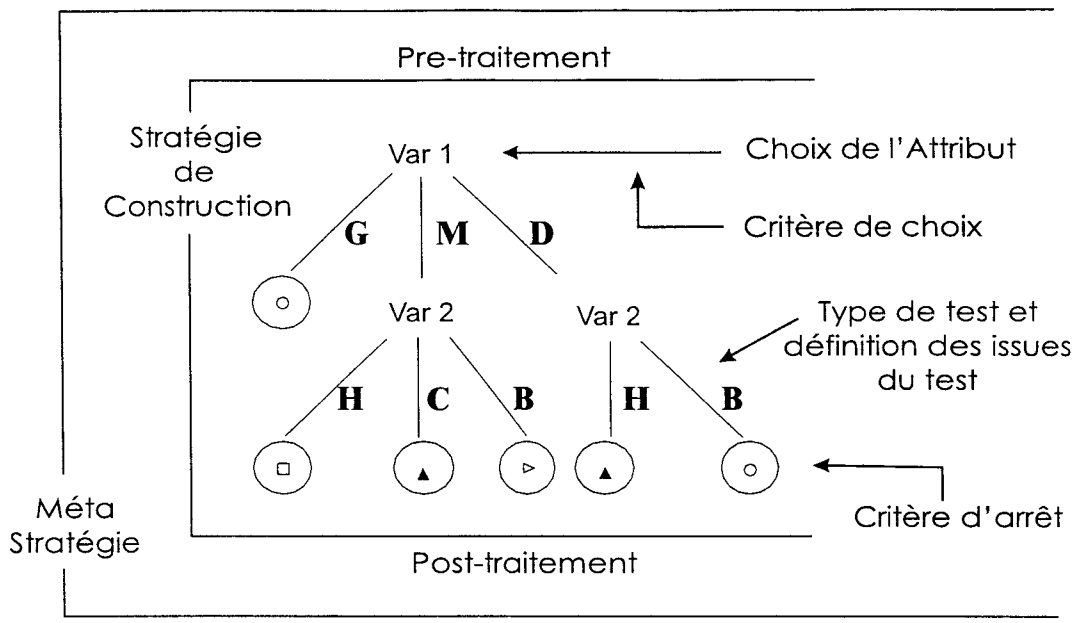



FIG. 5.2 – Schéma général de génération d'arbres de décision
Ici les attributs sont les variables d'origine de l'ensemble d'apprentissage de la Fig.5.1. Le type $\mathcal{S}$ de partitionnement induit est droit, à plusieurs modalités (les systèmes réels utilisent en général un partitionnement droit). Il n'y a pas dans cet exemple artificiel de pré ou post traitement, la stratégie $\mathcal{G}$ est récursive et il n'y a pas de méta-stratégie. Le critère $\mathcal{F}$ d'arrêt n'est pas spécifié, puisque les classes sont pures.

Celles-ci doivent minimiser une certaine mesure d'homogénéité intra-classe, tout en maximisant une mesure de la discrimination inter-classes.

## 5.2 Schéma général

De l'étude de différents algorithmes d'Induction par Arbre de Décision existants, nous avons développé un schéma général où chaque algorithme apparaît comme le choix de certaines heuristiques et de certains critères dans le problème général de trouver l'arbre de décision qui modélise au mieux une population d'apprentissage.

De façon générale, (voir Fig.5.2) les méthodes d'Induction par Arbre de Décision sont caractérisées par une stratégie globale $\mathcal{G}$ de génération de l'arbre, un critère $\mathcal{A}$ de sélection de l'attribut à tester, un critère $\mathcal{S}$ de sélection du test à réaliser à chaque nœud [DF96] et un critère $\mathcal{F}$ de fin de branche. En plus, des phases de pré et post traitement peuvent augmenter la puissance explicative des méthodes d'Induction par Arbre de Décision.

### 5.2.1 Stratégie de génération de l'arbre

Pour générer l'arbre, la méthode la plus naturelle semble être l'approche récursive [Qui86]. A chaque nœud correspond un ensemble d'individus, la totalité pour la racine. Si un nœud est une

feuille d'après $\mathcal{F}$ alors il lui est attaché une classe. Sinon, il lui est attribué un test, d'après $\mathcal{A}$ et $\mathcal{S}$, à appliquer sur son ensemble d'individus. Les résultats du test déterminent un découpage en des sous-ensembles qui seront attribués aux sous-arbres correspondants.

Cette approche peut être implémentée : ou bien par l'utilisation de la totalité de la population d'apprentissage disponible, ou bien par constructions successives sur des échantillons de taille croissante jusqu'à stabilisation du modèle construit [Qui96]. Ce choix s'apparente à une méta-stratégie de construction.

De l'une ou l'autre façon, l'algorithme peut mener à une explosion combinatoire si les données ne présentent pas de régularités compatibles avec $\mathcal{A}$, $\mathcal{S}$ et $\mathcal{F}$. Les arguments avancés jusqu'ici contre cette éventualité sont empiriques [Qui86] et heuristiques : en général on ne testera pas une variable deux fois lors d'un parcours d'arbre. Cela limite la taille des arbres à $|A|^n$ pour $n$ attributs à $|A|$ modalités.

Cette stratégie peut être remplacée [Utg95] par une approche purement agrégative qui ne modifie l'arbre que localement et ne pousse pas à la reconstruction de l'arbre dans sa totalité. Cette stratégie construit des arbres de façon progressive, et permet l'adaptation des modèles aux données nouvelles.

Une troisième approche développée [Tur95] consiste dans l'exploration par des algorithmes génétiques de l'espace des arbres de décision afin d'en trouver un qui soit optimal suivant un critère de qualité $\mathcal{Q}$.

Récemment, en parallèle à l'extension des méthodes de fusion de données, des méta-stratégies d'agrégation de résultats ont été proposées, qui construisent un certain nombre de modèles puis tirent leur conclusion finale en fonction de l'ensemble de réponses de ces modèles. Voir par exemple [OJ95] où l'on compare les élagages *a posteriori* avec la fusion d'ensembles d'arbres de décision. Dans [SL98], on utilise un réseau bayésien, construit avec un algorithme génétique, pour effectuer la fusion entre les conclusions d'un ensemble de classificateurs induits par différentes méthodes.

De plus, des méta-stratégies générales de construction comme le bagging ou le boosting peuvent être appliquées autour des méthodes d'induction par arbres de décision.

Le bagging [Bre94] consiste à construire des ensembles de modèles pour des sous-ensembles de la population d'apprentissage, et d'effectuer la classification effective par vote parmi les résultats des divers classificateurs. Ce principe est proche de celui du bootstrap [ST95] : les modèles construits sur des sous-échantillons sont utilisés par agrégation.

Le boosting [FS96] consiste à déterminer un ensemble de poids pour la population d'apprentissage, puis à réajuster ces poids en fonction de l'exactitude de la classification obtenue.

Ces deux méta-stratégies peuvent être vues comme des ajustements de l'espace de représentation de la population d'apprentissage et de l'espace des modèles. Dans le cas du bagging, on ne se place plus dans l'espace des modèles en arbre de décision, mais d'ensembles d'arbres de décision, en ajoutant un algorithme de vote. Le vote majoritaire a été utilisé jusqu'ici. Dans le cas du boosting, on se place dans un espace des individus qui est pondéré.

### 5.2.2 Critère d'arrêt

Le critère d'arrêt détermine et définit la sortie du classificateur. Pour des données consistantes (deux individus identiques sont identiquement classés), arrêter lorsqu'on a une classe homogène est un critère naturel et simple à implémenter. Par contre, en présence de données "bruitées", la consistance n'est pas assurée ; dans ce cas, il est possible de construire un arbre qui épuise les tests mais dont les branches terminales ne mènent pas à des classes homogènes. On attribue alors à la feuille la classe dominante ou bien la probabilité d'appartenance. Dans tous les cas, l'arrêt est établi lorsque le redécoupage de la population courante ne permet pas de gain pour la classification.

### 5.2.3 Critères de sélection du test

A chaque nœud de l'arbre, un test est réalisé dont les issues possibles déterminent le sous-arbre rattaché à ce nœud. La sélection de l'attribut à tester se fait en général parmi les attributs $A_k$ présents explicitement dans la base de données. Pour éviter une complexification extrême de l'arbre, chaque attribut n'est testé qu'une fois suivant un parcours descendant quelconque de l'arbre. Le choix du test à réaliser est déterminé à partir de la maximisation d'un certain indicateur : le $\chi^2$, le gain d'information, le coefficient de gain d'information, l'indice de Gini,...

Ces indicateurs sont tous des mesures de dépendance (en termes de fréquence donc de distribution) entre l'attribut $A_k$ testé et la classe correspondante de l'individu. Il s'agit de distances dans l'espace des distributions. Les arguments pour ou contre chaque mesure sont avancés *a priori* ou sur des exemples spécifiques, ce qui ne permet pas de dégager [Mur98] de meilleure mesure de distance dans l'absolu. Cette difficulté est aussi à l'origine des approches moyennantes et de fusion de données mentionnées ci-dessus.

### 5.2.4 Type de test

Le test est en général pour des attributs discrets du type `A` $= a$ `?` (pour des arbres binaires) ou `valeur de A ?`. Ce deuxième type pose des problèmes lorsque la mesure utilisée est entropique : celle-ci tend à favoriser les attributs ayant le plus grand nombre de modalités.

Les systèmes rencontrés jusqu'ici se bornent à ces deux types de test, ce qui restreint la recherche de la meilleure partition de la population d'apprentissage à des heuristiques viables en temps de calcul.

Le cas des attributs continus est différent : la plupart des algorithmes développés exploitent les attributs continus en définissant un point de séparation $P$ dans un test du type `X` $> P$ `?` à deux issues.

Des extensions vers le flou [Kos91, Zad65] ont été développées [ZS95, Mar98]. Elles établissent des tests d'appartenance à des ensembles flous pour des attributs continus. A partir du point de séparation, une heuristique établit les fonctions d'appartenance de l'attribut aux classes subséquentes, en général trois classes '>', '=','<'. Il a été proposé que de tels développements sont proches de

l'utilisation d'attributs supplémentaires non-linéaires, pourvu qu'un choix particulier d'opérateurs flous soit utilisé [IZR+97]. L'introduction en général du flou et l'imprécis dans les techniques d'induction est un thème d'actualité. Pour des définitions de ces notions, voir [FH95] ; pour des applications concrètes, voir [Mar98].



### 5.2.5 Espace de recherche des attributs

Le choix de limiter la recherche de l'attribut à tester aux $A_k$ force à réaliser un partitionnement "droit" de l'espace des distributions. Pour des attributs numériques, cela revient à utiliser uniquement des hyperplans séparateurs orthogonaux aux axes de l'espace. Trois approches au moins cherchent à détendre cette contrainte.

La première [PP97] considère non seulement les $A_k$ dans la recherche mais l'ensemble des partitions de $\{A_1, A_2, \ldots, A_n\}$. La recherche exhaustive dans cet espace est exponentielle, et l'on utilise des heuristiques agrégatives ou désagrégatives pour réaliser cette recherche en temps linéaire. Cette approche s'applique de façon très générale à tous les types d'attributs.

La deuxième [MKS94] s'applique exclusivement aux attributs continus et consiste à chercher une combinaison linéaire des attributs. Cette combinaison définit un hyperplan oblique qui réalise le partitionnement de l'espace de distribution. La recherche de cet hyperplan ne peut pas être envisagée de façon optimale, et des heuristiques d'optimisation locale sont mises en oeuvre, par exemple un type de recuit simulé pour OC1.

Enfin, l'introduction d'attributs qui sont des monômes multivariés des variables numériques d'origine permet d'établir des frontières de forme polynomiale, avec une augmentation de l'exactitude dans certains cas très mal couverts par les partitionnements droits ou obliques [IZR+97]. Cette exactitude est obtenue au détriment du temps de calcul.

### 5.2.6 Critère de choix de l'attribut

Le choix de l'attribut de test se fait d'après un critère de comparaison. Le plus courant est la mesure du gain d'information

$$g(Y|X) = H(Y) - H(Y|X)$$

avec

$$H(Y|X) = H(Y,X) - H(X)$$

où $H(Y|X)$ est l'entropie conditionnelle de $Y$ sachant $X$,

$H(Y,X)$ est l'entropie conjointe de $X$ et $Y$ (celle du couple $(X,Y)$) et

$H(X)$ est l'entropie [Cul72] de $X$ calculée à partir de la distribution de $X$ par :[1]

$$H(X) = -\sum_i p_i \log_2 p_i$$

si $p_i$ est la probabilité de la modalité $i$.

Précisément, l'attribut $X$ choisi sera celui qui maximise $g(Y|X)$ *c-à-d.* qui minimise $H(Y|X)$ si $Y$ est l'attribut qui détermine la classe (la variable à expliquer). Le gain d'information s'interprète comme la quantité d'information qu'apporte $X$ à $Y$ (c'est la transinformation externe de $X$ vers $Y$).

La mesure d'entropie est croissante en fonction du nombre de modalités de la distribution ; elle est maximale pour une distribution uniforme et peut donc s'interpréter comme une mesure de la distance à l'équiprobabilité. La dépendance par rapport au nombre de modalités est gênante [HS94] et a conduit à utiliser un critère dérivé : le coefficient de gain d'information défini par

$$g_r(Y|X) = \frac{g(Y|X)}{H(X)}$$

ce qui normalise la mesure de transinformation par rapport à sa valeur maximale possible.

Une autre mesure est proposée et défendue [DF96] comme critère : le $\chi^2$ de Pearson [Pea00], qui est une mesure de la distance d'une distribution à sa distribution marginale

$$\chi^2 = \sum_{ij} \frac{(p_{ij} - p_{i\bullet}p_{\bullet j})^2}{p_{i\bullet}p_{\bullet j}}$$

où $p_{i\bullet} = \sum_j p_{ij}$ et $p_{\bullet j} = \sum_i p_{ij}$ sont les distributions marginales et $p_{ij}$ est la distribution conjointe de $Y$ et $X$.

L'attribut à tester sera celui qui présente la plus grande dépendance vis-à-vis de $Y$, pour le sous-ensemble de la population d'apprentissage dont il est question.

### 5.2.7 Pré et post traitement

Certains systèmes amènent une heuristique de recherche d'un arbre optimal au traitement *a posteriori* de l'arbre construit. Ainsi, C4.5 optimise l'arbre brut par élagage des sous-arbres pour lesquels la substitution par une feuille n'introduit pas de supplément dans les erreurs de classification.

Lorsqu'il est prévisible que les méthodes d'Induction par Arbre de Décision ne toucheront pas à l'essence de la problématique (le temps pour nous), il est nécessaire de prétraiter les données. Pour nous, c'est l'aspect temporel et relatif des mesures qui n'est pas atteint intrinsèquement : les méthodes d'Induction par Arbre de Décision sont autant statiques, synchrones et stationnaires

[1] Il s'agit d'entropie au sens de Shannon (on utilise un noyau logarithmique), mais il ne s'agit pas de ce qu'on appelle "entropie de Shannon", qui en est dérivée.

que celles d'Analyse de Données. Il faudra donc coder en des modalités particulières les données brutes dont on dispose, de façon à faire apparaître pour l'algorithme le temps. Ces changements de représentation sont délicats car ils posent souvent des questions de définition précise de paramètres arbitraires. Typiquement, comment représenter une variation "moyenne" ?

Ce prétraitement nécessaire dans notre problème diminue de façon appréciable l'intérêt de l'Induction par Arbre de Décision en tant que système de modélisation "sans modèle *a priori*". C'est pourquoi il paraît impératif d'introduire le temps dans ces approches [PS96].

### 5.2.8 Critères de qualité

Pour évaluer les arbres de décision, un certain nombre de critères ont été mis en avant. Parmi ceux-ci, remarquons :

**L'exactitude** L'arbre de décision doit classifier correctement d'autres données du même phénomène si celui-ci a été bien modélisé. Nous avons à notre disposition un flux de données sur lequel tester cette précision.

Pour des données bruitées, au sens de la classification *c-à-d.* présentant des incohérences, les arbres construits présentent une erreur intrinsèque à la construction, indicateur de la précision avec laquelle ils sont capables d'abstraire les régularités dans les données.

**La complexité** La complexité de l'arbre construit peut être décrite de plusieurs façons : profondeur de l'arbre, nombre de noeuds, nombre de feuilles, mais la mesure de la longueur moyenne de parcours (la somme des longueurs de parcours descendant pondérées par la probabilité du parcours) donne un bon indicateur global de la complexité de l'arbre dans son utilisation par rapport au problème.

De plus, la complexité de l'arbre construit est en rapport direct avec sa lisibilité. On transpose souvent la représentation arborescente en une représentation en ensemble de règles.

Il ne s'agit toutefois pas du seul critère de lisibilité. Tout modèle est soumis à une évaluation du point de vue de son interprétation par un expert, utilisateur ou évaluateur. Dans une application telle que la nôtre, le rôle de l'expert est, tout comme celui du modélisateur, à définir explicitement, afin d'éviter les inadaptations de l'outil à sa fonction.

**Le coût** Le coût de (re)construction de l'arbre est un facteur important dans l'application à la surveillance. Il a été montré que la recherche d'un arbre minimal sous la plupart des hypothèses intéressantes constitue un problème NP-complet. Par exemple, d'après [Mur95], [TC92] montre que le problème de construire le plus petit arbre de décision qui permet de distinguer une caractéristique de plusieurs groupes est NP-complet. Tous les algorithmes proposés sont des heuristiques visant à trouver un arbre plus ou moins "bon" suivant les critères d'exactitude et de complexité. En ce sens, l'application locale d'heuristiques au problème de la partition au niveau d'un nœud

doit se faire avec le critère de coût en tête.

## 5.3 Variantes et algorithmes

Parmi les systèmes d'Induction par Arbre de Décision qui ont été suffisamment décrits et dont l'essentiel des algorithmes est implémenté, nous pouvons distinguer :

Le système ID3 [Qui86] implémente une construction récursive sur des échantillons de taille croissante, avec la contrainte d'un test par variable par parcours, et sélectionne les variables à tester parmi les attributs présents d'après un critère de coefficient de gain d'information. Les attributs doivent être qualitatifs, et le bruit est filtré au moment du choix de l'attribut à tester par l'utilisation d'un critère basé sur le chi-deux.

C4.5 [Qui92] introduit en plus l'utilisation de variables continues, la prise en compte de valeurs manquantes et l'élagage *a posteriori*. Sur les attributs continus, un point de coupe est déterminé entre deux valeurs après examen des points de coupe possibles (on utilise le fait que les valeurs continues en informatique sont toujours discrètes, et que la notion d'ordre total est ce qui les distingue des valeurs ensemblistes). Les valeurs manquantes sont ignorées au moment du calcul des distributions. Des améliorations ont été apportées par la suite à l'algorithme de base [Qui96], puis le système a fait la base d'un produit commercial baptisé C5.0. Voir `http://www.cse.unsw.edu.au/~quinlan/`.

OC1 [MKS94] étudie des attributs continus et réalise des partitions obliques en choisissant des hyperplans par un recuit simulé à partir d'une condition initiale aléatoire. Le code source de OC1 est disponible (à la date d'écriture de ce manuscrit) à partir de `www.cs.jhu.edu/~murthy/announce.html` librement pour des fins non commerciales.

ITI [Utg95] est basé sur une stratégie de reconstruction incrémentale de l'arbre. Cette stratégie est basée sur un opérateur de transposition, utilisé aussi dans le système DMTI qui réalise une construction incrémentale d'arbre par parcours de l'espace des arbres suivant un critère de qualité. Pour des chercheurs, le code source est disponible à partir de `http://www-ml.cs.umass.edu/iti/index.html`

ICET [Tur95] effectue la recherche de l'espace des arbres de façon génétique, suivant un critère de qualité qui incorpore non seulement les éléments déjà décrits, mais en plus introduit le coût de réalisation des tests. Cet aspect n'entre pas dans nos préoccupations immédiates.

M5' [YI97] est basé sur les publications décrivant le système M5 de Quinlan ; le système construit des arbres de régression par élagage constructif d'un arbre de classification préalablement construit.

IDF [MS98] utilise un critère d'estimation des distributions avant et après chaque partitionnement ; ce critère permet d'établir l'arrêt en même temps que le choix d'attribut.

NDT [IZR+97] introduit des variables supplémentaires, qui sont des monômes formés à partir des variables continues originales, et réalise le partitionnement de l'espace des variables augmenté avec une technique de combinaison linéaire proche de celle d'OC1 ; le résultat est un partitionne-

ment avec des frontières non-linéaires (pour que l'approche reste traitable, des frontières qui sont des coniques apparaissent comme suffisantes). NDT est la base d'un produit commercial de Data Mining.

LMDT [BU95] construit des formules locales de régression par recuit simulé, construisant donc des *arbres de régression* plutôt que de décision, à la manière de CART [BFSO84].

1R [Hol93] et T2 [AHM95] proposent une stratégie de génération basée sur des principes différents de ceux d'ID3 : ils privilégient tous deux la qualité (et donc le temps de calcul) du partitionnement, au détriment explicite de la profondeur de l'arbre, en s'appuyant sur des arguments empiriques de suffisamment bonne qualité des résultats et de lisibilité améliorée. En effet, 1R ne produit qu'un niveau de profondeur, T2 en génère deux. Voir [Elo94] pour une discussion contrastée des deux philosophies.

M5' fait partie du Waikato Environment for Knowledge Analysis (WEKA, [HDW94]), un système logiciel intégrant des méthodes de classification aussi bien supervisées que non, en fournissant des interfaces vers des systèmes d'évaluation, visualisation et transformation en clauses de Horn pour l'utilisation ultérieure. WEKA[2] est distribué sous la licence GPL[3] et contient des implémentations ou réimplémentations de bon nombre d'algorithmes. Les versions jusqu'à 2.3 sont écrites dans un mélange de Tcl/Tk, C, Prolog et tirent parti des interfaces offertes par les systèmes de visualisation XGobi[4] et dot[5] pour l'exploration des données et des modèles. En plus WEKA peut inter-opérer avec des systèmes tels que Autoclass [CS96][6] et C4.5.

Un autre ensemble d'outils est disponible sous une licence libérale : MLC++ [KLMP94] est une bibliothèque orientée objet écrite en C++. Elle propose une structure qui implémente des algorithmes classiques mais qui permet aussi d'en développer de nouveaux. Jusqu'à la version 1.3, MLC++ fait partie du domaine public ; les versions plus récentes sont disponibles pour la recherche sous des conditions peu contraignantes. Voir `http://www.sgi.com/Technology/mlc/`.

## 5.4 Limites d'application

L'application de ces méthodes à nos données est très problématique. En effet, leur utilisation brute en tant que génératrices de fonctions ne donne pas de résultats probants, tant en précision qu'en facilité d'interprétation. Initialement, nous avons utilisé les algorithmes tels qu'ils ont été implémentés à l'origine (C4.5, en l'occurrence). Leur adaptation aux problèmes posés par nos données nous a conduits à proposer des méthodes d'ajustement, par prétraitement et post-traitement. Nous ne nous sommes pas penchés sur la modification éventuelle des algorithmes pour inclure des aspects temporels.

[2] `http://www.cs.waikato.ac.nz/~ml/index.html`
[3] `http://www.gnu.org/copyleft/gpl.html`
[4] `http://www.research.att.com/~andreas/xgobi/`
[5] `http://www.research.att.com/sw/tools/graphviz/`
[6] `http://ic-www.arc.nasa.gov/ic/projects/bayes-group/group/autoclass/autoclass-c-program.html`

### 5.4.1 Données manquantes

L'inclusion de données manquantes pose un certain nombre de problèmes, pour l'utilisation des systèmes d'apprentissage tels quels. En effet, plusieurs de ces systèmes — C4.5 en particulier — autorisent les données manquantes pour les variables explicatives. Les absences de la variable de classification doivent être traités particulièrement.

Deux possibilités s'offrent à l'analyse. Dans la perspective conservatrice *cf.* §1.2.1, nous ne savons rien sur les données pour lesquelles la classe manque. Nous éliminons alors ces mesures pour opérer l'apprentissage.

Dans la perspective libérale, nous considérons que de l'information est portée par l'absence de données, et alors celles-ci forment la classe absent. Les résultats doivent alors être interprétés avec précaution : lorsque le nombre de mesures manquantes pour la classe dans une population d'apprentissage est élevé, ce sera essentiellement l'absence que la méthode essaiera d'expliquer.

### 5.4.2 Stationnarité locale

Dans l'hypothèse de fréquence d'échantillonnage suffisamment élevée, l'hypothèse supplémentaire de stationnarité locale — où "locale" reste à définir avec précision — permet le traitement des données au sein de fenêtres glissantes. Concrètement, nous appliquons le système d'apprentissage sur $\mathcal{T}$ mesures — les $\mathcal{T}$ dernières mesures pour le traitement en ligne — puis nous reconstruisons un modèle sur les $\mathcal{T}$ données suivantes.

Si la méthode d'apprentissage permet d'abstraire efficacement (en termes de précision et de complexité) le comportement mutuel des données sur la fenêtre de $\mathcal{T}$ mesures considérée, le changement des modèles au fur et à mesure du glissement de la fenêtre indiquera les changements dans les rapports mutuels sous-jacents entre les variables.

Le repérage des zones sur lesquelles cette configuration des rapports entre les variables est sensiblement constante fournira une représentation nouvelle du flot de données en terme de succession d'états, où chaque état sera caractérisé par un modèle construit par apprentissage. La figure 5.3 illustre ce schéma de traitement.

### 5.4.3 Variable à expliquer

La stationnarité envisagée ci-dessus est celle des distributions propres aux données et mutuelles entre données. Plus exactement, l'algorithme d'apprentissage s'occupe des distributions mutuelles des variables explicatives prises une à une par rapport à la variable à expliquer.

Nous disposons de certaines variables qui peuvent être considérées comme significatives pour déterminer l'état global du patient : $SpO_2$, $EtCO_2$,...Nous pourrions les assimiler à des variables de sortie du système. Pour les variables que nous considérons comme explicatives, certaines peuvent être vues comme des entrées — la plupart des mesures ventilatoires, en fonction du mode de ventilation utilisé — et d'autres comme variables d'état — les variables hémodynamiques en particulier.

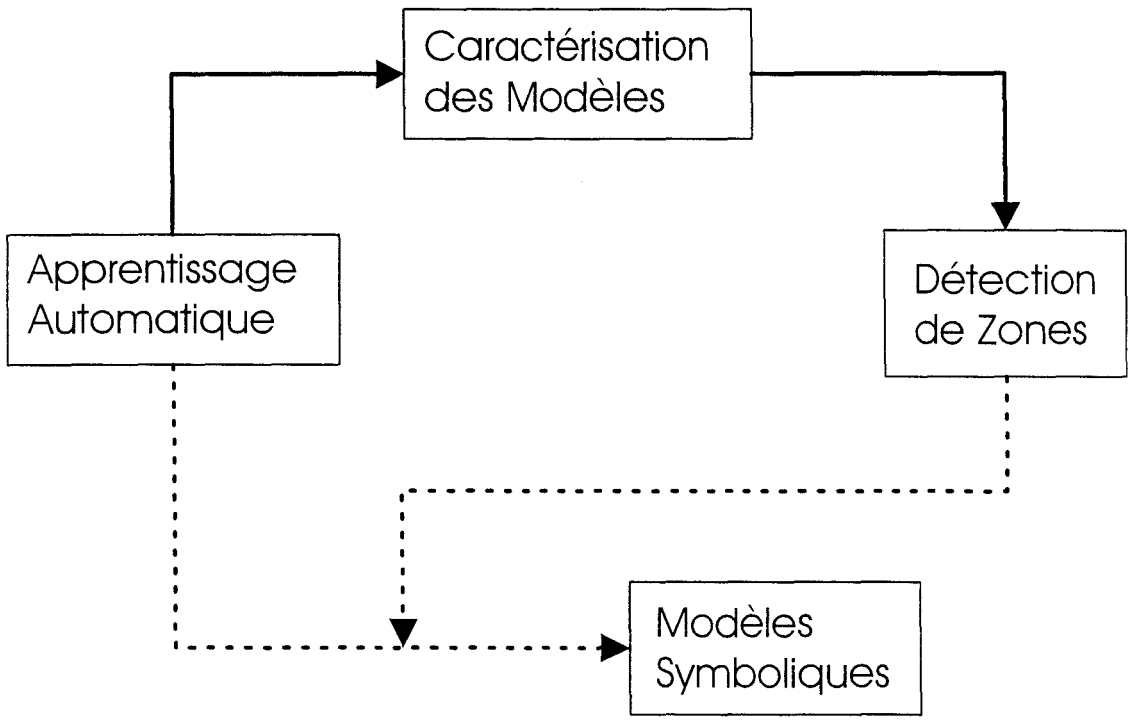



FIG. 5.3 – Schéma de repérage d'états à partir de l'apprentissage

A partir des données — éventuellement transformées, augmentées ou quantifiées en catégories — des modèles locaux sont construits par apprentissage automatique (à gauche). Des indicateurs suivent les caractéristiques de ces modèles (en haut) et permettent de repérer les changements dans les modèles construits donc des zones de continuité (à droite). L'application des méthodes d'apprentissage sur les zones ainsi établies permet d'obtenir finalement une succession de modèles sur des zones délimitées (en bas).

Nos expériences montrent que pour le problème de détermination de zones de stabilité des modèles, le choix exact de la variable à expliquer n'influe pas de façon déterminante. Cela mettrait en relief l'aspect multivarié de l'analyse, au sens où le choix d'une variable à chaque étape de la construction de l'arbre de décision ne se fait pas tout-à-fait indépendamment des autres.

## 5.5 Application

L'application des techniques d'apprentissage dans le cadre de notre méthodologie peut passer par trois biais différents :

- Par l'utilisation des données "statiques", en cherchant à expliquer les valeurs de la variable de classification à partir des données des autres variables. Cette application présente l'inconvénient d'être sujette au synchronisme des données : on suppose que les variations sont vues aux mêmes instants dans le temps. Or, cela n'est pas vérifié en général.
- Par l'utilisation de la tendance et la stabilité calculées avec les techniques du Chap.2 basées sur la recherche de l'échelle caractéristique (Chap.3). On cherchera à expliquer d'un côté les valeurs de la variable de classification à partir des valeurs, tendances et stabilités des autres données. D'un autre côté, on expliquera la *dynamique* de la variable à expliquer à partir des mêmes variables explicatives.
- Par l'application des moyens de symbolisation mis en place au Chap.4, en cherchant à expliquer les symboles associés à la variable à expliquer en fonction des symboles de dynamique des autres variables. Nous aurons alors des modèles de la dynamique de la variable à expliquer.

### 5.5.1 Traitement par fenêtres

Chacune de ces modalités sera déclinée en un traitement fenêtré. Concrètement, nous posons une hypothèse de stationnarité locale et appliquons l'algorithme d'apprentissage sur des population prises sur une fenêtre mobile de taille fixe $\mathcal{T}$.

Cette taille, faute de critères objectifs bien définis, a été déterminée de manière empirique. Plus exactement, nous avons pris un nombre de points de mesure (1000) considéré suffisant pour la construction, du point de vue de l'approximation des probabilités par les fréquences. Faire varier $\mathcal{T}$ de la moitié au double de cette quantité ne change pas qualitativement les résultats obtenus.

Afin de suivre l'évolution des modèles considérés, nous utilisons un certain nombre d'indicateurs de caractéristiques des arbres de décision construits. On peut considérer deux types de mesures sur les arbres de décision : des mesures de qualité et des mesures de structure. Les deux types peuvent être utilisés afin de détecter des changements de régime.

### 5.5.2 Mesures de Qualité

Parmi les mesures de qualité possibles [DF96], nous nous intéressons à la complexité de l'arbre obtenu et à l'erreur de prédiction de celui-ci. La complexité sera mesurée en nombre de noeuds, ce

qui est une bonne mesure pour des arbres binaires comme les nôtres. Comme erreur de prédiction, nous retenons l'erreur intrinsèque de construction.

### 5.5.3 Mesures de Structure

Des mesures de qualité comme les précédentes restent globales par rapport à l'arbre de décision ; d'autres mesures, portant sur la structure de l'arbre, apportent des connaissances plus fines. On peut établir cette structure dans deux espaces : celui, morphologique, des arbres de décision en tant que partition de l'espace des variables explicatives et celui, fonctionnel, des arbres de décision en tant que générateurs de distributions pour la variable à expliquer.

#### Morphologie des Arbres de Décision

Nous proposons une représentation simple de la morphologie d'un arbre de décision par une mise à plat de chacune des variables explicatives. Ceci devient même nécessaire lorsque (comme dans notre cas) les arbres de décision comportent un nombre important de noeuds. Concrètement, nous calculons un *indice de présence* :

$$\mathcal{I}(y) = \sum_{y \text{ à la profondeur } p} 2^{-p}$$

Cet indice prend son sens de l'algorithme de construction des arbres de décision : plus une variable apparaît tôt dans l'arbre, plus elle détient d'information concernant la variable à expliquer.

Certes, des mesures plus significatives peuvent être proposées. Toutefois, l'indice $\mathcal{I}(y)$ est extrêmement facile à calculer sur des arbres déjà construits. De plus, pour les arbres générés à partir de nos données, $\mathcal{I}(y)$ est fortement corrélé à l'entropie de Shannon de la variable à expliquer conditionnée par la variable explicative (à une normalisation près).

Cet indice, ou d'autres de même nature permettent surtout d'avoir une visualisation immédiate des importances relatives de chacune des variables explicatives par rapport à la variable à expliquer.

#### Distributions Induites par un Arbre de Décision

Une caractérisation différente des arbres de décision peut être faite du point de vue fonctionnel. Un arbre de décision sera ainsi considéré comme un générateur de la distribution de la variable à expliquer à partir des distributions des variables explicatives. Pour étudier les variations des arbres de décision au cours du temps, on pourra alors, pour un arbre de décision de référence, procéder à des simulations Monte-Carlo dans l'espace des variables explicatives afin d'obtenir une distribution-type de la variable à expliquer. Ces distributions-type fournissent par leur variation une autre façon d'étudier l'évolution du phénomène modélisé.

Ces approches, bien que complémentaires des précédentes sont prohibitives en temps de calcul.

## 5.6 Résultats

Nous n'avons pas effectué d'application et de validation exhaustives. Ceci pour deux raison majeures.

D'un côté, la base de données Aiddiag a été établie dans des conditions de terrain. Un nombre très réduit d'enregistrements correspond à des protocoles d'investigation clinique bien définis. Par conséquent, les données dont on dispose proviennent d'une population de patients hétérogène, autant en caractéristiques démographiques qu'en pathologie et thérapeutique.

D'un autre côté, la méthodologie propose des moyens d'interprétation. Elle doit donc être évaluée dans cette perspective. Or, le nombre d'enregistrements pour lesquels on dispose d'informations complémentaires est aussi réduit. Nous avons donc peu de renseignements sur l'environnement et les actions subies par les patients dont nous avons les données.

Encore une fois, l'appel à l'expert se fait sentir pour la validation des approches, dans le cadre de protocoles cliniques établis et avec en appui l'ensemble de la documentation retraçant le séjour d'un patient.

Nous présentons par la suite un exemple illustratif d'application, sur un enregistrement documenté [CCPV99].

L'application nécessite de traitements ajustés aux caractéristiques des données, en particulier en nombre de modalités et de valeurs manquantes. Ainsi, nous avons éliminé :

- les données non physiologiques,[7]
- les données pour lesquelles l'enregistrement ne comprend que des données manquantes, ou présentes de façon ponctuelle (dans ce cas, le filtrage de tendance fournira uniquement des données manquantes),
- les données pour lesquelles le calcul de $\tau$ donne des résultats aberrants (*cf.* §3.5),
- les données ne possédant qu'une valeur.

Nous avons alors opéré des constructions d'arbres de décision sur des fenêtres (de taille 1000 points, soit $83min$ ou $1h23min$) glissantes, suivant les trois modalités introduites ci-dessus.

Les Figs. 5.4, 5.5 et 5.6 (pages 110 à 112) représentent les indices de présence définis aux §5.5.3, les mesures de complexité et d'erreur, en fonction de la position dans le temps des fenêtres.

Les zones que l'on peut découper visuellement à partir de ces mesures peuvent être mises en correspondance avec des actions extérieures et des changements notoires documentés manuellement dans le dossier patient.

Précisément, les intervalles (en minutes) [0; 20] [20; 60] [60; 130] [130; 210] [210; 300] [300; 330] et [330; 360] ont des points de rupture contemporains d'actions extérieures : aspirations vers $20min$ et $60min$, des soins (dont un changement des niveaux d'oxygénation) sont appliqués vers $130min$ et $330min$, un nouveau changement dans la ventilation a lieu à $210min$, finalement à $300min$ une stabilisation de la fréquence cardiaque s'instaure (mais nous n'avons pas suffisamment d'éléments

[7] La plate-forme Aiddiag est munie d'un système d'enregistrement sonore avec lequel on espère détecter un certain nombre d'événements ayant lieu dans la chambre [JRC+97] Ces données ne correspondent pas de façon immédiate à notre problématique.

pour préciser son origine, intrinsèque ou extérieure).

Ceci fournit des arguments optimistes quant à la pertinence de la méthodologie dans sa globalité. La correspondance est en plus qualitativement respectée entre les trois représentations proposées. Quelques remarques sont de mise :

- La complexité et l'erreur évoluent en même sens. En effet, lorsque la méthode est adaptée aux données, elle fournit des modèles simples et précis. A l'inverse, lorsque les données ne se prêtent pas à la modélisation, les modèles sont complexes et l'erreur augmente.
- Les modèles construits sont de complexité et d'erreur comparables quelquesoit le mode de calcul.
- Certaines zones sont repérées par les trois modes de traitement, mais quelques unes n'apparaissent que par les traitements faisant intervenir la dynamique.
- Le niveau de contraste entre zones est beaucoup plus net avec les traitements numériques.

La complexité et l'erreur sont difficiles à comparer entre les modes numériques de représentation et le mode symbolique. En effet, pour ce dernier les arbres de décision ont des nœuds pour lesquels en général quatre embranchements existent. Les arbres construits à partir de variables numériques sont *toujours* binaires.

De la même façon, la différence de contraste sur les indices de présence entre les traitements numériques et le symbolique peut être attribuée au fait que l'algorithme cherche à utiliser une variable *une seule fois* sur chaque parcours descendant de l'arbre, dans le cas symbolique.

Les indicateurs que nous avons suivis ne représentent pas de façon adéquate les modèles, lorsque ceux-ci sont symboliques.

Il semblerait donc dans le cadre de l'approche Fig. 5.3 adéquat de suivre les indicateurs construits sur des modèles numériques pour la détermination des périodes de stabilité. Ces états sont établis et caractérisés en terme d'influence des variables explicatives sur une variable à expliquer globale. La recherche de modèles explicites à l'intérieur de chaque état pourra passer par l'approche symbolique.

## 5.7 Conclusion

Nous nous sommes proposé d'établir, au moyen de techniques d'apprentissage, des zones de stabilité dans les enregistrements issus du système Aiddiag. Pour ce faire, nous avons proposé et exploré l'application d'un traitement par fenêtre glissante. A chaque pas de fenêtre, nous appliquons un algorithme d'apprentissage. Le suivi des caractéristiques des modèles construits, reprenant la structure des données, permet de repérer les changements dans le système étudié — pour autant que celui-ci soit effectivement représenté par les données dont on dispose.

Nous avons passé en revue les méthodes disponibles d'apprentissage, en justifiant nos choix par la dissymétrie de notre ensemble de données et par la volonté de limiter le traitement au prétraitement introduit dans les chapitres précédents et à un post-traitement des résultats. De cette façon, le cadre méthodologique reste général et indépendant de la méthode choisie.

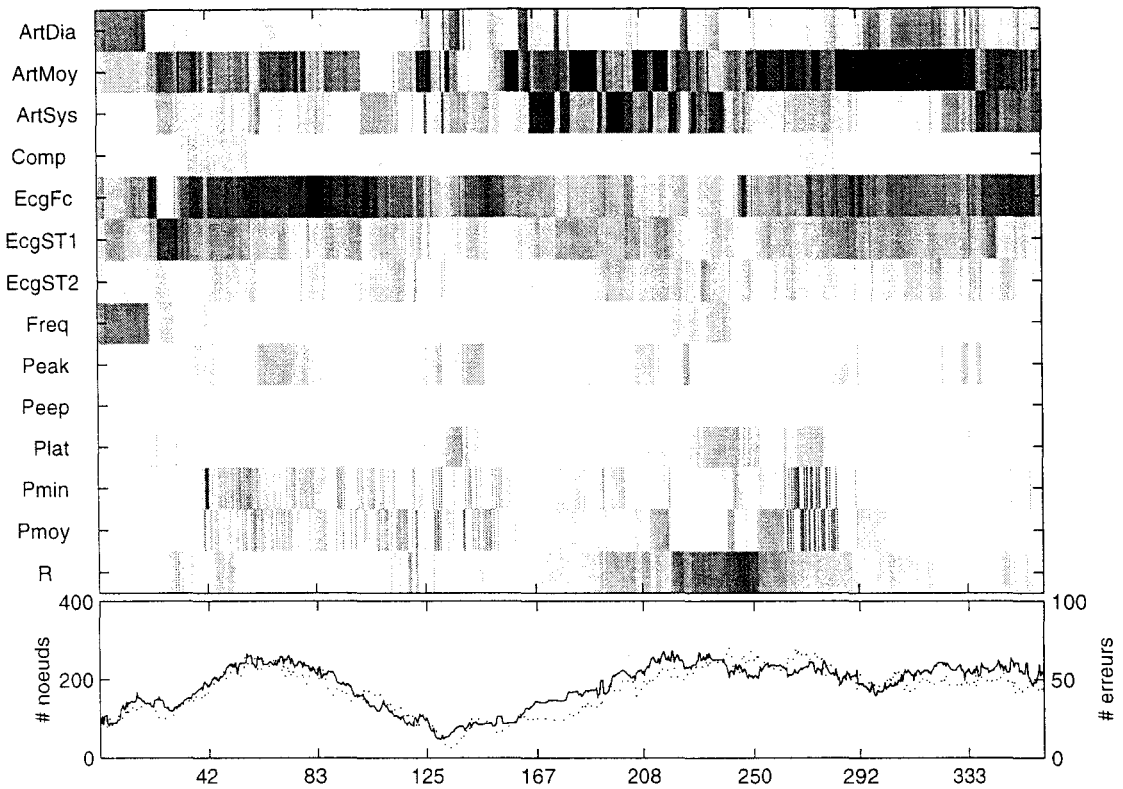



FIG. 5.4 – Exemple d'application du traitement fenêtré, cas statique

En abscisse, le temps, en *min*. Chaque bande représente l'indice de présence $\mathcal{I}$ pour la variable correspondante. La fenêtre glissante utilisée est de $83min$. En bas, la complexité des arbres construits, en nombre de nœuds (trait plein), et l'erreur à la construction, en pourcentage de la taille des populations d'apprentissage (pointillé).

Les arbres construits expliquent les valeurs de $SpO_2$ à partir des valeurs numériques des autres variables. On repère des zones de continuité des modèles construits.

En utilisant le système classique C4.5 nous avons procédé au traitement fenêtré. Nous avons comparé qualitativement les résultats obtenus à partir des données numériques brutes, à partir des données de tendance et stabilité numériques et à partir des symboles définis au Chap.4 (critère $\mathsf{val}\mathsf{m}_{\mathsf{arg}}\mathsf{p}_{\mathsf{rio}}$).

Nous pouvons repérer effectivement des zones, que l'on peut faire correspondre avec des événements extérieurs tels qu'ils ont été relevés manuellement. Le repérage s'est fait de façon visuelle uniquement. Les conditions et les moyens d'automatiser la tâche n'ont pas été étudiés.

La validation de l'approche globale, utilisant les méthodes du Chap.2 à partir de la détermination du temps caractéristique défini au Chap.3 pour définir des symboles comme au Chap.4 et appliquer le principe de repérage de zones de stabilité proposé dans ce chapitre n'a pu être validé à ce stade. Toutefois, les voies de cette validation peuvent être proposées.

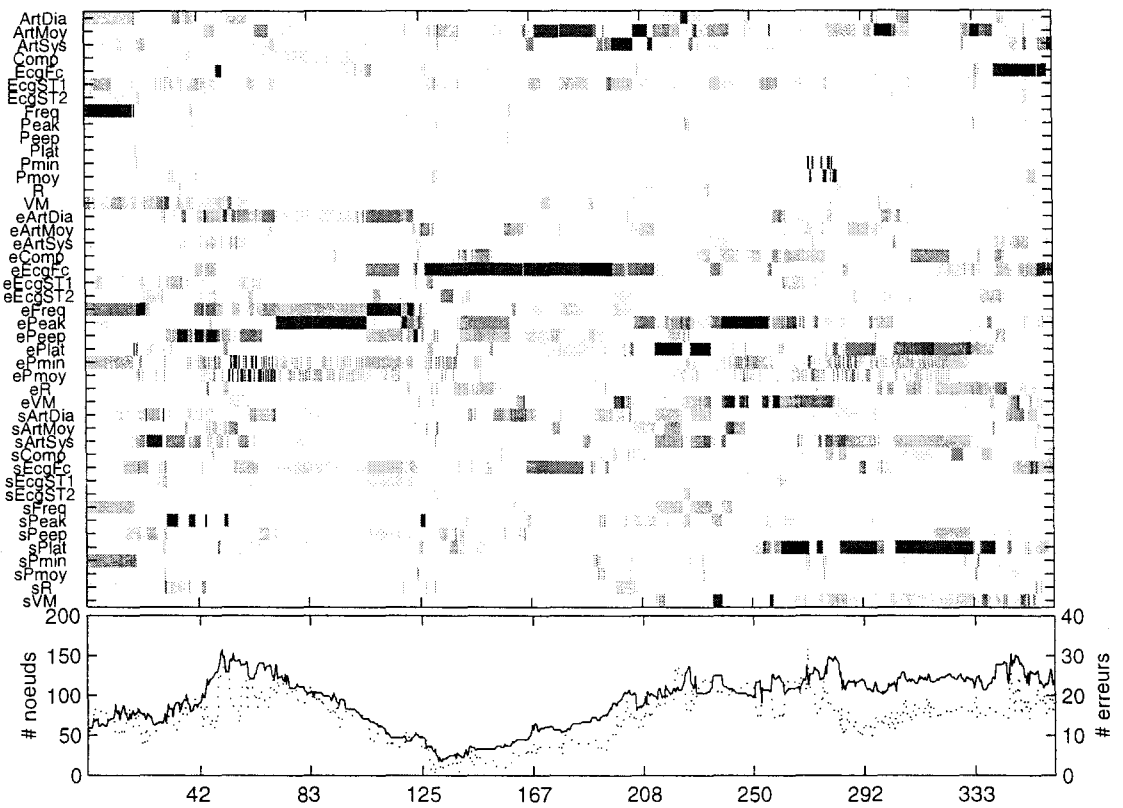



FIG. 5.5 – Exemple d'application du traitement fenêtré, cas dynamique

En abscisse, le temps, en *min*. Chaque bande représente l'indice de présence $\mathcal{I}$ pour la variable correspondante. La fenêtre glissante utilisée est de $83min$. En bas, la complexité des arbres construits, en nombre de nœuds (trait plein), et l'erreur à la construction, en pourcentage de la taille des populations d'apprentissage (pointillé).

Les arbres construits expliquent les valeurs de $SpO_2$ à partir des valeurs numériques des autres variables, de leur tendance numérique (préfixe e) et de leur indicateur de stabilité numérique (préfixe s). Les zones que l'on peut repérer sont plus contrastées que dans le cas statique. De plus, l'interprétation de l'indice de présence en terme de lien permet de dégager nettement les variables qui influent le plus à chaque instant sur les valeurs de $SpO_2$.

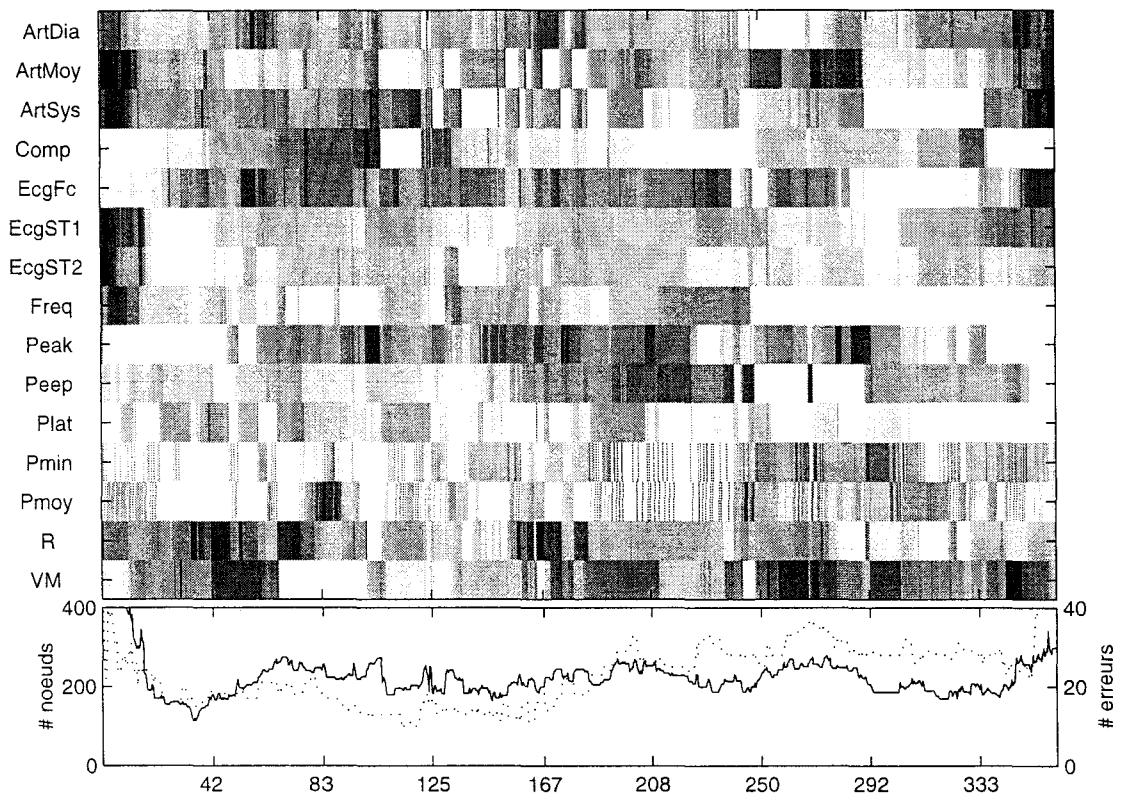



FIG. 5.6 – Exemple d'application du traitement fenêtré, cas symbolique

En abscisse, le temps, en *min*. Chaque bande représente l'indice de présence $\mathcal{I}$ pour la variable correspondante. La fenêtre glissante utilisée est de $83min$. En bas, la complexité des arbres construits, en nombre de nœuds (trait plein), et l'erreur à la construction, en pourcentage de la taille des populations d'apprentissage (pointillé).

Les arbres construits expliquent les symboles de $SpO_2$ à partir des symboles des autres variables. Le contraste est moindre que dans les cas des variables numériques brutes. Les symboles utilisés sont ceux définis par $d_{ist}m_{arg}p_{rio}$ au §4.3.1. Deux catégories supplémentaires ont été introduites : manquant et indisponible, ce dernier reprenant les cas où le calcul de tendance n'est pas possible.

*Yo no sé muchas cosas*
*es verdad*
*digo tan solo lo que he visto.*

León Felipe

# 6

# Discussion

Du positionnement de la méthodologie dans l'existant. Des critiques et des limites qu'elle inspire.

## Sommaire

## 6.1 Quelle information peuvent fournir les données ?

A la base de ce travail, et du projet Aiddiag lui-même, nous nous sommes posé la question fondamentale : en l'absence de modèles de référence, peut-on utiliser les données du monitorage pour affiner la perception du personnel soignant sur l'état et l'évolution des patients ?

Nous sommes en effet partis du principe suivant ($\mathbb{P}_0$) : le personnel soignant *utilise* les données du monitorage, telles quelles, pour le diagnostic et le traitement.

Mais, cette prise en compte se fait de façon intermittente. Les valeurs des paramètres physiologiques sont en effet "lus" uniquement au moment des visites, donc avec des périodes de l'ordre de l'heure dans un service de réanimation. Les données servent par ailleurs au monitorage, *c-à-d.* au suivi de l'état du patient d'après les dépassements des valeurs de seuils établis, et à la mise en place de mécanismes d'alerte en cas de complication ou évolution délétère.

Par ailleurs, d'autres variables *auxquelles nous n'avons pas accès* sont intégrées par le personnel soignant : depuis les résultats d'analyses en laboratoire jusqu'à la couleur de la peau.

Ce que propose de façon la plus immédiate le système d'aide au diagnostic est l'historique d'évolution des données, en plus de leur présentation ponctuelle. Dans ce cadre, nous avons proposé dans ce travail :

1. Des représentations du résumé temporel des événements survenus à chaque paramètre physiologique, en Chap.4.
2. Une technique permettant de repérer des périodes de stabilité dans le lien entre les paramètres physiologiques.
3. Des modèles explicites de ce comportement dans chaque période.

L'évaluation de ces apports par rapport au cadre général de l'aide à la décision nécessite d'un travail de terrain supplémentaire. La difficulté d'évaluation est croissante pour les trois points signalés ci-dessus.

Le premier intègre d'une certaine façon, ses propres moyens de relativisation, donc de validation.

Les deuxième et troisième, intimement liés, posent des difficultés supplémentaires. Celles-ci sont à distinguer entre :

- Les difficultés méthodologiques ;
- Les difficultés liées à la qualité des données de départ.

Le premier point, abordé tout au long de ce travail, a fait l'objet de validations à partir des données, mais devra aussi être évalué par rapport à la connaissance d'expert.

Le deuxième, par contre, pose un certain nombre d'interrogations, déjà soulevées au §1.2. Leur résolution dépasse le cadre de ce travail, au sens où elles font partie de notre principe de base $\mathbb{P}_0$. Toutefois, la méthodologie développée ici pose des jalons pour étudier le problème de la pertinence des données par rapport à la détection d'états *cliniques*. En effet, la recherche des conditions (configurations du rapport entre les variables) dans lesquelles tels gestes cliniques sont employés pourrait passer par le choix judicieux de variables à expliquer et l'application des méthodes présentées.

### 6.1.1 Bruit, quantification, prétraitement, données manquantes

Les données dont nous disposons ont fait l'objet de prétraitements, dont une forte quantification. Celle-ci vient à mettre en défaut un nombre important de techniques d'analyse. Par exemple, il semble correct de lui attribuer (avec la non-stationnarité) l'échec de modélisations univariées en modèles autorégressifs de type ARMA [Gir93]. Les modèles construits sont en effet, pour des ensembles de mesures de l'ordre de ceux que nous avons traités (soit plusieurs heures) pratiquement d'ordre un, de la forme $y_t = y_{t-1} + \epsilon$.

Du fait de la quantification, une forte autocorrélation apparaît dans les résidus de toute approche de modélisation. Il est possible que l'*introduction* artificielle de bruit dans les données permette de saisir ces phénomènes, mais cela dépasse le cadre de ce travail.

De même, les méthodes d'analyse fréquentielle sont mal adaptées aux variations brutales introduites par la quantification. Nous détaillons davantage ce point au §6.2.

### 6.1.2 Échantillonnage

Le mode d'échantillonnage employée pour la collecte des données est celui défini dans la plateforme Aiddiag. Il s'agit, suivant les moniteurs et les paramètres, soit :

- d'une valeur instantanée fournie par le moniteur sur requête, qui correspond en général à la valeur courante calculée par un filtrage (souvent inconnu) sur les mesures continues ; soit
- de la valeur moyenne de la mesure d'un paramètre de forme à partir d'un signal, souvent un paramètre calculé cycle par cycle *cf.* Annexe B ; soit
- de la dernière valeur disponible pour le paramètre en question au moment de la requête de l'échantillon.

Ce mode permet d'avoir une mesure synchrone de tous les paramètres. Il permet donc d'intégrer les échantillonnages de mesures continues ($T°$, $SpO_2$...) et les paramètres issus de l'extraction de caractéristiques de signaux semi-périodiques (à partir des signaux de $EtCO_2$, ECG, débit, pression aérienne et pression artérielle). Si cela est pratique pour tout traitement, et pour la visualisation en particulier, un certain nombre de traitements sont invalidés.

Par exemple, les études de la littérature qui utilisent les signaux de pression aérienne ou d'ECG emploient un mode d'échantillonnage cycle à cycle. Cela pose un certain nombre de problèmes de détection et de reconnaissance de formes, qui sont résolus de façon plus ou moins satisfaisante par les algorithmes des moniteurs [Kos96, CME+91]. Le mode de (sur)échantillonnage qu'ont suivi nos données invalide la pertinence de ces approches. Par exemple, le rapport $H_f/L_f$ calculé sur un flux de données $RR$ obtenu cycle à cycle (*cf.* FigB.1) est un indicateur normalisé et admis dans la pratique, en particulier du fait qu'il peut être interprété en termes physiologiques.[1] Ce rapport, calculé à partir de nos données n'a pas fourni d'informations proches de celles de la littérature.

[1] Les hautes fréquences étant caractéristiques des actions du système sympathique, les basses fréquences sont représentatives des actions para-sympathiques. Le rapport $H_f/L_f$ donne un indicateur de l'équilibre entre ces deux sous-systèmes.

L'échantillonnage des paramètres cycle par cycle à partir du *signal* permettrait d'opérer un certain nombre de ces traitements, en particulier à partir des signaux de $CO_2$, ECG, débit, pression aérienne, pressions artérielles. Il se poserait alors les problèmes largement étayés dans la littérature, de reconnaissance de forme et détection de défauts. De plus, pour nous, il se pose des problèmes de synchronisation des données. On proposera au Chap.7 des voies d'intégration de données calculées cycle à cycle.

Par ailleurs, la question de l'intégration de mesures enregistrées à des périodes beaucoup plus longues se posera à court terme. En effet, les résultats des analyses en laboratoire parviennent à des périodes de l'ordre de quelques heures, voire de la journée. Les mesures de pression artérielle au brassard parviennent toutes les dizaines de minutes, voire toutes les quelques heures. L'intégration de ces mesures, qui font intervenir des notions de tendance qualitativement différentes de celles que nous traitons, devra passer par des moyens d'abstraction temporelle.

### 6.1.3 Des données suffisantes ?

Dans la chaîne de traitement, nous avons séparé les traitements univariés des traitements multivariés. Ces derniers entrent en compte lors de la détermination des zones de stabilité, puis lors de la génération de modèles locaux.

La stabilité, telle que nous l'avons définie en §5.4.3 est celle des rapports mutuels entre variables, en terme de distributions. Les zones de stabilité que la méthodologie permet de repérer sont celles pour lesquelles les rapports entre variables, mesurés par leurs distributions mutuelles deux à deux, sont "sensiblement" les mêmes.

La relation entre cette stabilité et celle que les praticiens peuvent définir n'est pas immédiate, surtout si les données ne sont pas "représentatives". S'attaquer à ces deux questions nécessite d'un travail *a priori* vaste, et qui dépasse notre cadre, voire nos hypothèses fondamentales de travail. Dans la perspective de l'aide à la décision, nous restons dans une approche de transformation des données pour améliorer leur interprétation. La représentativité des données posera un problème lorsque nous essaierons d'aller plus loin, vers la *modélisation* effective.[2] Nous nous rapprochons de la modélisation et du problème de représentativité lorsque nous essayons de dégager des zones de stabilité. Pour les étapes précédentes de la chaîne de traitement, ce problème n'est pas fondamental.

D'un autre côté, nous avons caractérisé les données *a priori* suivant un seul critère : l'existence de paramètres caractéristiques de l'état global du patient. Nous n'avons pas opéré de distinction entre variables d'entrée, d'état et de sortie pour le système observé.

Nous pensons que c'est ce style de caractérisation des variables qui peut aider dans leur prise en compte dans les systèmes d'aide à la décision. Ici nous avons abordé des caractérisations d'assez haut niveau : nous proposons une échelle caractéristique de filtrage pour l'obtention de la tendance. Mais nous avons aussi essayé de caractériser les données suivant leur adaptation aux méthodes de

[2] Même si cette perspective ressort de nos travaux, elle ne l'est que dans un sens modeste : nous nous bornons sciemment à la modélisation comme changement de représentation du flux de données ; nous ne passons que très prudemment à l'inférence des propriétés du système observé.

traitement.

Plus précisément, nous avons considéré les facteurs absence et quantification comme étant des prérequis à toute possibilité d'analyse. Nous avons intégré ces facteurs uniquement de façon qualitative dans notre approche. Néanmoins, il ressort de cette prise en compte la possibilité de caractériser effectivement les flux de données en terme de quantification, et de typifier les données manquantes en terme d'absences ponctuelles ou prolongées, interruptions, intermittences...

La caractérisation suivant ces facteurs devra se faire, comme pour l'échelle de filtrage, en tenant compte de l'échelle d'analyse et des conditions de stabilité des indicateurs (par rapport à la population d'apprentissage, aux enregistrements...).

## 6.2 Échelles et Ondelettes

La prise en compte de l'échelle des phénomènes a été mise à jour et abordée particulièrement par les décompositions en ondelettes, et surtout les ondelettes temps-échelle [Mey94, pp.173–189]. Voir [Hub95] pour une introduction aux décompositions par ondelettes. Cette famille de décompositions fournit en plus des moyens de visualisation riches.

La projection que nous faisons sur les vecteurs $[s_{n,1} \dots s_{n,n}]'$ et $[p_{n,1} \dots p_{n,n}]'$ n'est pas équivalente à une décomposition, au sens des ondelettes. En effet, les propriétés fondamentales définissant une famille de fonctions comme permettant une décomposition — donc une reconstruction — ne sont pas vérifiées par ces filtres. Voir [Mey90] pour les propriétés fondamentales des ondelettes. D'après Morlet-Grassman, une ondelette est une fonction dont la transformée de Fourier $\Psi(\xi)$ vérifie :

$$\int_0^\infty \left|\Psi(t,\xi)^2\right| \frac{dt}{t} = 1$$

Les filtres en rampe dont nous nous servons ne vérifient pas cette propriété.

Suivant la caractérisation de Meyer [Mey94], nous pouvons expliquer cela par le fait que nos filtres n'oscillent pas suffisamment, et ne sont pas suffisamment atténués sur les bords, pour pouvoir accéder à la propriété de décomposition.

Notre but n'est pas de déterminer l'*oscillation* locale des données en fonction de l'échelle d'observation. Nous avons éliminé *a priori* et par suite d'expériences préliminaires les approches basées sur les représentations fréquentielles.

Peu de données présentent un comportement fréquentiel intéressant, avec l'exception notable du *signal* de pression intra-cranienne [AC90]. Nous ne disposons que de la valeur moyenne de ce signal sur un nombre très réduit (trois) de patients. Le monitorage de ce paramètre n'est réalisé que pour des pathologies particulières.

Par ailleurs, le spectre de puissance du signal *RR* est riche d'informations et est aujourd'hui établi comme moyen de surveillance, en anesthésie particulièrement [CME+91]. Pourtant, le mode

d'échantillonnage employé par les moniteurs ne fait parvenir à nos données qu'une information transformée, pour laquelle les traitements désormais classiques de détermination du rapport $H_f/L_f$ deviennent inefficaces. En effet, ils sont basés sur un traitement cycle à cycle à partir du signal de l'ECG.

Après un traitement exploratoire sur un certain nombre de paramètres et de patients tirés au hasard, le contenu fréquentiel n'a fourni aucune information caractérisant ces flux de données : fréquences dominantes, changements de contenu spectral haute-basse fréquence, caractérisation de la couleur du spectre...

L'utilisation de techniques d'ondelettes pour la recherche d'une échelle caractéristique revient à la recherche d'une fréquence/échelle caractéristique, au moment de l'agrégation dans le temps [Wil97]. Mais ce serait perdre la richesse de la décomposition. Une approche possible, qui n'a pas été développée est celle de décomposer le signal puis n'utiliser que les composantes de haute énergie, à la manière de la compression par ondelettes [SS96].[3] Sur chacune des composantes de plus haute énergie, on peut reconstruire le signal, puis opérer comme ci-dessous avec chaque composante dominante.

Cette approche nécessiterait la réécriture des algorithmes pour tenir compte des données manquantes. En plus, elle présente l'inconvénient de multiplier le nombre de variables sur lesquelles on opère.

Il est possible de diminuer ce nombre, *c-à-d.* la dimension de l'espace de décomposition. En condensant, à la manière de §4.1 les composantes autres que la dominante, on peut fournir un indicateur de l'énergie qui n'est pas représentée par la composante dominante. On reviendrait ainsi à une représentation bidimensionnelle reprenant les notions de tendance et stabilité. Sauf que cette tendance représenterait à chaque instant le comportement oscillant le plus saillant. Or, *cf.* §2, cette notion de tendance dominante en tant qu'oscillation ne rentre pas dans les sémantiques identifiées pour exprimer la dynamique des variables physiologiques. Il serait intéressant d'appliquer ces principes sur des données au comportement oscillatoire bien identifié par la pratique médicale, mais ce n'est pas l'objet de ce travail.

D'un autre côté, *cf.* §1.2, nous avons affaire à des paramètres caractéristiques extraits de signaux. L'analyse par ondelettes dans la littérature s'est toujours occupée de signaux : mesures à haute fréquence et de haute précision (bruit ou pas). Voir [Aka97] pour des travaux récents appliqués au domaine biomédical. Nos séries chronologiques peuvent être considérées de haute fréquence par rapport à l'échelle temporelle des phénomènes physio-pathologiques. Mais la quantification qu'elles subissent dans les moniteurs introduit des artefacts haute fréquence qui rendent difficile l'étude par les méthodes de décomposition temps-fréquence.

Certaines données pourraient correspondre à la notion typique de signal — en terme de précision/quantification en particulier. Toutefois, il ne serait pas aisé de les privilégier par rapport aux autres variables, en leur appliquant un traitement échelle par échelle comme ci-dessus. L'intégration finale n'en serait que plus difficile. De plus le traitement que nous avons exposé devra de toute

[3] L'essentiel de l'énergie est dans les hautes fréquences/petites échelles, parfois uniquement à cause d'artefacts issus de la quantification.

façon être réalisé pour les autres paramètres.

## 6.3 Jusqu'où peut-on se baser sur les données exclusivement ?

Nous nous sommes proposé comme programme d'établir des moyens d'analyse en retardant le plus en aval l'introduction de la connaissance d'expert.

Ceci partant du constat, mis en évidence par [CRC+89] et les travaux menés au sein du projet Aiddiag [RCJ+94], de cinq problèmes fondamentaux dans la modélisation de la connaissance d'expert :

Généralité Les modèles explicités sont souvent des modèles généraux. Ils portent davantage sur la connaissance formalisée de la science publique[4] que sur l'expérience et la pratique individuelles concrètes.

Complétude Souvent les modèles considérés ne recouvrent pas l'ensemble des situations et, en particulier, passent sous silence les éléments vus comme évidents.

Consistance Les modèles explicités par les experts sont souvent issus à la fois de ces modèles généraux, adaptés aux cas particuliers et à l'expérience de chaque expert. Les divergences entre les deux origines apparaissent dans le corpus de connaissances comme : au mieux, des cas particuliers ; au pire, comme des incohérences.

Cohérence Les avis d'experts différents peuvent diverger, malgré le point de généralité soulevé ci-dessus.

Explicitation La connaissance, même lorsqu'elle ne pose pas les problèmes précédents, est formalisée de façon à prendre en compte, entre autres, les mesures de paramètres physiologiques. Mais cette formalisation n'est pas immédiatement transposable en des traitements à partir des mesures réalisées par les moniteurs. En effet, les notions relatives, floues, de formes d'évolution, de tendance, de lien mutuel, de valeur anormale,...ont besoin d'être transcrites en des critères précis pour pouvoir être implémentées dans un système d'aide à la décision.

On essaye de contourner ces problèmes, ou plutôt de retarder leur apparition le plus en aval possible de la chaîne de traitement à partir des données. Cela nécessite l'ajustement des méthodes existantes, la proposition de méthodes nouvelles et la caractérisation et l'évaluation pour la validation des choix effectués.

Plus précisément, nous avons misé sur l'utilisation de méthodes d'apprentissage, de façon à introduire des modèles symboliques du comportement des variables à partir des données.

Les méthodes disponibles (Chap.5) ne présentent pas de moyens d'intégration immédiate de notions de dynamique. Nous avons donc proposé des moyens de prendre en compte la dynamique à partir des notions de tendance et stabilité.

[4] Au sens de Holton [Hol78] : il s'agit du corpus de connaissances établi et partagé par l'ensemble d'une communauté scientifique. C'est la remise en cause d'éléments de ce corpus qui constitue les révolutions scientifiques au sens de Kuhn [Kuh62].

Le problème fondamental de non-stationnarité a été abordé par le traitement fenêtré, suivi d'une reconnaissance de la continuité des modèles. Cette reconnaissance s'est faite de façon visuelle jusqu'ici.

Pour arriver à ce stade, il a fallu effectuer des choix de critères. Pour l'extraction de la tendance et de l'échelle caractéristique, nous avons proposé des méthodes basées sur les données. Celles-ci ont été basées sur des traitements et des validations *a posteriori*, mais avec des choix de départ et des partis-pris. En particulier, des principes clairs de :

Simplicité Nous avons proposé une chaîne de traitements élémentaires, où chacun peut être utilisé de façon indépendante dans d'autres chaînes. Ainsi les méthodes de calcul efficace de tendance (Chap.2, Annexe C, Annexe D) sont génériques au traitement de données échantillonnées. La notion de temps caractéristique introduite au Chap.3 peut servir de base à des filtrages n'ayant pas pour but l'extraction de symboles. La décomposition tendance *vs.* stabilité, qui permet une transformation numérique-symbolique, peut servir de moyen de symbolisation dans d'autres systèmes de raisonnement automatique que celui présenté.

Intelligibilité Nous avons privilégié des traitements faisant intervenir des notions proches de celles communément employées par les praticiens du domaine. Ainsi, les approximations de tendance se font au premier ordre. Pour la détermination des échelles caractéristiques, des notions de risque statistique sont utilisées, au détriment de notions d'accélération. Des représentations symboliques en arbres de décision sont utilisées au lieu d'autres formalismes.

Lisibilité Nous avons insisté sur les possibilités de visualisation pour l'interprétation, de façon à minimiser la surcharge cognitive introduite par des systèmes de représentation des données éventuellement peu familiers. Chaque étape produit des moyens propres de visualisation pour la validation.

Ces principes ont pour objectif commun la mise à disposition, à chaque étape du traitement, des éléments permettant leur adoption par l'utilisateur médical. En effet, la décomposition du problème en phases de traitement permet de s'occuper séparément de chaque étape, indépendamment des suivantes. Les efforts dans la lisibilité et l'intelligibilité portent sur la capacité de l'utilisateur de relativiser, interpréter et éventuellement mettre en défaut les résultats fournis par le système.

Dans l'ensemble, nous avons posé des hypothèses sur des critères cognitifs généraux et spécifiques au domaine. Mais nous n'avons pas introduit de connaissance de spécialiste.

Cette approche trouve ses limites assez tôt. Ainsi, la transformation numérique-symbolique que nous avons proposée en lignes générales ne peut être validée et spécifiée (catégories, modalités exactes, *cf.* §4.4) uniquement à partir des données. De même, le traitement utilisé au Chap.5.5 ne peut être validé sans faire appel à un expert du domaine.

Nous nous sommes donc trouvés confrontés, dès le passage du numérique au symbolique, au besoin d'introduire de la connaissance d'expert. Il est difficile pour nous, en l'état actuel, de proposer des moyens de retarder davantage le passage par l'expert.

*Rendre visite, pas résoudre.*

Friedrich Dürrenmatt, *Le Retraité*

# 7

# Conclusions et Perspectives

Où l'on met en évidence que le chemin restant à parcourir est long et vaste.

## Sommaire

Nous sommes partis d'enregistrements des paramètres physiologiques d'un patient en réanimation adulte. Nous avons proposé une chaîne de traitement permettant de retracer l'historique de cet enregistrement en terme d'une succession d'états de stabilité. Cette stabilité a été définie comme étant celle des caractéristiques d'un modèle symbolique construit à partir des données.

Nous avons à chaque étape — calcul de la tendance à partir d'une échelle caractéristique, conversion symbolique, apprentissage — établi des moyens d'évaluation et de visualisation pour l'aide à la décision. Nous avons appliqué et illustré ces traitements sur un certain nombre d'enregistrements de la base Aiddiag.

Un certain nombre de points essentiels restent en perspective :

1. Les techniques de détection des zones de stabilité ;
2. Les critères de choix de la technique d'apprentissage *cf.* §5 ;
3. Les paramètres de conversion numérique-symbolique *cf.* §4 :
   (a) Les catégories à utiliser ;
   (b) L'introduction à ce niveau de la connaissance du domaine, par exemple les seuils de normalité ;
   (c) Les moyens de détermination des seuils, éventuellement par des méthodes automatiques.
4. Les modalités d'intégration de données échantillonnées à des fréquences différentes ;
5. Les paramètres d'intégration de la chaîne de traitement dans le système Aiddiag.

Voyons ces points plus en détail.

## 7.1 Quelle définition de "stabilité" ?

Dans l'aproche présentée, la caractérisation des enregistrements en tant que succession d'états a été réalisée de façon visuelle. Si cette approche peut être justifiée lors de l'utilisation par un expert, il faudra proposer des moyens de détection automatique des changements de modèle.

Pour cela, plusieurs voies sont ouvertes.

D'un côté, par l'utilisation des indices de présence, le flux de données originel a été converti en un flux de données dans lequel on cherche des ruptures de niveau. On est dans le cadre des approches classiques de détection à partir d'un signal. Ces approches ont été établies pour la détection en ligne.

D'un autre côté, par des moyens de classification non supervisée, ou semi-supervisée — *c-à-d.* par l'utilisation d'outils de classification non supervisée et la validation avec l'aide d'un expert — ces zones peuvent être repérées de façon automatique ou semi-automatique. Si cette approche est pratique hors ligne, les méthodes de classification adaptées à un fonctionnement incrémental ou en ligne sont rares.

## 7.2 Précision du choix des techniques d'apprentissage

Nous nous sommes basés dans ce travail exclusivement sur des techniques d'apprentissage supervisé. Les raisons pour cela ont été données au §5.4.3. Pourtant, une approche de classification dans le même cadre — tendance, symbolisation, détection de zones de stabilité — devrait fournir des résultats permettant de saisir la validité de l'approche dans son ensemble.

D'un autre côté, la question du choix de la méthode exacte de classification supervisée, C4.5 en l'occurrence s'est faite aussi par des critères *a priori* de robustesse, documentation, souplesse. Ce choix doit encore être validé *a posteriori* par la comparaison avec les autres méthodes revues au Chap.5. En effet, il n'y a pas de résultat définitif et général concernant les mérites comparés des différentes méthodes existantes d'induction par arbres de décision. Nous avons un problème précis avec des données particulières, il semble donc adéquat de chercher la méthodologie qui s'ajuste le mieux à nos problèmes.

## 7.3 Détermination des paramètres de conversion numérique-symbolique

Nous nous sommes trouvés confrontés au problème de définition des catégories et des moyens de représentation de ces catégories. Faute de critère objectif de détermination de ces paramètres, il faudra explorer le recours à la connaissance d'expert. Plus précisément, nous pourrons avoir recours à des catégories définies *a priori* par des experts, en fonction des séquences qu'ils chercheront à identifier.

Ce moyen de catégorisation pourra être introduit à partir de la projection tendance *vs.* stabilité. Dans ce cas, le reste de la méthodologie reste inchangé. D'un autre côté, la catégorisation pourra intervenir de façon indirecte, par ajustement semi-automatique des seuils, en fonction d'une classification introduite *après l'étape d'apprentissage* par l'expert. Ceci introduit une boucle de rétroaction dans la méthodologie exposée ici, entre la sortie des systèmes d'apprentissage et les paramètres de conversion numérique-symbolique.

## 7.4 Intégration des données acquises sous d'autres modalités

Nous avons détaillé une méthodologie qui se prête à l'étude de séries chronologiques synchrones. Pourtant, nous voudrions intégrer ce travail dans un objectif plus poussé : la construction de modèles locaux tenant compte de l'ensemble des données disponibles. Ceci comprend les données déjà employées, mais aussi les résultats d'analyses à partir de prélèvements : sang, ponctions,... Ces données sont disponibles, pour des raisons techniques, à des périodes d'échantillonnage de l'ordre de quelques heures. L'intégration de ces données extraites hors ligne avec le flux en ligne

dont nous nous sommes occupés pose des problèmes de resynchronisation.

En effet, pour pouvoir dégager des traitements impliquant des données acquises en haute et en basse fréquence, il faudra procéder à la mise en place : soit de systèmes de raisonnement temporel au niveau symbolique, soit de moyens d'interpolation et/ou extrapolation au niveau numérique. Dans les deux cas, des recherches supplémentaires sont nécessaires.

En ce qui concerne les données physiologiques extraites des signaux, leur calcul est effectué cycle à cycle, sur des grandeurs cycliques comme l'ECG, le débit aérien, les pressions,... Les valeurs dont nous disposons sont celles de la dernière mesure en cycle à cycle, ou bien des grandeurs dérivées par filtrage de celles-là.

Disposer des signaux en provenance des capteurs nous permettrait d'avoir accès à des données cycle à cycle, sans filtrage (ou alors en déterminant nous-mêmes le filtrage). Les traitements univariés peuvent être effectués sur des données calculées cycle à cycle. Pourtant, la fréquence de mise à disposition de ces données serait variable, fonction directe de la fréquence du signal cyclique sous-jacent.

Du fait du filtrage pour obtenir des tendances, nous introduisons une certaine interpolation et un lissage. Ceux-ci permettraient de réaliser un sur-échantillonnage, à la période des acquisitions des autres paramètres.

## 7.5 Implémentation dans la plate-forme Aiddiag

Nous avons opéré dans une perspective de mise en ligne des traitements développés. Ceci est évident pour les méthodes d'extraction de tendance et de conversion numérique-symbolique. De même, le choix de la méthode d'apprentissage tient compte du compromis entre l'efficacité de l'implémentation d'une part, et d'autre part la sophistication des modes de représentation et la richesse de l'espace de recherche. Ainsi nous avons préféré des systèmes d'induction utilisant des partitionnements droits aux méthodes plus fines de partitionnement oblique ou non-linéaire.

Pourtant, l'implémentation effective intégrée à la plate-forme Aiddiag nécessite encore de régler en ligne certains problèmes. Notamment, la caractérisation *a priori* des données en fonction de leur pertinence dans l'un ou l'autre traitement nécessite de moyens de détection que nous avons esquissés mais le problème ne sera résolu que par la confrontation avec les conditions de terrain. En effet, nous avons vu que les caractéristiques des données — en particulier la répartition des données manquantes, la quantification, le nombre de modalités — peuvent mener à des comportements aberrants. Ceux-ci sont faciles à détecter *a posteriori*, mais il sera obligatoire de les détecter avant traitement, sous peine d'introduire des artefacts pour l'interprétation subséquente.

## 7.6 Perspectives

Nous avons proposé une méthodologie d'analyse motivée par les objectifs de l'aide à la décision et modulée par les caractéristiques des donnés disponibles dans le cadre du monitorage clinique.

Elle a été exposée comme un moyen de synthétiser hors ligne le séjour d'un patient tel qu'il est vu à partir des données du monitorage. Les conditions d'extension vers un système de détermination de changements d'état en ligne ont été posées, et des voies de solutions esquissées.

L'approche modulaire qui a été exposée permet une application partielle, par paliers. En effet, les techniques de régression par filtrage (particulièrement dans leur version polynomiale), le concept et la méthode d'extraction de l'échelle caractéristique, l'utilisation de la projection en tendance *vs.* stabilité pour la visualisation et la symbolisation, et le schéma de suivi de modèles dynamiques admettent des variantes, des ajustements mais surtout ils peuvent être transposés dans des contextes autres que le nôtre.

Le choix de travailler avec des éléments qui puissent être isolés nous a permis d'effectuer des validations, des ajustements partiels et des choix techniques qui influencent peu ou prou l'ensemble de la chaîne. De cette façon, nous avons un schéma global dans lequel l'ajustement des différentes pièces est possible. L'implémentation à terme d'un système complet en ligne basé sur cette chaîne méthodologique devra passer par l'adéquation de chaque élément à la problématique. Ce travail nécessitera des conditions de référence contrôlées pour la validation, ainsi que des mécanismes d'intégration des critères d'expert en terme d'accessibilité du système, d'aisance d'interprétation et d'utilité.

# A

# Paramètres Physiologiques Disponibles

Nous présentons une liste des paramètres couramment disponibles dans les enregistrements de la plate-forme d'acquisition Aiddiag. Il s'agit des variables repésentant des paramètres physiologiques. Pour certaines, il s'agit de mesures échantillonnées d'une quantité. Pour d'autres, il s'agit de paramètres caractéristiques tirés de la forme typique des signaux. Voir Annexe B pour plus de détails.

Nous présentons aussi un exemple d'enregistrement de ces paramètres en Fig. A.1. Celui-ci permet d'illustrer les caractéristiques des données, détaillées en §1.2.

| Paramètres Respiratoires | | | |
|---|---|---|---|
| Paramètre | Nom courant | Unité | Précision |
| $V_T$ | Volume Courant | l | 0.01 |
| $f_R$ | Fréquence Respiratoire | $min^{-1}$ | 1 |
| $V_E$ | Ventilation Minute | $l.min^{-1}$ | 0.1 |
| $P_{aw}max$ | Pression Maximale | mbar | 1 |
| $P_{aw}plat$ | Pression Plateau | mbar | 1 |
| PEP | Pression Expiratoire Positive | mbar | 1 |
| $P_{aw}min$ | Pression Minimale | mbar | 1 |
| $P_{aw}moy$ | Pression Moyenne | mbar | 1 |
| $FIO_2$ | Fraction d'Oxygène Inspiré | % | 1 |
| Crs st | Compliance Statique | $ml.mbar^{-1}$ | 1 |
| Raw | Résistance | $mbar/(l.s^{-1})$ | 1 |
| $V_E$spont | Ventilation Spontanée | $l.min^{-1}$ | 0.1 |
| $f_R$spont | Fréquence Spontanée | $min^{-1}$ | 1 |
| $T°$ | Température | °C | 0.1 |

| Paramètres Sanguins | | | |
|---|---|---|---|
| Paramètre | Nom courant | Unité | Précision |
| pH | pH sanguin | | 0.01 |
| $PCO_2$ | Pression Partielle de $CO_2$ | mmHg | 0.1 |
| $PO_2$ | Pression Partielle de $O_2$ | mmHg | 1 |
| $T°$ | Température | °C | 0.1 |
| $HCO_3^-$ | Concentration Carbonate | $mmol.l^{-1}$ | 0.01 |
| BE | Base Excess | $mmol.l^{-1}$ | 0.01 |
| $SaO_2$ | Saturation d'Oxygène | % | 0.01 |

| Paramètres des moniteurs cardio-vasculaires | | | |
|---|---|---|---|
| Paramètre | Nom courant | Unité | Précision |
| ECG FC | Fréquence Cardiaque | $min^{-1}$ | 1 |
| ECG ANO | Anomalies Détectées | | |
| $ST_1/ST_2$ | Alignement segment ST | mV | 0.01 |
| TA SYS | Tension Artérielle Systolique | mmHg | 1 |
| TA DIA | Tension Artérielle Diastolique | mmHg | 1 |
| TA Moy | Tension Artérielle Moyenne | mmHg | 1 |
| AP SYS/DIA/Moy | Tension Artère Pulmonaire | mmHg | 1 |
| ART SYS/DIA/Moy | Tension Artérielle | mmHg | 1 |
| PRS SYS/DIA/Moy | Mesure Secondaire de Pression | mmHg | 1 |
| PVC | Pression Veineuse Centrale | mmHg | 1 |
| OD | Pression Oreillette Droite | mmHg | 1 |
| OG | Pression Oreillette Gauche | mmHg | 1 |
| PIC | Pression Intra-cranienne | mmHg | 1 |
| $SpO_2$ | Oxymétrie de pouls | % | 1 |
| FinExp RESP | Fréquence Respiratoire | $min^{-1}$ | 1 |
| FinExp % AGENT | Conc. gaz autres que $CO_2$ | % | 0.1 |
| FinExp % | Conc. $CO_2$à l'expiration | % | 0.1 |
| RESP FREQ | Fréquence Respiratoire | $min^{-1}$ | 1 |

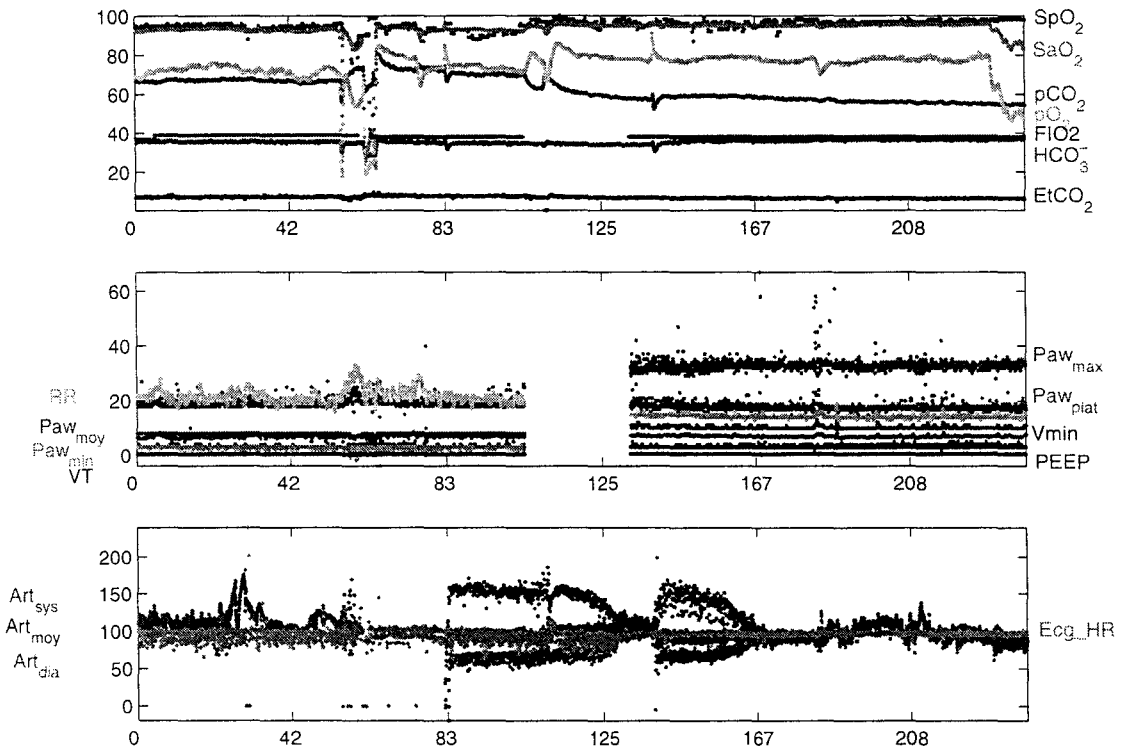



FIG. A.1 – Exemple d'enregistrement Aiddiag

En haut, les paramètres disponibles sur les échanges gazeux. Au milieu, les paramètres respiratoires. En bas, les paramètres hémodynamiques. Les unités sont celles des paramètres, *cf.* les tables précédentes. En abscisse, le temps en minutes.

Notez les zones de données manquantes, ainsi que la forte quantification, visible sur les paramètres de faible plage de variation.

# B

# Extraction des paramètres à partir des signaux

Parmi les paramètres physiologiques dont nous disposons, un certain nombre sont issus de la détection de paramètres caractéristiques de la forme typique des signaux. Nous schématisons ici l'extraction de paramètres caractéristiques à partir des signaux de pression aérienne, de débit (Fig. B.2) et d'ECG (Fig. B.1).

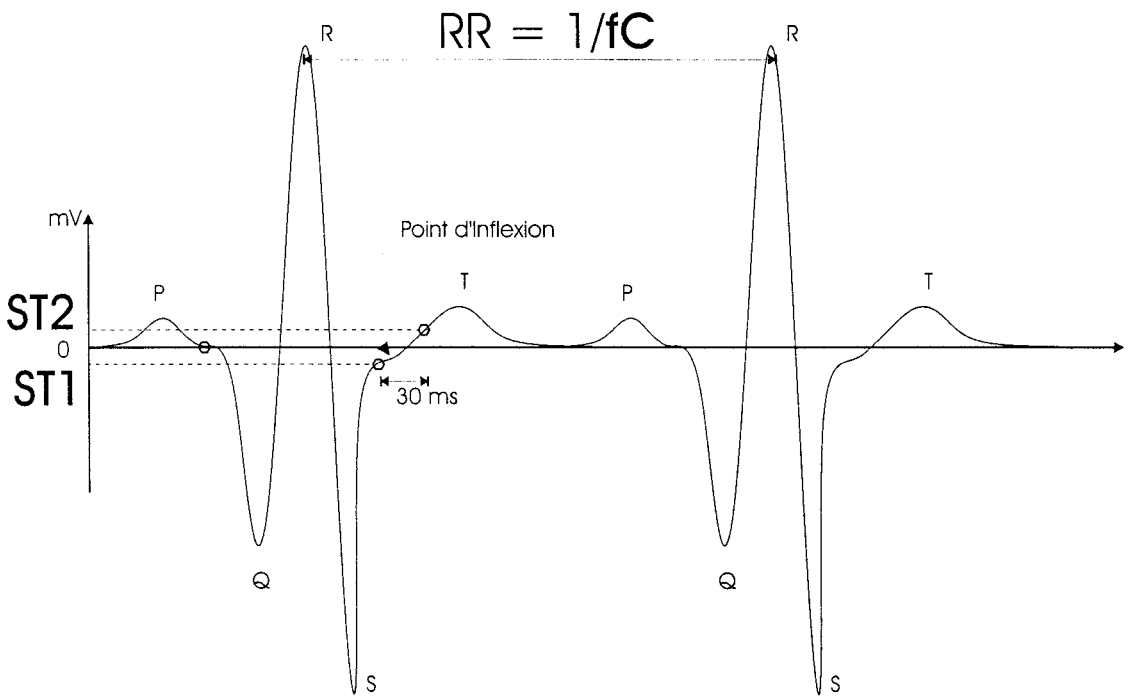



FIG. B.1 – Extraction de $F_c$, $ST_1$ et $ST_2$ à partir de l'ECG

A partir du signal d'ECG, la détection des points $R$ permet d'estimer la fréquence cardiaque $F_c$. Les points $ST_1$ et $ST_2$ servent comme points de référence pour estimer l'allure du segment $ST$. On prend le repère $0V$ à partir de $PQ$ : il y a toujours une zone de stabilisation entre $P$ et $Q$. On détecte le point d'inflexion qui suit le pic S. Il détermine $ST_1$ ; $30ms$ (ou $60ms$ suivant le constructeur) plus tard, on mesure $ST_2$.

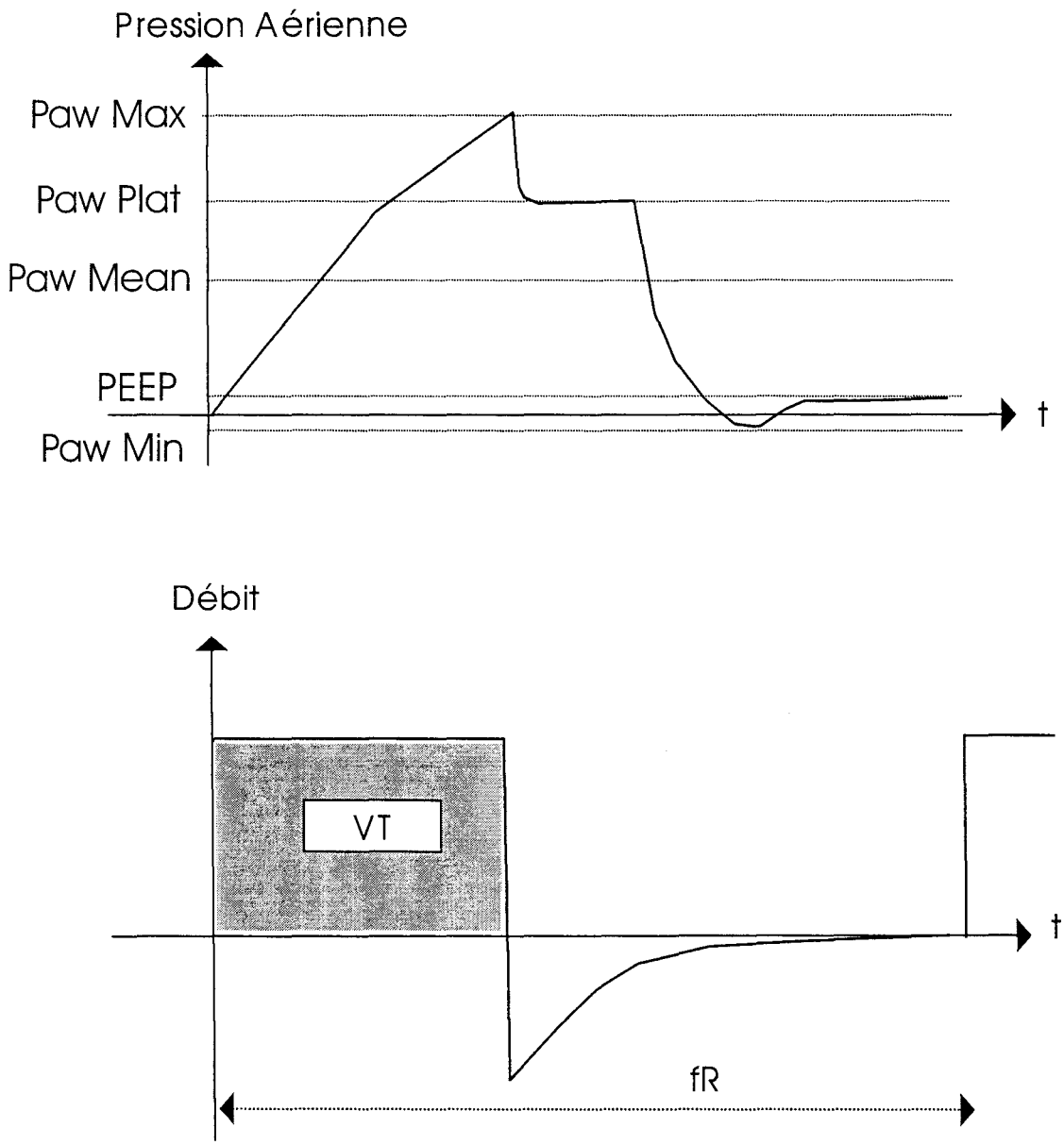



FIG. B.2 – Paramètres ventilatoires à partir du débit et de la pression aérienne
A partir des signaux de débit et de pression, un certain nombre de paramètres sont extraits des formes typiques des signaux. Ici les courbes reprennent les formes typiques de celles de la ventilation contrôlée en débit. Les paramètres de pression dérivés du signal de pression sont : maximale, minimale, de fin d'expiration (Positive End-Expiratory Pressure), de plateau et moyenne. A partir du débit est calculé le volume courant $V_T$ et le temps total, duquel est dérivé la fréquence respiratoire $f_R$.

# C

# Extensions de la régression en tant que filtrage à l'approximation polynomiale

Nous proposons un travail semblable à celui présenté en §2.3 et §2.4 pour des approximations polynomiales d'ordre supérieur à un. Nous détaillons le cas du polynôme d'ordre deux et posons le principe de calcul pour des polynômes de degré quelconque.

## C.1 Approximation parabolique

Pour une approximation polynomiale d'ordre deux (parabolique), on a :

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 2^2 & 2 & 1 \\ \vdots & \vdots & \vdots \\ n^2 & n & 1 \end{bmatrix}$$

Alors :

$$X'X = \begin{bmatrix} \frac{n(2n+1)(n+1)(3n^2+3n-1)}{30} & \frac{n^2(n+1)^2}{4} & \frac{n(2n+1)(n+1)}{6} \\ \frac{n^2(n+1)^2}{4} & \frac{n(2n+1)(n+1)}{6} & \frac{n(n+1)}{2} \\ \frac{n(2n+1)(n+1)}{6} & \frac{n(n+1)}{2} & n \end{bmatrix}$$

Et donc :

$$(X'X)^{-1} = \frac{3}{n(n-1)(n-2)} \begin{bmatrix} \frac{60}{(n+1)(n+2)} & -\frac{60}{n+2} & 10 \\ -\frac{60}{n+2} & \frac{4(2n+1)(8n+11)}{(n+1)(n+2)} & -6(2n+1) \\ 10 & -6(2n+1) & 3n^2+3n+2 \end{bmatrix}$$

L'approximation parabolique de la tendance des données $Y$ à l'instant $t$ et à l'échelle $n$ est donnée par :

$$\begin{bmatrix} Q_{n,t} \\ L_{n,t} \\ C_{n,t} \end{bmatrix} = \begin{bmatrix} q_{n,1} & \cdots & q_{n,j} & \cdots & q_{n,n} \\ l_{n,1} & \cdots & l_{n,j} & \cdots & l_{n,n} \\ c_{n,1} & \cdots & c_{n,j} & \cdots & c_{n,n} \end{bmatrix} \begin{bmatrix} y_{t-n+1} \\ \vdots \\ y_{t-1} \\ y_t \end{bmatrix}$$

où $\forall j \in [1, n]$

$$q_{n,j} = 30\,\frac{6j^2 - 6(n+1)j + (n+1)(n+2)}{(n^2-4)(n^2-1)n}$$

$$l_{n,j} = -6\,\frac{30(n+1)j^2 - 2(8n+11)(2n+1)j + 3(2n+1)(n+2)(n+1)}{(n^2-4)(n^2-1)n}$$

$$c_{n,j} = 3\,\frac{10j^2 - 6(2n+1)j + 3n^2 + 3n + 2}{(n-1)(n-2)n}$$

Ces coefficients peuvent être précalculés. On obtient alors un filtre linéaire qui permet de déterminer les coefficients de l'approximation parabolique à un ensemble de données régulièrement échantillonnées.

Le calcul est réalisé en temps linéaire par rapport à $n$, dans l'implémentation la plus immédiate.

## Implémentation incrémentale

On peut aussi reprendre la recherche d'une procédure incrémentale pour traiter ces trois composantes.

Puisque $q_{n,j}$ (*resp.* $l_{n,j}$, $c_{n,j}$) sont quadratiques en $j$, l'expression de $q_{n,j+1}$ en fonction de $q_{n,j}$ ne sera pas une constante dépendant uniquement de $n$ ; il faudra passer aux différences secondes,

pour calculer $q_{n,j+1}$ de façon directe en fonction de $q_{n,j-1}$. De façon plus lourde qu'au §2.4, on aura une expression permettant de calculer les coefficients de la parabole de façon incrémentale et en temps constant par rapport à $n$.

Précisément :

$$q_{n,j+1} - 2\,q_{n,j} + q_{n,j-1} = \frac{360}{n\,(n^2-1)\,(n^2-4)}$$

Alors les coefficients du terme quadratique de l'approximation parabolique à l'instant $t$ et à l'échelle $n$ sont donnés par :

$$\begin{aligned} Q_{n,t+1} - 2\,Q_{n,t} + Q_{n,t-1} = K_q\,(\mathbb{S}_{n,t} - y_t - y_{t-n+1}) \quad &+ \quad q_{n,1}\,y_{t+1} \\ &+ \quad (q_{n,2} - 2\,q_{n,1})\,y_t \\ &+ \quad q_{n,n}\,y_{t-n} \\ &+ \quad (q_{n,n-1} - 2\,q_{n,n})\,y_{t-n+1} \end{aligned}$$

Le même traitement peut être effectué pour les coefficients linéaire et constant. En reprenant les coefficients des termes en $y_t$ en une seule matrice :

$$\mathcal{K} = \begin{bmatrix} q_{n,1} & q_{n,2} - 2\,q_{n,1} - K_q & q_{n,n} & q_{n,n-1} - 2\,q_{n,n} - K_q \\ l_{n,1} & l_{n,2} - 2\,l_{n,1} - K_l & l_{n,n} & l_{n,n-1} - 2\,l_{n,n} - K_l \\ c_{n,1} & c_{n,2} - 2\,c_{n,1} - K_c & c_{n,n} & c_{n,n-1} - 2\,c_{n,n} - K_c \end{bmatrix}$$

nous obtenons :

$$\begin{bmatrix} Q_{n,t+1} \\ L_{n,t+1} \\ C_{n,t+1} \end{bmatrix} = 2 \begin{bmatrix} Q_{n,t} \\ L_{n,t} \\ C_{n,t} \end{bmatrix} - \begin{bmatrix} Q_{n,t-1} \\ L_{n,t-1} \\ C_{n,t-1} \end{bmatrix} + \begin{bmatrix} K_q \\ K_l \\ K_c \end{bmatrix} \mathbb{S}_{n,t} + \mathcal{K} \begin{bmatrix} y_{t+1} \\ y_t \\ y_{t-n+1} \\ y_{t-n} \end{bmatrix}$$

avec les coefficients :

$$K_q = q_{n,j+1} - 2\,q_{n,j} + q_{n,j-1} = \frac{360}{n\,(n^2-1)\,(n^2-4)}$$

$$K_l = l_{n,j} - 2\,l_{n,j-1} + l_{n,j-2} = -\frac{360}{n\,(n-1)(n^2-4)}$$

$$K_c = c_{n,j} - 2\,c_{n,j-1} + c_{n,j-2} = \frac{60}{n\,(n-1)(n-2)}$$

et

$$\mathcal{K} = \frac{3}{n}\begin{bmatrix} \frac{10}{(n+1)(n+2)} & \frac{-10}{(n-1)(n-2)} & \frac{10}{(n+1)(n+2)} & \frac{-10}{(n-1)(n-2)} \\ -2\frac{6n+7}{(n+1)(n+2)} & 6\frac{2n+1}{(n-1)(n-2)} & -2\frac{4n+3}{(n+1)(n+2)} & 2\frac{4n+7}{(n-1)(n-2)} \\ 3 & -\frac{74-33n+n^2}{(n-1)(n-2)} & -\frac{38-11n+n^2}{(n-1)(n-2)} & 1 \end{bmatrix}$$

Les coefficients $K_c$, $K_l$, $K_q$ et $\mathcal{K}$ sont fonction de $n$, et peuvent être précalculés hors-ligne.

## C.2 Approximation polynomiale

Pour une approximation polynomiale de degré $N$, on a :

$$\mathcal{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 \\ 2^N & 2^{N-1} & \cdots & 2^2 & 2 & 1 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ n^N & n^{N-1} & \cdots & n^2 & n & 1 \end{bmatrix}$$

Et donc :

$$\mathcal{M} = \mathcal{X}'\mathcal{X} = \begin{bmatrix} \sum_{k=1}^{n} k^{2N} & \sum_{k=1}^{n} k^{2N-1} & \cdots & \sum_{k=1}^{n} k^{N} \\ \sum_{k=1}^{n} k^{2N-1} & \sum_{k=1}^{n} k^{2N-2} & \cdots & \sum_{k=1}^{n} k^{N-1} \\ \vdots & \vdots & & \vdots \\ \sum_{k=1}^{n} k^{N+1} & \sum_{k=1}^{n} k^{N} & \cdots & \sum_{k=1}^{n} k \\ \sum_{k=1}^{n} k^{N} & \sum_{k=1}^{n} k^{N-1} & \cdots & n \end{bmatrix}$$

Cette matrice $(N+1) \times (N+1)$ est de rang plein donc inversible. La démonstration (donnée à la fin de ce chapitre) se fait par récurrence sur $N$, sachant que la méthode classique pour calculer $\sum_{k=1}^{n} k^N$ fait intervenir des termes en $n^{N+1}$ et une combinaison linéaire de termes de degré plus bas.

L'inversion peut être réalisée comme ci-dessus de façon symbolique, mais cela devient encombrant pour un $N$ arbitraire. L'inversion numérique par contre peut être établie hors ligne, pour une utilisation incrémentale, qui fera intervenir $N$ mesures passées.

En notant $\mathcal{M}^{-1}_{i,j}$ les éléments de $\mathcal{M}^{-1}$, on aura un filtre linéaire de coefficients :

$$\mathcal{X} = \mathcal{M}^{-1}\,\mathcal{X}' = [M_1\ M_2\ \ldots\ M_n]$$

$$\mathcal{X} = \begin{bmatrix} \sum_{k=1}^{N+1} \mathcal{M}_{1,k}^{-1} 1^{N-k+1} & \dots & \sum_{k=1}^{N+1} \mathcal{M}_{1,k}^{-1} j^{N-k+1} & \dots & \sum_{k=1}^{N+1} \mathcal{M}_{1,k}^{-1} n^{N-k+1} \\ \sum_{k=1}^{N+1} \mathcal{M}_{2,k}^{-1} 1^{N-k+1} & \dots & \sum_{k=1}^{N+1} \mathcal{M}_{2,k}^{-1} j^{N-k+1} & \dots & \sum_{k=1}^{N+1} \mathcal{M}_{2,k}^{-1} n^{N-k+1} \\ \vdots & & \vdots & & \vdots \\ \sum_{k=1}^{N+1} \mathcal{M}_{N+1,k}^{-1} 1^{N-k+1} & \dots & \sum_{k=1}^{N+1} \mathcal{M}_{N+1,k}^{-1} j^{N-k+1} & \dots & \sum_{k=1}^{N+1} \mathcal{M}_{N+1,k}^{-1} n^{N-k+1} \end{bmatrix}$$

Chaque colonne $M_j$ de cette matrice $(N+1) \times n$ est un vecteur de polynômes en $j$, de degré $N$, aux coefficients dépendant de $n$. Notons $\mathcal{X}_{i,j}$ les éléments de la matrice $\mathcal{X}$ ($M_j = [\mathcal{X}_{1,j}\ \mathcal{X}_{2,j} \dots \mathcal{X}_{N+1,j}]'$).

Lorsqu'on applique le filtre $\mathcal{X}$ sur $[y_{t-n+1} \dots y_{t-1}\ y_t]'$, on obtient les coefficients du polynôme de meilleure approximation :

$$\begin{bmatrix} m^{(1)} \\ m^{(2)} \\ \vdots \\ m^{(N+1)} \end{bmatrix}_{n,t} = [M_1\ M_2 \dots M_n] \begin{bmatrix} y_{t-n+1} \\ \vdots \\ y_{t-1} \\ y_t \end{bmatrix}$$

$$\begin{bmatrix} m^{(1)} \\ m^{(2)} \\ \vdots \\ m^{(N+1)} \end{bmatrix}_{n,t} = \begin{bmatrix} \mathcal{X}_{1,1} & \dots & \mathcal{X}_{1,j} & \dots & \mathcal{X}_{1,n} \\ \mathcal{X}_{2,1} & \dots & \mathcal{X}_{2,j} & \dots & \mathcal{X}_{2,n} \\ \vdots & & \vdots & & \vdots \\ \mathcal{X}_{N+1,1} & \dots & \mathcal{X}_{N+1,j} & \dots & \mathcal{X}_{N+1,n} \end{bmatrix} \begin{bmatrix} y_{t-n+1} \\ \vdots \\ y_{t-1} \\ y_t \end{bmatrix}$$

Donc les coefficients du polynôme d'approximation peuvent être calculés par :

$$m_{n,t}^{(k)} = \sum_{i=1}^{n} \mathcal{X}_{i,k}\, y_{t-n+i}$$

où les $\mathcal{X}_{i,k}$ sont fonction de $n$ et peuvent être précalculés. Le calcul prend un temps linéaire par rapport à $n$.

## Implémentation incrémentale

Le calcul de façon incrémentale de ce filtrage repose sur le fait que les coefficients de filtrage sont polynomiaux de degré $N$. Lorsque le filtrage sera effectué, bon nombre de termes pourra être traité par différenciation comme constants. Passons au détail :

Soit $\Delta_i^N$ l'opérateur de différenciation formelle sur l'indice $i$ défini par :

$$\begin{cases} \Delta_i^1(x) = x_i - x_{i-1} \\ \Delta_i^{n+1} = \Delta_i^n - \Delta_{i-1}^n \end{cases}$$

Pour un ensemble de variables $x_i$ indicées par $i$.

On montrera (à la fin de ce chapitre) :

$$\Delta_i^N(x) = \sum_{k=0}^{N} (-1)^k \binom{N}{k} x_{i-k}$$

On pourra dériver une expression incrémentale sachant que, pour chaque jeu de coefficients, les différences $N$-èmes par rapport au temps sont des constantes. Les différences $N+1$-èmes sont nulles, ce qu'on peut aussi exploiter, avec des résultats comparables à ceux obtenus ici, en terme de temps de calcul.

On a, puisque $M_j$ est un polynôme de degré $N$ en $j$ avec des coefficients qui sont des fractions rationelles en $n$ :

$$\forall n \quad \exists K \quad \forall j \in [N+1, n] \qquad \Delta_j^N(M) = \sum_{k=0}^{N} (-1)^k \binom{N}{k} M_{j-k} = K$$

$K$ est un vecteur de fractions rationnelles en $n$. En notant $K_i$ les éléments de $K$, et pour $n > N$, le calcul de $\Delta_j^N(M)$ permet de trouver l'expression générale suivante :

$$m_{n,t+N}^{(i)} = \sum_{p=0}^{N-1} (-1)^p \binom{N}{p} m_{n,t+p}^{(i)} + \sum_{q=1}^{N} K_{i,q}^{-} y_{t-n+q} + K_i \sum_{q=N+1}^{n} y_{t-n+q} + \sum_{q=1}^{N} K_{i,q}^{+} y_{t+q}$$

Ce qu'on peut interpréter comme un ajustement des valeurs passées de $m^{(i)}$ par trois composantes. La première contient les termes des "vieilles" valeurs de $y_t$, que l'on élimine du calcul. La deuxième contient les termes constants issus de la différenciation. C'est par la présence de ces termes que l'on peut utiliser une approche efficace pour des $n \geq N$ : les constantes $K_i$ sont précalculées, puis on se sert uniquement d'une somme partielle, qu'il est facile d'implémenter de façon incrémentale. La troisième composante prend en charge les valeurs "récentes" de $y_t$. Ces trois composantes dépendent des coefficients suivants, qui sont fonction de $N$ et $n$ :

$$K_{i,q}^{-} = \sum_{r=1}^{q} (-1)^r \binom{N}{r} \mathcal{X}_{i,q-r+1}$$

$$K_{i,q}^{+} = \sum_{r=q}^{N} (-1)^r \binom{N}{r} \mathcal{X}_{i,q-r+n}$$

$$K_i = \sum_{r=0}^{N} (-1)^r \binom{N}{r} \mathcal{X}_{i,r}$$

Nous avons ainsi une expression permettant un calcul incrémental en temps constant (par rapport au nombre de points sur lesquels on opère) d'une approximation polynomiale à une série chronologique, lorsque celle-ci est issue d'un échantillonnage à pas constant. Ce calcul est l'équivalent d'un calcul de régression linéaire sur des régresseurs monomiaux.

## C.3 Résultats intermédiaires

### Expression directe de la différence $N$-ème

Donnons la démonstration du résultat :

$$\Delta_i^N(x) = \sum_{k=0}^{N} (-1)^k \binom{N}{k} x_{i-k}$$

Par récurrence sur $N$ :

Lorsque $N = 2$ :

$$\begin{aligned} \Delta_i^2(x) &= \Delta_i^1(x) - \Delta_{i-1}^1(x) \\ &= (x_i - x_{i-1}) - (x_{i-1} - x_{i-2}) \\ &= x_i - 2\,x_{i-1} + x_{i-2} \end{aligned}$$

$$\Delta_i^2(x) = \binom{2}{0} x_i - \binom{2}{1} x_{i-1} + \binom{2}{2} x_{i-2}$$

Et, si

$$\Delta_i^N(x) = \sum_{k=0}^{N} (-1)^k \binom{N}{k} x_{i-k}$$

alors

$$
\begin{aligned}
\Delta_i^{N+1}(x) &= \Delta_i^N(x) - \Delta_{i-1}^N(x) \\
&= \sum_{k=0}^{N}(-1)^k\binom{N}{k}x_{i-k} - \sum_{k=0}^{N}(-1)^k\binom{N}{k}x_{i-1-k} \\
&= \sum_{k=0}^{N}(-1)^k\binom{N}{k}x_{i-k} - \sum_{k=1}^{N+1}(-1)^{k-1}\binom{N}{k-1}x_{i-k} \\
&= \binom{N}{0}x_i + \sum_{k=1}^{N}\left[(-1)^k\binom{N}{k} - (-1)^{k-1}\binom{N}{k-1}\right]x_{i-k} - (-1)^N\binom{N}{N}x_{i-(N+1)} \\
&= x_i + \sum_{k=1}^{N}(-1)^k\left[\binom{N}{k} + \binom{N}{k-1}\right]x_{i-k} + (-1)^{N+1}x_{i-(N+1)}
\end{aligned}
$$

$$
\Delta_i^{N+1}(x) = \sum_{k=0}^{N+1}(-1)^k\binom{N+1}{k}x_{i-k}
$$

D'où le résultat, par récurrence.

## Les matrices $\mathcal{M}$ sont inversibles

Pour démontrer ce résultat, nous procédons en deux temps.

1. Nous montrons d'abord que les sommes $\sum_{k=1}^{n} k^N$ sont des polynômes en $n$, de degré $N+1$.

   Formellement, cela s'écrit $\sum_{k=1}^{n} k^N \in \mathbb{R}[N+1]$.

2. Ensuite nous montrerons que les matrices

$$
\mathcal{A}_N = \begin{bmatrix} A_{2N} & \dots & A_{N+1} & A_N \\ \vdots & & \vdots & \vdots \\ A_{N+1} & \dots & A_2 & A_1 \\ A_N & \dots & A_1 & A_0 \end{bmatrix}
$$

   où les $A_j \in \mathbb{R}[j]$ sont non-nuls et ne présentent pas de racines positives, sont inversibles pour $n > 0$.

**Les éléments de $\mathcal{M}$ sont des polynômes**

Notons $S_m(n) = \sum_{k=1}^{n} k^m$. On cherche à montrer

$$\forall m \in \mathbb{N} \quad S_m(n) \in \mathbb{R}[m+1]$$

Pour $N = 1$ :

$$S_1(n) = \sum_{k=1}^{n} k^1 = \frac{1}{2}(n+1)n \qquad \in \mathbb{R}[2]$$

Pour $N = 2$ :

$$S_2(n) = \sum_{k=1}^{n} k^2 = \frac{1}{6}(2n+1)(n+1)n \qquad \in \mathbb{R}[3]$$

Supposons $\forall m \leq N-2 \quad S_m(n) \in \mathbb{R}[m+1]$, alors :

$$\begin{array}{rcccccl}
(n+1)^N & = & \sum_{k=0}^{N} \binom{N}{k} n^k & = & n^N & + & \binom{N}{1} n^{N-1} \quad +\cdots+1 \\
n^N & = & \sum_{k=0}^{N} \binom{N}{k} (n-1)^k & = & (n-1)^N & + & \binom{N}{1} (n-1)^{N-1} \quad +\cdots+1 \\
(n-1)^N & = & \sum_{k=0}^{N} \binom{N}{k} (n-2)^k & = & (n-2)^N & + & \binom{N}{1} (n-2)^{N-1} \quad +\cdots+1 \\
 & \vdots & & & & & \\
3^N & = & \sum_{k=0}^{N} \binom{N}{k} 2^k & = & 2^N & + & \binom{N}{1} 2^{N-1} \quad +\cdots+1 \\
2^N & = & \sum_{k=0}^{N} \binom{N}{k} 1^k & = & 1^N & + & \binom{N}{1} 1^{N-1} \quad +\cdots+1
\end{array}$$

La somme de ces équations produit :

$$(n+1)^N = \binom{N}{1} S_{N-1}(n) + \binom{N}{2} S_{N-2}(n) + \cdots + \binom{N}{k} S_{N-k}(n) + \cdots + n + 1$$

d'où :

$$S_{N-1}(n) = \frac{1}{N}\left((n+1)^N - \sum_{k=2}^{N} \binom{N}{k} S_{N-k}(n) - 1\right)$$

or, par hypothèse de récurrence, tous les $S_{N-k}(n)$ sont de degré inférieur ou égal à $N-1$ pour $k \in [2, N]$. Donc, le terme de plus haut degré dans l'expression de $S_{N-1}(n)$ est $\frac{n^N}{N}$. Donc $S_{N-1}(n) \in \mathbb{R}[N]$. D'où le résultat, par récurrence.

### Les matrices $\mathcal{A}_N$ sont inversibles

Montrons ensuite que les matrices $\mathcal{A}_N$ sont inversibles.

Pour $N = 0$ : la fraction rationnelle $A_0^*(n) = 1/A_0(n)$ n'a pas de pôles strictement positifs, donc $\forall n > 0 \quad A_0^*(n)\, A_0(n) = 1$.

Supposons le résultat vrai pour $\mathcal{A}_N$. Pour montrer que $\mathcal{A}_{N+1}$ est inversible pour $n > 0$, il suffit de montrer que ses colonnes sont indépendantes.

Soit $B_i = [A_{N+1+i}\ A_{N+i} \dots A_{i+1}\ A_i]'$ les colonnes de $\mathcal{A}_{N+1}$ :

$$\mathcal{A}_{N+1} = \begin{bmatrix} A_{2N+2} & A_{2N+1} & \dots & A_{N+2} & A_{N+1} \\ A_{2N+1} & A_{2N} & \dots & A_{N+1} & A_N \\ \vdots & & & \vdots & \vdots \\ A_{N+2} & A_{N+1} & \dots & A_2 & A_1 \\ A_{N+1} & A_N & \dots & A_1 & A_0 \end{bmatrix} = [B_{N+1}\ B_N \dots B_1\ B_0]$$

Soit $\alpha_0 \dots \alpha_{N+1} \in \mathbb{R}$ tels que

$$\sum_{i=0}^{N+1} \alpha_i\, B_i = 0$$

Cela s'écrit aussi :

$$\forall j \in [0, N+1] \quad \sum_{i=0}^{N+1} \alpha_i A_{N+1+i-j} = 0$$

Or, par hypothèse $A_m \in \mathbb{R}[m]$ ; donc le seul terme de degré $2N+2$ provient de $A_{2N+2}$. Puisque $n > 0$ et que les racines des $A_m$ ne sont pas positives, il en résulte $\alpha_{N+1} = 0$. D'où :

$$\forall j \in [0, N+1] \quad \sum_{i=0}^{N} \alpha_i A_{N+1+i-j} = 0$$

Si on laisse de côté le cas $j = 0$, on retrouve :

$$\forall j \in [0, N] \quad \sum_{i=0}^{N} \alpha_i A_{N+i-j} = 0$$

Ce qui est l'écriture d'une relation de dépendance linéaire entre les colonnes de $\mathcal{A}_N$. Or, par hypothèse de récurrence, cette matrice est inversible. Donc

$$\forall i \in [0, N] \quad \alpha_i = 0$$

D'où l'on conclut que tous les $\alpha_i$ sont nuls. Par conséquent, les colonnes de $\mathcal{A}_{N+1}$ sont indépendantes, donc la matrice est inversible.

Par récurrence, toutes les $\mathcal{A}_N$ sont inversibles, lorsque les éléments n'ont pas de racines strictement positives.

Les quantités $S_m(n)$ sont strictement croissantes en $n$, par construction. Or $\forall m \quad S_m(n)|_{n=0} = 0$, donc les $S_m(n)$ sont des polynômes de degré $m+1$ en $n$, sans racines strictement positives. Donc les matrices $\mathcal{M}$ sont inversibles, en vertu des résultats précédents.

# D

# Implémentation du filtre de régression pour tenir compte des données manquantes

Il est possible de précalculer des coefficients de filtrage pour des matrices $X_m^* = (X_m' X_m)^{-1} X_m'$ construites en tenant compte des configurations possibles de données manquantes. Soit

$$\mathcal{B}_n = \{0,1\}^n$$

On notera $b_i$ le $i$-ème élément de $b \in \mathcal{B}_n$ : $b = b_1\ b_2\ \ldots\ b_n$.

Soit

$$\mathcal{M}_n = \left\{ \{n | b_n = 1\} \quad | \ b \in \mathcal{B}_n \right\}$$

l'ensemble des combinaisons que l'on peut construire avec des nombres compris entre 1 et $n$, plus l'ensemble vide.

Alors pour tout $m \in \mathcal{M}_n \setminus \{\emptyset\}$, on note $L(m)$ le vecteur construit dans $\mathbb{R}^{|m|}$ avec les éléments de $m$ triés dans l'ordre croissant.

$$\forall m \in \mathcal{M}_n \quad X_m = [L(m) \quad \mathbf{1}]$$

Et on peut donc précalculer $X_m^*$ pour tout $m$ non vide.

Ensuite, il suffit d'empiler dans le tampon mémoire les données, en ignorant les données manquantes. Les coefficients de filtrage, éléments de $X_m^*$, seront ceux correspondant au $m$ issu de l'élément $b$ de $\mathcal{B}_n$ tel que $b_i = 1$ pour les éléments présents et $b_i = 0$ pour les éléments manquants. Dans le cas $b = 0^n$, on ne fait rien, et on retourne un indicateur de valeur manquante.

Il semble délicat d'implémenter la procédure complète, vu que

$$|\mathcal{B}_n| = |\mathcal{M}_n| = 2^n$$

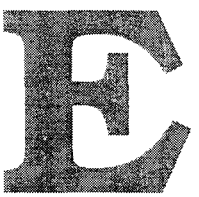
On se contentera donc d'une arithmétique intégrant NaN qu'on interprétera comme indicateur de valeur manquante — ce qui n'est pas son rôle, en principe.

Pour les approches incrémentales, il ne suffit pas d'utiliser NaN pour les données manquant en entrée puis exploiter la propagation des NaNs au reste du calcul. La mise à jour d'une valeur NaN ne peut pas se faire de façon algébrique : les résultats auront indéfiniment la valeur NaN.

Il faut marquer les valeurs manquantes par NaN, pouvoir émettre en sortie du traitement des NaN pour tout calcul faisant intervenir des valeurs manquantes. De plus, lorsque des données manquantes n'apparaissent *plus* dans la zone de calcul considérée, il faut réinitialiser la procédure incrémentale, en utilisant la méthode par filtrage, par exemple.

*Nous nous étions proposé de situer le débat sensiblement plus haut et, pour tout dire, au coeur même de cette hésitation qui s'empare de l'esprit lorsqu'il cherche à définir le mot "hasard".*

André Breton, *L'Amour Fou*



# Un test de signification adapté à l'extraction de tendance

La littérature propose [Sap90, p.367] un test de signification pour la régression. Il consiste à vérifier, de manière statistique, que les résultats de régression obtenus ont peu de chances de provenir de distributions indépendantes. Ce test utilise une distribution de Student ou, de façon tout-à-fait équivalente, une distribution de Fisher.

Nous proposons ici un test alternatif, pour une hypothèse proche de l'hypothèse classique, en utilisant des méthodes de statistique numérique. Les résultats sont comparables avec ceux issus du test classique.

De même, nous comparons les résultats obtenus par le test classique avec ceux obtenus par application de tests non paramétriques. Ceux-ci ne nécessitent d'aucune hypothèse sur les distributions mises en jeu. Les résultats restent comparables à ceux issus du test classique, mais nécessitent de temps de calcul en ligne nettement plus importants.

Enfin, nous allons illustrer et discuter l'influence du risque employé dans la détremination des indicateurs $\mathcal{N}(n)$ et $\mathcal{Z}(n)$.

## E.1 Un test alternatif

Nous poposons ici un test alternatif au test classique, reposant sur des méthodes purement numériques.

### E.1.1 Réécriture de l'hypothèse nulle

Le test classique de signification pour la régression a été établi vis-à-vis d'une hypothèse nulle d'indépendance de deux variables aléatoires pour lesquelles l'espérance de la variable régressée sachant le régresseur suit une distribution indépendante et gaussienne. Dans cette hypothèse, la statistique $T$ définie par :

$$T = \frac{\mathcal{R}_{n,t}}{\sqrt{1 - \mathcal{R}_{n,t}^2}} \sqrt{2\,n - 1}$$

suit une distribution de Student à $n - 2$ degrés de liberté ; $\mathcal{R}_{n,t}$ est le coefficient de corrélation (de Pearson ou de Bravais-Pearson), défini en §2.5.

Dans notre application à l'estimation d'une dérivée, la variable temps, le régresseur, n'est pas une variable aléatoire. En toute rigueur, nous ne sommes pas en train d'établir une régression mais un *modèle linéaire.*

L'hypothèse nulle pour un test de signification peut alors être réécrite : $T$ suit la distribution d'une variable aléatoire $T'$, calculée à partir de la régression d'une variable aléatoire distribuée de façon gaussienne, par rapport au temps.

La statistique $T$ a été estimée dans les points intéressants par simulation numérique [Mac96].

Pour la génération des valeurs de la variable gaussienne, nous avons utilisé le générateur de nombres pseudo-aléatoires Mersenne Twister [MN98]. Il s'agit d'un générateur de très longue période ($2^{19937} - 1$), avec un ordre élevé d'équidistribution (623), qui utilise néanmoins peu de ressources en mémoire et en temps de calcul. De plus, il est disponible, code source compris à `http://math.keio.ac.jp/~matumoto/emt.html` ; le code est utilisable et modifiable suivant une licence très libérale.

S'agissant d'un générateur *binaire* à la base, l'ordre d'équidistribution permet d'estimer sa qualité en tant que générateur de nombres réels. Il s'agit d'une mesure de la qualité de couverture du générateur.

L'ordre d'équidistribution est calculé de la façon suivante. Soit $D$ un entier. Le flux binaire produit par le générateur est divisé (arithmétiquement) par $2^{32}$, afin d'obtenir un nombre réel à 32 bits de précision. Puis pour une période complète du générateur, les nombres correspondants vont être placés cycliquement dans les dimensions de 1 à $D$. La dimension $D$ maximale pour laquelle les nombres ainsi placés sont équidistribués — sauf en zéro, en toute rigueur — constitue l'ordre d'équidistribution. Il s'agit d'une mesure de la régularité de couverture des réels pour un générateur binaire — et dépend par conséquent du nombre de bits de précision qu'on désire, ici 32.

### E.1.2 Estimation de la distribution de référence

La distribution $T'$ est calculée par simulation. Pour pouvoir estimer la qualité de la simulation, nous avons exploré la variance des points qui nous intéressent dans la distribution de $T'$.

Pour $N$ jeux de tirage de $L$ nombres aléatoires distribués suivant une loi normale[1], nous avons calculé les quantiles de $T'$ correspondant aux risques $0.1 \cdot 2^{-k}$ $k \in [0 \ldots 7]$. La variance de cette quantité — qui suit presque toujours une loi normale de façon significative — est stable après $N > 400$ et suit la loi empirique suivante :

$$\sigma_{T'} = c_\alpha \, \alpha + c_{L^{-1}} \, L^{-1} + c_{\tau^{-1}} \, \tau^{-1}$$

$$\begin{aligned} c_\alpha &= -0.7238 \pm 0.0055 \\ c_{L^{-1}} &= 327.8031 \pm 7.5813 \\ c_{\tau^{-1}} &= -0.7566 \pm 0.1105 \end{aligned}$$

qui a été établie par régression multiple et rééchantillonnage jacknife [ST95], aussi appelé *leave-one-out* à partir d'un ensemble de 15 séries d'échantillons. Les coefficients $c$ sont donnés en médiane plus ou moins un écart-type.

Ainsi, pour un risque donné (désormais $10^{-2}$) nous avons une bonne approximation (d'écart-type $10^{-5}$ pour des valeurs de l'ordre de l'unité) de $T'$ après quelques 500 échantillons de longueur 37000, pour l'ordre de grandeur de $\tau$ qu'on est capables de gérer : $\tau \leq 540$ (soit $45min$).[2]

En Fig.E.1 nous avons illustré la différence entre $T$ et $T'$, pour les petits $n$ ; pour des $n$ plus importants, les deux distributions approchent la loi normale, en vertu du théorème central-limite [Sap90, pp. 43,62–63].

Les calculs ont été réalisés de façon "exacte" pour un quadrillage de $\tau \in \{10, 12, \ldots, 80\}$ et $\alpha \in \{0.1 \cdot 2^{-k} \mid k \in \{0 \ldots 7\}\}$. Pour les valeurs hors quadrillage, nous avons entraîné un réseau de neurones [Hay94][3] de type perceptron à une couche cachée, dont les coefficients des couches sont donnés en Fig.E.2. Le réseau a été entraîné par l'algorithme de Levenberg-Marquardt [Dav93], pour une précision de $10^{-3}$ (obtenue après 1150 pas de l'algorithme).

### E.1.3 Un test suffisant ?

Nous pouvons donc construire un test mieux adapté à notre problématique. Les différences constatées par rapport aux valeurs critiques du test classique sont probablement à mettre à dos des moyens de normalisation qu'on utilise. En effet, pour le calcul du coefficient de corrélation[4], la pente de la droite calculée a été nomalisée par rapport à l'écart-type de la variable à régresser et par rapport à l'écart-type du régresseur. Celui-ci étant en fait le temps, cette normalisation ne ramène pas toutes les échelles à la même. Il ne s'agit là que d'une piste pour expliquer la différence constatée, et qui n'est vraie que pour des petits $\tau$.

[1] La distribution normale est déduite du tirage uniforme par la méthode de [AD73] ; voir [Knu97, pp. 122–132] pour plus de détails et d'autres approches.

[2] Etant donnée la puissance de calcul à notre disposition et les implémentations actuelles de ces algorithmes.

[3] Voir aussi la Foire Aux Questions sur les réseaux de neurones à `ftp://ftp.sas.com/pub/neural/FAQ.html`

[4] Il s'agit là d'un abus de langage : nous ne traitons pas de variables aléatoires, donc parler de régression linéaire est abusif. Un vocabulaire basé sur les notions de modèle linéaire et d'ajustement de paramètres serait plus adapté.

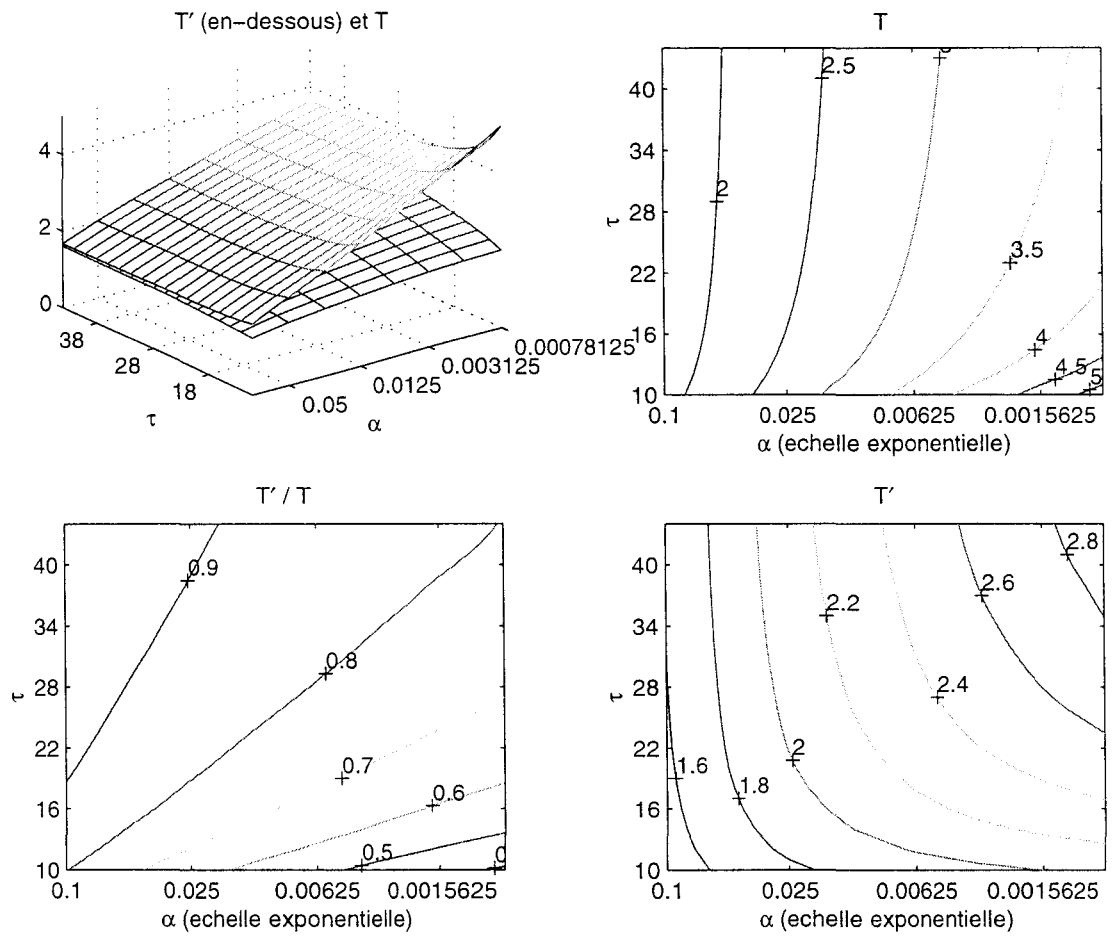



FIG. E.1 – Distribution $T$ de Student et Distribution $T'$

En haut à gauche, une représentation tridimensionnelle des valeurs des deux distributions, en fonction de $\alpha$ et de $\tau$ (qui est, pour la loi $T$ le nombre de degrés de liberté plus 2). La loi $T'$ possède des valeurs inférieures à $T$, et s'en rapproche pour des $\tau$ croissants et des $\alpha$ croissants. En haut à droite, la distribution $T$ en diagramme de contour, en bas à droite la distribution $T'$. En bas à gauche, le diagramme de contour du quotient des valeurs des deux distributions.

La valeur de la fonction de répartition inverse de $T'$, pour $\tau$ degrés de liberté et au risque $\alpha$ vaut :

$$t'(\alpha,\tau) = \sum_{i=1}^{M} \gamma_i \, \phi \left( \sum_{j=1}^{p} w_{ij} \, \xi_i - \theta_i \right) - \delta$$

avec $M = 10$ neurones dans la couche cachée ; $p = 2$ entrées ($\xi_1 \equiv \alpha$ et $\xi_2 \equiv \tau$) et des fonctions d'activation sigmoïdes

$$\phi(x) = \frac{2}{1 + e^{-2x}} - 1$$

Les coefficients valent :

$$W = \begin{bmatrix} 113.94 & -0.0055532 \\ 25.338 & -0.058055 \\ -73.905 & -0.035396 \\ 44.513 & 0.9764 \\ 27.214 & 0.21151 \\ -106.43 & 0.0048209 \\ -12.427 & -0.041191 \\ 8.9914 & 0.1255 \\ -43.565 & 1.5175 \\ -12.877 & 0.00047305 \end{bmatrix} \quad \Theta = \begin{bmatrix} 1.0637 \\ -2.9212 \\ -1.9088 \\ -3.6069 \\ -6.147 \\ -1.6202 \\ 0.70751 \\ -0.50501 \\ 0.23792 \\ 0.11908 \end{bmatrix} \quad \Gamma = \begin{bmatrix} 25.185 \\ -0.12281 \\ -20.076 \\ 11.025 \\ -0.0045603 \\ 79.203 \\ -0.24154 \\ 0.49539 \\ 10.822 \\ 0.80783 \end{bmatrix}$$

et

$$\delta = 0.37691$$

FIG. E.2 - Le réseau d'approximation de la distribution $T'$

On peut néanmoins s'interroger encore sur le bien-fondé de ce nouveau test. En effet, jusqu'où est-il raisonnable de supposer que la distribution de référence (celle qui correspondrait au hasard) est effectivement gaussienne ? Cette hypothèse est mise en défaut *a priori* par la forte quantification des données.

Si nous voulons déterminer un test qui soit vraiment sensible à l'ordre temporel des données (puisque c'est cet aspect-là qui détermine la signification de la régression), il faudrait pouvoir ajuster le test aux distributions *réelles* des données.

Cela complexifie énormément le problème : nous ne pouvons plus tenir compte d'une seule distribution de référence considérée comme représentative du cas ou seul le hasard intervient. Il faudrait établir un jeu de distributions de référence, pour lesquelles nous pouvons calculer les statistiques $T'$ correspondantes, puis choisir au vol la distribution de référence qui est la plus proche de la distribution réelle, ordre temporel en moins. Cela poserait évidemment des problèmes de temps de calcul hors ligne et de stockage en mémoire. Il faut de plus être capable de choisir la distribution de référence au vol, à partir des caractéristiques des données, ce qui est un problème en soi.

Une approche possible est celle de faire le test à partir d'une approche de données brouillées ou *surrogate data* [TGL+92, ST95]. Celle-ci consisterait à tester exclusivement l'ordre temporel des données, en brouillant cet ordre et recalculant les statistiques sur un ensemble de réalisations brouillées, qui ont donc la même distribution mais sont indépendantes dans le temps.

Le problème majeur de cette approche est qu'elle n'est pratiquable que pour des ensembles de données suffisamment importants pour pouvoir générer un nombre important d'ensembles dérivés brouillés. Un deuxième problème est que le nombre de ces ensembles reste à déterminer.

Le premier écueil est assez limitant dans notre cas, puisqu'on cherche à caractériser la signification d'un nombre de mesures variant de quelques-unes à plusieurs centaines — la variation que nous utilisons de la taille de la fenêtre. Quant au deuxième, sa résolution ne semble pas évidente et nécessiterait d'une étude du même genre que celle effectuée pour la détermination du test du §E.1.2.

Du fait de ces limitations, nous allons considérer le test établi précédemment comme suffisant, étant un bon compromis entre la possibilité de mise en ligne — il suffit de calculer les valeurs de référence une fois pour toutes hors-ligne pour chaque $\tau$ et pour les $\alpha$ souhaités ; en-ligne aucune procédure de sélection n'est nécessaire — et l'adaptation aux données — l'hypothèse d'une distribution gaussienne n'étant mise en défaut que par la quantification, et l'ajustement le plus important à priori, celui de la distribution uniforme du temps ayant été réalisé. Nous réserverons des ajustements ultérieurs pour un travail approfondi postérieur.

## E.2 Tests non paramétriques

Afin de nous affranchir de toute hypothèse concernant les distributions des variables, nous allons employer des tests non paramétriques pour la recherche des indicateurs $\mathcal{N}(n)$ et $\mathcal{Z}(n)$.

Nous allons utiliser les tests de Spearman [Sap90, pp.141–143] et de Kendall [Sap90, pp.143–145]. Tous deux sont basés sur des indicateurs de la "monotonicité conjointe" de deux échantillons. Ces indicateurs sont invariants pour toute transformation monotone des variables originelles.

### E.2.1 Test de corrélation des rangs de Spearman

Spearman propose d'étudier le rapport des échantillons de deux variables aléatoires $X$ et $Y$ en terme de *rangs* de leurs valeurs. La relation monotone entre $X$ et $Y$ sera mesurée par le coefficient de corrélation des rangs.

On associe à chaque valeur $x_i$ de $X$ son rang $r_i^x$, et de même pour $Y$. Les $r^x = \{r_i^x \mid i = 1 \dots n\}$ et $r^y = \{r_i^y \mid i = 1 \dots n\}$ sont des permutations de $[1\,2 \dots n]$.

Le coefficient de corrélation entre $r^x$ et $r^y$ vaut alors, après simplification :

$$r_{r^x r^y} = 1 - \frac{6}{n\,(n^2-1)} \sum_{i=1}^{n} (r_i^x - r_i^y)^2$$

Cette statistique sera comparée avec la valeur obtenue en considérant l'équiprobabilité des $n!$ permutations de $[1\,2 \dots n]$. Cette valeur de référence est tabulée et est appelée typiquement "$r$ de Spearman". Pour des $n$ assez grands, la valeur de référence est distribuée comme une loi gaussienne centrée d'écart-type $\frac{1}{\sqrt{n-1}}$.

### E.2.2 Test de corrélation des rangs de Kendall

Kendall propose d'étudier le rapport des classements des rangs pour les deux variables. Pour cela, le coefficient dit "$\tau$ de Kendall" est calculé :

$$\tau = \frac{2\,S}{n(n-1)}$$

avec

$$S = \sum_{i>j} (x_i - x_j)(y_i - y_j)$$

$S$ est la différence entre le nombre de classements concordants entre $X$ et $Y$ le nombre de classements discordants entre $X$ et $Y$. Pour $\tau = 1$, tous les classements sont identiques, pour $\tau = -1$ tous les classements sont inversés. Un moyen pratique pour calculer $\tau$ pour un échantillon consiste à trier $X$ et $Y$ suivant l'ordre des $X$. Cela produit $\tilde{X}$ qui est trié dans l'ordre croissant et $\tilde{Y}$. Alors, on calcule :

$$R = \sum_{i>j} \mathbb{1}(\tilde{y}_i > \tilde{y}_j) = \sum_{i>j} \left| \{i \mid y_i > y_j\} \right|$$

où la fonction $\mathbb{1}(c)$ retourne 1 lorsque $c$ est une proposition vraie et zéro autrement. Alors :

$$S = 2R - \frac{n(n-1)}{2}$$

et

$$\tau = \frac{4R}{n(n-1)} - 1$$

Lorsque $X$ et $Y$ sont indépendantes, $\tau$ est distribuée approximativement comme une loi gaussienne centrée d'écart-type $\sqrt{\frac{2(2n+5)}{9n(n-1)}}$, avec une bonne approximation dès que $n > 8$.

Le principe de calcul de $\tau$ peut être généralisé à plusieurs variables, par l'utilisation du "$W$ de Kendall" [Sap90, pp.146–147].

### E.2.3 Résultats

Nous avons appliqué les deux tests précédents dans les mêmes conditions qu'au chapitre 2 : pour chaque point de mesure d'une donnée et pour chaque échelle. Les résultats sont illustrés en Fig. E.3. Les divers tests fournissent qualitativement les mêmes résultats, à une transformation affine près.

Au passage, notez que les tests non paramétriques sont apparemment moins exigeants que le test paramétrique de Pearson/Student. Ceci est issu d'un artefact de calcul : lorsque les données ne varient pas, le calcul du coefficient de corrélation donne NaN, qui n'entre pas dans le décompte. Les tests non paramétriques s'accomodent du cas des valeurs constantes.

Les temps de calcul, par contre sont prohibitifs. En effet, les tests non paramétriques nécessitent un tri préalable des données. Leur temps de calcul est donc au moins de $O(n \log n)$, le temps de calcul en moyenne du Quicksort [Hoa62]. En ligne cela devient très vite excessif. Le test classique utilisant le coefficient de régression peut être effectué en temps de calcul constant, pour tout $n$. A titre d'anecdote, l'obtention des résultats de la figure E.3 a nécessité plusieurs *jours* de calcul sur un ordinateur équipé de processeur Celeron à 333MHz. Pour le test basé sur la régression, les résultats sont obtenus avec des temps de calcul de l'ordre de la *minute*.

## E.3 Influence du risque

La valeur exacte du risque utilisé pour déterminer les indicateurs $\mathcal{N}(n)$ et $\mathcal{Z}(n)$ n'influe pas qualitativement sur les résultats. Il apparaît que la plus ou moins grande permissivité agit sur un nombre à peu près constant de points. Il s'ensuit pour $\mathcal{N}(n)$ que la diminution du risque — *c-à-d.* l'augmentation de l'exigence — introduit un facteur additif sur $\mathcal{N}(n)$.

La figure E.4 montre la forme de $\mathcal{N}(n)$ pour des riques $\alpha \in \{.5, .1, .05, .01, \ldots, .0001\}$. La forme étant qualitativement conservée, les maxima le sont aussi.

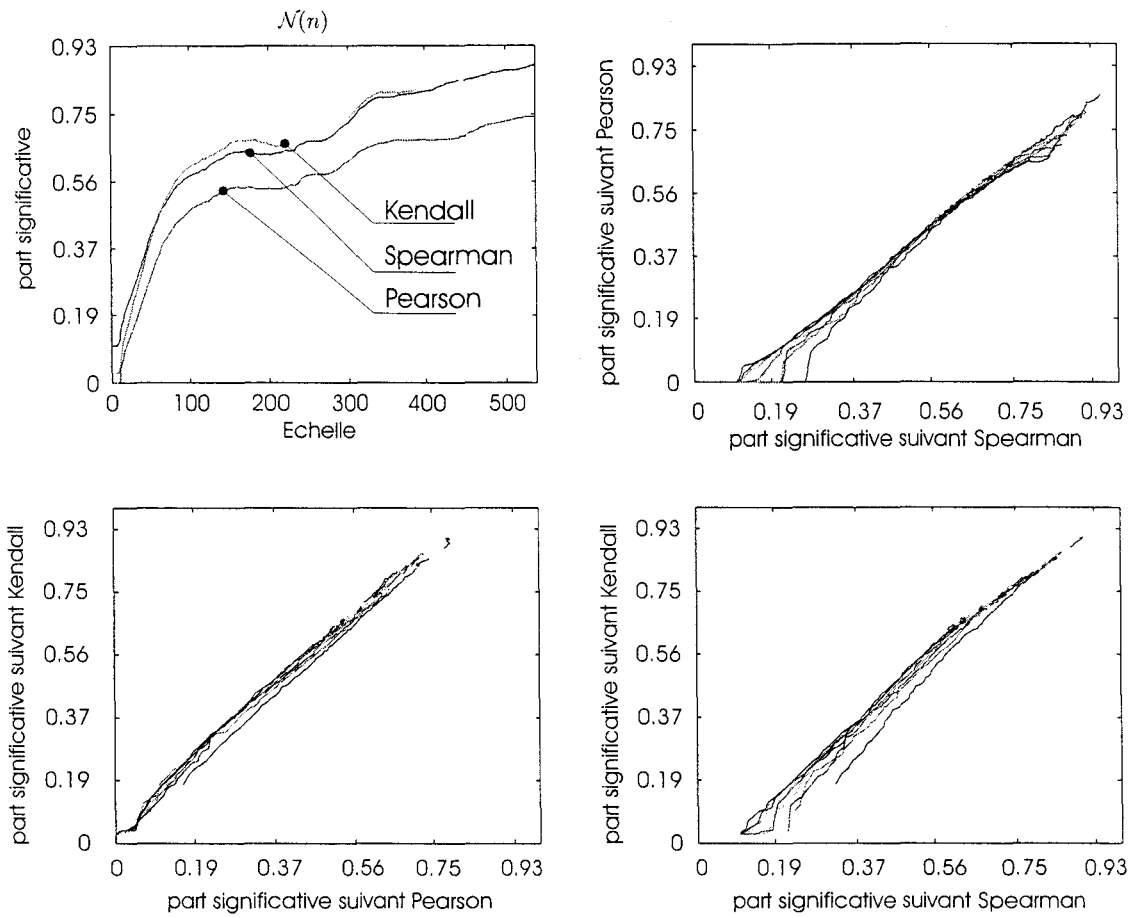



FIG. E.3 – Comparaison de $\mathcal{N}(n)$ pour le test retenu et des tests non paramétriques

En haut à gauche, $\mathcal{N}(n)$ au risque $10^{-2}$ calculé suivant :
- le test utilisant le coefficient de corrélation de Pearson et la distribution de référence de Student ;
- le test non paramétrique basé sur le coefficient de corrélation des rangs de Spearman ;
- le test non paramétrique basé sur le coeffcient de Kendall.

En haut à droite (*resp.* bas-gauche, bas-droite) la relation entre $\mathcal{N}(n)$ suivant les tests basés sur les coefficients de corrélation de Spearman et Pearson (*resp.* Pearson et Kendall, Spearman et Kendall) pour des risques de .5, .1, .05, .01, ..., .0001.

Les comportements sont qualitativement les mêmes entre les trois tests.

Les courbes segmentées pour Kendall proviennent d'interruptions dans le calcul, qui a pris plusieurs *jours*.

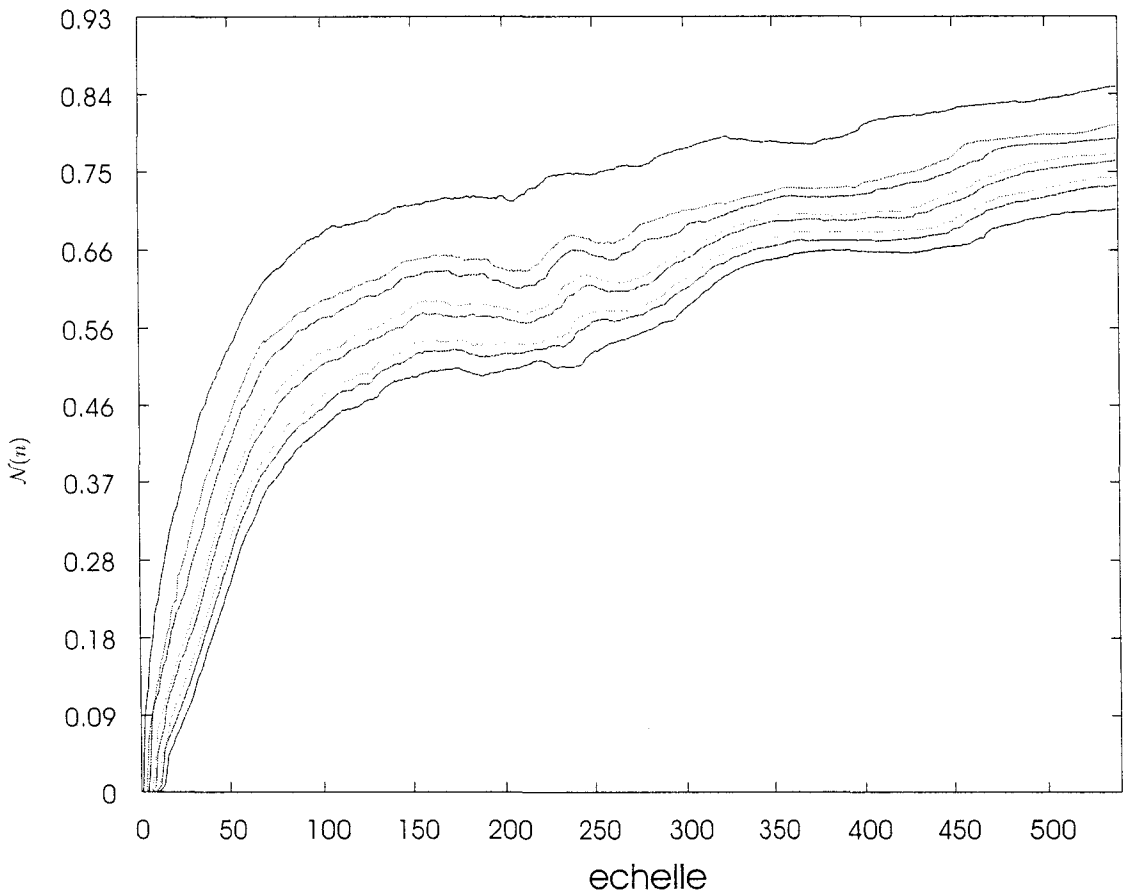



FIG. E.4 – Influence du risque sur $\mathcal{N}(n)$

L'indicateur $\mathcal{N}(n)$ pour des risques $\alpha \in \{.5, .1, .05, .01, \ldots, .0001\}$. Les risques les plus permissifs sont ceux des courbes supérieures. La forme et les maxima sont conservés pour des risques suffisamment petits, dans l'ordre de grandeur des usages courants : inférieurs à $10^{-2}$.

# F

# Liste des Notations

Nous présentons la liste des notations employées au long de ce document, regroupées par thématique. Ce découpage correspond grossièrement au découpage en chapitres.

Tendance et régression Chap.2

| | |
|---|---|
| $\mathcal{S}_{n,t}$ | la valeur de la pente de régression au temps $t$ et pour l'échelle $n$ |
| $\mathcal{P}_{n,t}$ | la valeur de l'ordonnée à l'origine |
| $\mathbb{S}_{n,t}$ | les sommes partielles des $n$ derniers $y_t$ |
| $\sigma^2_{n,t}$ | la variance des $n$ derniers $y_t$ |
| $s_{n,j}$ | le $j$-ème coefficient du filtre linéaire pour la pente à l'échelle $n$ |
| $p_{n,j}$ | *idem.* ordonnée à l'origine |
| $\mathcal{R}_{n,t}$ | le coefficient de régression correspondant à $\mathcal{S}_{n,t}$ et $\sigma^2_{n,t}$ |

Recherche d'une échelle caractéristique Chap.3

| | |
|---|---|
| $\mathcal{K}_{n,t}$ | indicateur booléen "significatif en $t$,$n$" ? |
| $\mathcal{N}(n)$ | nombre de points de régression significative en fonction de l'échelle |
| $\mathcal{Z}(n)$ | nombre de zones significatives en fonction de l'échelle |
| $\tau_{\mathcal{N}}$ | échelle caractéristique définié à partir de $\mathcal{N}(n)$ |
| $\tau_{\mathcal{Z}}$ | *idem.* $\mathcal{Z}(n)$ |

Transformation numérique-symbolique Chap.4

| | |
|---|---|
| $\Sigma$ | la variable qui désigne un symbole |
| $\mathcal{E}_{\mathrm{ss}}$ | donnée artificielle, somme de deux sinus |
| $\mathcal{E}_{\mathrm{cm}}$ | *idem.* constante par morceaux |
| $\mathcal{E}_{\mathrm{sd}}$ | *idem.* sinus décalé |
| $\mathcal{E}_{\mathrm{cp}}$ | *idem.* constante perturbée par des impulsions |

| Construction d'arbres de décision Chap.5 | |
|---|---|
| $\mathcal{T}$ | taille des fenêtres de construction des arbres de décision |
| $O_i$ | le $i$-ème individu parmi $T$ |
| $A_j$ | le $j$-ème attribut parmi $N$ |
| $a_{jk}$ | les modalités de l'attribut $A_j$ |
| $\mathcal{G}$ | stratégie globale de génération |
| $\mathcal{S}$ | critère de sélection d'un test |
| $\mathcal{A}$ | critère de sélection d'attributs |
| $\mathcal{F}$ | critère d'arrêt |
| $\mathcal{Q}$ | critère de qualité |
| $\mathcal{P}$ | seuil de séparation (attributs numériques) |
| $H(x)$ | entropie de Shannon de la variable $x$ |
| $H(y\|x)$ | entropie consitionnelle de $y$ sachant $x$ |
| $H(y,x)$ | entropie conjointe de $y$ et de $x$ |
| $g(y\|x)$ | gain d'information de $y$ sachant $x$ |
| $g_r(y\|x)$ | "gain ratio"de $y$ sachant $x$ |
| $\mathcal{I}(y)$ | indice de couplage de $y$ dans un arbre de décision |

| Autres | |
|---|---|
| $p_a$ | probabilité de l'élément $a$ d'une variable aléatoire |
| $p_{ab}$ | probabilité conjointe de $a$ et $b$ |
| $r_{XY}(\eta)$ | fonction de corrélation croisée au décalage $\eta$ entre les séries chronologiques $X$ et $Y$ |
| NaN | Not a Number |
| $O(x)$ | grand O de Landau |
| $\|E\|$ | cardinal de l'ensemble $E$ |
| $\lfloor x \rfloor$ | partie entière de $x$ |
| $M^*$ | pseudo-inverse de la matrice $M$ |
| $T'$ | distribution alternative à celle de Student pour le test de régression |

# Bibliographie


[AC90] R.K. Avent and J.D. Charlton. A critical review of trend-detection methodologies for biomedical monitoring systems. *Crit. Rev. Biomed. Eng.*, 17 :621–659, 1990.

[AD73] J.H. Ahrens and U. Dieter. Extensions of forsythe's method for random sampling from the normal distribution. *Math. Comput.*, 27 :927 - 937, Oct 1973.

[AHM95] P. Auer, R.C. Holte, and W. Maass. Theory and applications of agnostic pac-learning with small decision trees. In *Proc. 7th Int. Machine Learning Conf.*, 1995.

[Aka97] Metin Akay, editor. *Time-Frequency And Wavelets In Biomedical Signal Processing.* IEEE Press Series on Biomedical Engineering. IEEE Press, 1997.

[BBJ+79] J.E.W. Beneken, J.A. Blom, F.F. Jorritsma, A. Nandorff, and J. Spierdijk. Prognosis, trend and prediction in patient management. *J. Biomed. Eng.*, 1, 1979.

[BC90] Trevor Bench-Capon. *Knowledge Representation : An Approach to Artificial Intelligence.* Academic Press, 1990.

[BFSO84] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees.* CRC Press, 1984.

[BO95] Rohan A. Baxter and Jonathan J. Oliver. Mdl and mml : Similarities and differences. Technical Report 207, Department of Computer Science, Monash University, 1995.

[Bre94] Leo Breiman. Bagging predictors. *Machine Learning*, 24, 1994.

[Bro89] J. Glenn Brookshear. *Theory of Computation.* Benjamin-Cummings, 1989.

[BU95] C. E. Brodley and P. E. Utgoff. Multivariate decision trees. *Machine Learning*, 19 :45–77, 1995.

[Cal96] Daniel Calvelo. Analyse et modélisation des systèmes complexes : Surveillance et aide au diagnostic en réanimation. Mémoire de DEA, L.I.F.L Université des Sciences et Technologies de Lille, Juin 1996.

[CCPR97] D. Calvelo, M.C. Chambrin, D. Pomorski, and P. Ravaux. Arbres de décision et analyse dynamique de données médicales : introduction de la notion d'echelle. In *Colloque de Recherche Doctorale AGIS'97 "Automatique, Genie Informatique, Image, Signal"*, Angers, 1997.

[CCPR99] D. Calvelo, M.C. Chambrin, D. Pomorski, and P. Ravaux. Icu patient state characterisation using machine learning in a time series framework. In W. Horn, Y Shahar, G Lindberg, S Andreassen, and J Wyatt, editors, *Artificial Intelligence in Medicine. Proceedings of the joint european conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99*, number 1620 in Springer's Lecture notes in computer science. Springer, 1999.

[CCPV99] D. Calvelo, M.C. Chambrin, D. Pomorski, and C. Vilhelm. Decision support using machine learning : towards intensive care unit patient state characterization. In *Proc. European Control Conference ECC'99*, Karlsruhe, 1999.

[CCVP99] D. Calvelo, M.C. Chambrin, C. Vilhelm, and D. Pomorski. Apprentissage de modèles dynamiques pour l'aide a la décision en monitorage médical. In *JDA'99 "Journées Doctorales d'Automatique"*, pages 361–364, Nancy, 1999.

[Cib87] Ph. Cibois. *L'Analyse Factorielle.* Coll. Que sais-je? P.U.F., 1987.

[CME+91] C. Carpeggiani, A. Macerata, M. Emdin, C. Michelassi, R. Balocchi, and A. L'Abbate. *Spectral Analysis of Heart Rate Variability Signal, Methodological and Clinical Aspects*, chapter Power Spectral Analysis of Heart Rate Variability. Quaderni di Medicina del Lavoro e Medicina Riabilitativa. La Goliardica Pavese, 1991.

[Coi94] Enrico Coiera. *Monitoring in Anaesthesia and Intensive Care*, chapter Automated Signal Interpretation, pages 32–42. W.B. Saunders Co Ltd, 1994.

[CPM96] Improving control of patient status in critical care : the improve project. *Computer Programs and Methods in Biomedicine*, 51, 1996.

[CRC+89] M.C. Chambrin, P. Ravaux, C. Chopin, J. Mangalaboyi, P. Lestavel, and F. Fourier. Computer-assisted evaluation of respiratory data in ventilated critically ill patients. *International Journal of Clinical Monitoring and Computing*, 6 :211–215, 1989.

[CRC+ss] M.C. Chambrin, P. Ravaux, D. Calvelo, A. Jaborska, C. Chopin, and B. Boniface. Multicentric study of monitoring alarms in adult icu : a descriptive analysis. *Intensive Care Medecine*, In Press.

[Cru92] James P. Crutchfield. *Nonlinear Modeling and Forecasting*, volume XII, chapter Semantics and Thermodynamics, pages 317–360. Addison-Wesley, 1992.

[CS96] P. Cheeseman and J. Stutz. *Advances in Knowledge Discovery and Data Mining*, chapter Bayesian Classification (AutoClass) : Theory and Results. AAAI Press/MIT Press, 1996.

[Cul72] G. Cullmann. *Codage et Transmission de l'Information.* Eyrolles, 1972.

[Dav93] Paul Davis. Levenberg-marquart methods and nonlinear estimation. *SIAM News*, 26(6), October 1993.

[DF96] P. Dean and A. Famili. Comparative performance of rule quality measures in an induction system. *Applied Intelligence Journal*, 1996.

[DPG+97] M. Dojat, F. Pachet, Z Guessoum, D. Touchard, A. Harf, and L. Brochard. Néoganesh : a working system for the automated control of assisted ventilation in icus. *Artificial Intelligence in Medicine*, 11 :97–117, 1997.

[Elo94] T. Elomaa. In defence of c4.5 : Notes on learning one-level decision trees. In William W. Cohen and Haym Hirsh, editors, *Machine Learning : Proceedings of the Eleventh International Conference (ML-94)*, pages 62–69. Morgan Kaufmann, 1994.

[FH95] Nir Friedman and J.Y. Halpern. Plausibility measures : A user's guide. In *Proc. Eleventh Conf. on Uncertainty in Artificial Intelligence (UAI 95)*, pages 175–184, 1995.

[FS96] Yoav Freund and Robert Schapire. Experiments with a new boosting algorithm. In *Machine Learning : proc. of the Thirteenth Intl. Conf.*, 1996.

[Gir93] R. Giraud. *L'Econométrie.* Coll. Que sais-je? P.U.F., 1993.

[Hay94] Simon Haykin. *Neural Networks : A Comprehensive Foundation.* MacMillan, 1994.

[HDW94] G. Holmes, A. Donkin, and I.H. Witten. Weka : A machine learning workbench. In *Proc. Second Australia and New Zealand Conference on Intelligent Information Systems*, 1994.

[HK96] I.J. Haimovitz and I. Kohane. Managing temporal worlds for medical trend diagnosis. *Art.Intel.Med.*, 8(3), 1996.

[HM99] Jim Hunter and Neil McIntosh. Knowledge-based event detection in complex time series. In W. Horn, Y Shahar, G Lindberg, S Andreassen, and J Wyatt, editors, *Artificial Intelligence in Medicine. Proceedings of the joint european conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99*, number 1620 in Springer's Lecture notes in computer science, pages 281–290. Springer, 1999.

[HME+97] W. Horn, S. Miksch, G. Egghart, C. Popow, and F. Paky. Effective data validation of high-frequency data : Time-point-, time-interval-, and trend-based methods. *Computer in Biology and Medicine, Special Issue : Time-Oriented Systems in Medicine*, 27(5) :389–409, 1997.

[Hoa62] C.A.R. Hoare. Quicksort. *Computer Journal*, 5(1), 1962.

[Hol78] Gerald Holton. *The Scientific Imagination.* Cambridge University Press, 1978.

[Hol93] R. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1) :63–91, 1993.

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 1933.

[HS94] M. Holsheimer and A. Siebes. Data mining, the search for knowledge in databases. Technical Report CS-R9406, CWI, 1994.

[Hub95] Barbara Burke Hubbard. *Ondes et Ondelettes. La Saga d'un Outil Mathématique.* Sciences d'Avenir. Pour la Science, 1995.

[IZR+97] A. Ittner, J. Zeidler, R. Rossius, W. Dilger, and M. Schlosser. Feature space partitionning by non-linear and fuzzy decision trees. In *Proceedings of the 7th International Fuzzy Systems Association World Congress(IFSA'97)*, 1997.

[Jan94] J.-S. R. Jang. Structure determination in fuzzy modeling : A fuzzy cart approach. In *Proc. of IEEE international conference on fuzzy systems*, June 1994.

[JRC+97] A. Jaborska, P. Ravaux, M.C. Chambrin, D. Calvelo, and C. Vilhelm. Ambient sound as a simple but useful parameter in icu monitoring. *Medical & Biological Engineering & Computing*, 35 :619, 1997.

[Ker96] E.T. Keravnou. Temporal reasoning for diagnosis in a causal probabilistic knowledge base. *Artificial Intelligence in Medicine*, 8(3) :235–266, 1996.

[KK93] Peter R. Keller and Mary M. Keller. *Visual Clues. Practical Data visualization.* IEEE Press, 1993.

[KLMP94] R. Kohavi, R. Long, D. Manley, and K. Pflegger. Mlc++ : A machine learning library in c++. In *Tools with Artificial Intelligence*, pages 740–743. IEEE Computers Society Pres, 1994.

[KM90] Y. Kodratoff and R. Michalski, editors. *Machine Learning : An Arificial Intelligence Approach*, volume III. Morgan Kaufmann, 1990.

[KMK92] Matthew Koebbe and Gottfried Mayer-Kress. Use of recurrence plots in the analysis of time-series data. In Martin Casdagli and Stephen Eubank, editors, *Nonlinear modeling and forecasting*, pages 361–378. Addison-Wesley, 1992.

[KMSK95] E.M. Koski, A. Mäkivirta, T. Sukuvaara, and A. Kari. Clinicians' opinions on alarm limits and urgency of therapeutic responses. *Int.J. Clin. Monit. Comput.*, 12(2) :85–88, may 1995.

[Knu97] Donald E. Knuth. *The Art of Computer Programming : Seminumerical Algorithms*, volume 2. Addison-Wesley, 3 edition, 1997.

[Kod97] Y. Kodratoff. L'extraction de connaissances à partir des données : un nouveau sujet pour la recherche scientifique. *READ Revue Electronique sur l'Apprentissage par les Donées*, 1(1) :1–28, 1997.

[Kos91] Bart Kosko. *Neural Networks and Fuzzy Systems.* Prentice-Hall, 1991.

[Kos96] A. Koski. Modelling ecg signals with hidden markov models. *Artif. Intel. in Med*, 8(5), october 1996.

[Kuh62] Thomas S. Kuhn. *The Structure of Scientific Revolutions.* University of Chicago Press, 1962.

[LHRG96] J. E. Larsson, B. Hayes-Roth, and D. Gaba. Guardian : Final evaluation. Technical Report KSL-96-25, Knowledge Systems Laboratory, Stanford University, August 1996.

[Mac96] David J.C. MacKay. Introduction to monte carlo methods. In *proceedings International School On Neural Nets "E. Caianiello"*, 1996.

[Mar98] Ch. Marsala. *Apprentissage inductif en présence de données imprécises : Construction et utilisation d'arbres de décision flous.* Thèse de doctorat, Université Paris 6 LIP6, 1998.

[MD96] E.J. Manders and B.M. Dawant. Data acquisition for an intelligent bedside monitoring system. In *Proceedings of the 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 957–958, 1996.

[Mey90] Yves Meyer. *Ondelettes et Opérateurs I.* Hermann, 1990.

[Mey94] Yves Meyer. *Les Ondelettes, algorithmes et applications.* Armand Colin, 2eme edition, 1994.

[MHPP96] S. Miksch, W. Horn, C. Popow, and F. Paky. Utilizing temporal data abstraction for data validation and therapy planning for artificially ventilated newborn infants. *Artificial Intelligence in Medicine*, 8 :543–576, 1996.

[MHPP97] S. Miksch, W. Horn, C. Popow, and F. Paky. *Intelligent Data Analysis in Medicine and Pharmacology*, chapter Time-Oriented Analysis of High-Frequency Data in ICU Monitoring, pages 17–36. Kluwer Academic Press, 1997.

[MKS94] S.K. Murthy, S. Kasif, and S. Salzberg. A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 1994.

[MN98] M. Matsumoto and T. Nishimura. Mersenne twister : A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. on Modeling and Computer Simulation*, 8(1) :3–30, January 1998.

[Mol91] Abraham A. Moles. *Les Sciences de l'Imprécis*. Seuil, 1991.

[MPCP93] F.A. Mora, G. Passariello, G. Carrault, and J-P. Le Pichon. Intelligent patient monitoring and management systems : A review. *IEEE Eng. Med. Bio.*, december 1993.

[MS98] P. Munteanu and J.F. Serignat. Idf : induction d'arbres de décision par l'approximation des fréquences. *Revue Electronique sur l'Apprentissage par les Données*, 2(1) :22–38, 1998.

[MSHP99] Silvia Miksch, Andreas Seyfang, Werner Horn, and Christian Popow. Abstracting steady qualitative descriptions over time from noisy, high-frequency data. In W. Horn, Y Shahar, G Lindberg, S Andreassen, and J Wyatt, editors, *Artificial Intelligence in Medicine. Proceedings of the joint european conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99*, number 1620 in Springer's Lecture notes in computer science, pages 281–290. Springer, 1999.

[Mur95] S.K. Murthy. *On growing better decisision trees from data*. PhD thesis, Johns Hopkins university, 1995.

[Mur98] S.K. Murthy. Automatic construction of decision trees from data : A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4) :345–389, december 1998.

[OJ95] Jonathan J. Oliver and David J.Hand. On pruning and averaging decision trees. In *Proceedings of the Twelth International Conference on Machine Learning*, pages 430–437. Morgan Kaufmann, 1995.

[Pal96] Milan Palus. Nonlinearity in normal human eeg : Cycles, temporal asymmetry, nonstationarity and randomness, not chaos. *Biological Cybernetics*, 75(5) :389–396, 1996.

[Pea00] Karl Pearson. *Philosophical Magazine*, 50 :157–175, 1900.

[PP97] Paul-Benoît Perche and Denis Pomorski. Decision tree induction methods using an entropy criterion - ii- local approaches. In D. W. Pearson, editor, *Proceedings of the Second International ICSC Symposium on Soft Computing SOCO'97*, pages 294–299, Nîmes, 17-19 septembre 1997. ICSC Academic Press.

[PS96] D. Pomorski and M. Staroswiecki. Analysis of dynamical systems based on information theory. In *Proc. World Automation Congress (WAC'96)*, 1996.

[PT94] Seth M. Posner and Edward R. Tufte. Graphical summary of patient status. *The Lancet*, 344 :386–389, 1994.

[Qui86] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1, 1986.

[Qui92] J. R. Quinlan. *C4.5 : Programs for Machine Learning.* Morgan Kaufmann, 1992.

[Qui96] J.R. Quinlan. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 4, 1996.

[RCJ+94] P. Ravaux, M.C. Chambrin, A. Jaborska, C. Vilhelm, and M. Boniface. Aiddiag : Un système d'aide au diagnostic utilisant l'acquisition de la connaissance. *Biometric Bulletin*, 11(3), 1994.

[RR93] B. Le Roux and H. Rouanet. *Analyse de Données Multidimensionnelles.* Dunod, 1993.

[RVBC92] P. Ravaux, C. Vilhelm, M. Boniface, and M.C. Chambrin. A neural approach to knowledge-based systems. In *Proc. IEEE EMBS'92 Conference*, pages 928–930, 1992.

[Sap90] G. Saporta. *Probabilités, Analyse de Données et Statistique.* technip, 1990.

[SC98] Yuval Shahar and Carlo Combi. Timing is everything : Time-oriented clinical information systems. *Western Journal of Medicine*, 168 :105–113, 1998.

[Sch96] D. Schwartz. *Méthodes Statistiques à l'Usage des Médecins et des Biologistes.* Statistique en Biologie et en Médecine. Flammarion, 4eme edition, 1996.

[Sha94] Y. Shahar. A knowledge-based method for temporal abstraction of clinical data. Ph.d. dissertation, Stanford University, 1994.

[SL98] B. Sierra and P. Larrañaga. Predicting the survival in malignant skin melanoma using bayesian networks automatically induced by genetic algorithms. an empirical comparision between different approaches. *Artificial Intelligence in Medicine*, 14(1) :215–230, 1998.

[SS96] W. Sweldens and P. Schröder. Building your own wavelets at home. In *Wavelets in Computer Graphics*, pages 15–87. ACM SIGGRAPH Course notes, 1996.

[SSM96] B. Schölkopf, A. Smola, and K.-R. Mülller. Nonlinear component analysis as a kernel eigenvalue problem. Technical Report 44, Max-Planck-Institut für biologische Kybernetik, Tübingen, 1996.

[SSN+93] Tommi Sukuvaara, Matti Sydänmaa, Hannu Nieminen, Arno Heikelä, and Erkki M.J. Koski. Object-oriented implementation of an architecture for patient monitoring. *IEEE Engineering in Medicine and Biology*, 1993.

[ST95] J. Shao and D. Tu. *The Jacknife and Bootstrap.* Springer's series in Statistics. Springer, 1995.

[Ste96] F. Steimann. The interpretation of time-varying data with diamon-1. *Artif. Intel. in Med.*, 8(4), Aug 1996.

[TC92] Pei-Lei Tu and Jen-Yao Ching. A new decision-tree classification algorithm for machine learning. In *Proceedings of the IEEE International Conference on Tools with AI*, pages 370–377, 1992.

[TG97] Luís Torgo and João Gama. Regression using classification algorithms. *Intelligent Data Analysis*, 1(4), 1997.

[TGL+92] James Theiler, Bryan Galdrikian, André Longtin, Stephen Eubank, and J. Doyne Farmer. Using surrogate data to detect nonlinearity in time series. In Martin Casdagli and Stephen Eubank, editors, *Nonlinear modeling and forecasting*, pages 163–188. Addison-Wesley, 1992.

[TT96] R.G. Turcott and M.C. Teich. Fractal character of the electrocardiogram : Distinguishing heart-failure and normal patients. *Ann. Biomed. Eng.*, 24 :269–293, 1996.

[Tuf97] Edward Tufte. *Visual Explanations : Images and Quantities, Evidence and Narrative.* Graphics Press, 1997.

[Tuk77] John W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, 1977.

[Tur95] P.D. Turney. Cost-sensitive classification : Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 2, 1995.

[Utg95] P.E. Utgoff. Decision tree induction based on efficient tree restructuring. Technical Report 95-18, Department of Computer Science, University of Massachussets, 1995.

[VRC+00] C. Vilhelm, P. Ravaux, D. Calvelo, A. Jaborska, M.-C. Chambrin, and M. Boniface. Think ! : a unified numerical – symbolic knowledge representation scheme and reasonning system. *Artificial Intelligence Journal*, 116 :67–85, 2000.

[WG94] Andreas S. Weigend and Neil A. Gershenfeld, editors. *Time Series Prediction : Forecasting the Future and Understanding the Past*, volume XV of *Santa Fe Studies in the Sciences of Complexity.* Addison-Wesley, 1994.

[Wil97] William J. Williams. *Time-Frequency And Wavelets In Biomedical Signal Processing*, chapter Recent Advances in Time Frequency Representations : Some Theoretical Foundations. IEEE Press Series on Biomedical Engineering. IEEE Press, 1997.

[YI97] Wang Y. and Witten I.H. Induction of model trees for predicting continuous classes. In *Proc. European Conference on Machine Learning*, 1997.

[Zad65] Lotfi Zadeh. Fuzzy sets. *Information and Control*, 8(3) :338–353, 1965.

[Zad96] Lotfi Zadeh. Fuzzy logic and the calculi of fuzzy rules and fuzzy graphs : A precis. *Multiple-Valued Logic*, 1 :1–38, 1996.

[Zha97] Zhengyou Zhang. Parameter estimation techniques : A tutorial with application to conic fitting. *Image and Vision Computing Journal*, 15(1) :59–76, 1997.

[ZS95] J. Zeidler and M. Schlosser. Fuzzy handling of continuous attributes in decision trees. In *Proc. ECML-95 Mlnet Familiarization Workshop "Statistics, Machine Learning and Knowledge Discovery in Databases"*, 1995.

# Ressources techniques

L'ensemble de ce travail a été réalisé avec bon nombre d'outils. Maple a donné un coup de main sur la mise en forme des côtés mathématiques de l'approche. Même si les graphiques sortent tout droit de Matlab et s'ils ont été parfois sophistiqués par Corel Draw, l'essentiel des outils employés sont des logiciels libres. C'est-à-dire des logiciels dont la distribution est libre de droits, sous forme de code source. La philosophie essentielle derrière ce genre de licence est celle du partage de la connaissance et de la participation par l'accès à la connaissance. Il est regrettable que cette tendance, bien que recevant récemment les feux de l'actualité, ne soit pas plus largement adoptée.

Les langages :

**Perl** – `http://www.perl.org` – Le langage où écrire un programme devient immédiatement un exercice de style. L'essentiel des petits programmes de conversion de formats et de gestion des ressources utilisés dans ce travail a été écrit en Perl.

**Python** – `http://www.python.org` – Le langage où l'élégance prime. Les routines de traitement et d'exploration des données issues de la plate-forme Aiddiag ont été écrites en Python.

**C/C++** – `http://www.gnu.org/software/gcc/gcc.html` – Le standard lorsque l'efficacité est essentielle. Les traitements les plus lourds ont été prototypés en Octave puis reprogrammés en C/C++.

**Sather** – `http://www.gnu.org/software/sather` – Un langage malheureusement peu connu, qui est un bonheur à écrire. Il est tout aussi facile de prototyper que d'écrire des versions finales en Sather. Les routines de traitement rapide écrites en C l'ont été aussi en Sather, avec des gains de clarté étonnants.

Les outils :

**Octave** – `http://www.che.wisc.edu/octave` – Un outil de calcul numérique compatible avec Matlab.

**TEX** – `http://www.tug.org` – La merveille de la mise en page.

**emacs** – `http://www.gnu.org/software/emacs` – Le monstre de l'édition de texte.

**xgobi** – `http://www.research.att.com/~andreas/xgobi/` – Un bijou de l'analyse exploratoire de données.

Le système :

**Linux** – `http://www.linux.org` – L'un des UNIX libres qui est une alternative très sérieuse aux systèmes commerciaux.

**Debian** – `http://www.debian.org` – La distribution de Linux la plus proche de la philosophie du libre.

Les références :

**SAL** – `http://sal.kachinatech.com` – Scientific Applications for Linux, le répositoire de référence de systèmes scientifiques.

**freshmeat** – `http://freshmeat.net` – Le coffre aux trésors des logiciels libres.

Tous ces logiciels sont libres : ils donnent accès à leur code source, ils sont librement distribuables. Le site `ftp://ftp.lip6.fr` par exemple donne accès à tous ces outils et bien d'autres. Il est clair que nous vivons dans les temps de l'Internet. Cette recherche a été aussi un exercice dans le travail basé presque exclusivement sur "le Net".