LA RECHERCHE DES HARMONIQUES
UNE NOUVELLE FONCTION DU LOGICIEL CORICO

Michel Lesty

Coryent Conseil, 28, rue Sainte Adélaïde 78000 Versailles France
Couriel : michel.lesty@coryent.com

Résumé: *On présente une nouvelle manière de décomposer une série en ses harmoniques, avec prise en compte des ruptures de tendance éventuelles et des points atypiques. La méthode fondée sur les corrélations partielles et non sur les moyennes mobiles ne requiert pas une cadence régulière d'échantillonnage. Elle est validée sur plusieurs cas d'école, puis appliquée à une étoile.*

Mots clés: *Séries temporelles, Régression multiple, Corrélations partielles, Décomposition Harmonique, Ruptures de tendances, Prévision, Etoiles variables.*

1. Introduction

Un précédent article a montré des applications du logiciel CORICO ainsi que le principe de cette méthode fondée sur les corrélations partielles [LES99]. Nous présentons ici une nouvelle fonction apparue dans la version 3.0, qui permet la décomposition harmonique des séries chronologiques d'une manière à la fois simple et robuste. Elle ne recourt pas à l'analyse de Fourier, limitée aux composantes sinusoïdales, ni au processus ARIMA, *réclamant une cadence régulière d'échantillonnage,* ni aux lissages de courbes par moyennes mobiles qui augmentent les risques de créer des saisonnalités parasites [GOU83].

2. Les phénomènes cycliques

Dans la régression multiple on cherche à expliquer une variable Y à l'aide de *plusieurs autres* variables qui constituent les conditions de l'expérience. Dans la prévision basée sur la décomposition harmonique, il arrive que l'on cherche à expliquer Y par *une seule* variable (le temps par exemple), ou, plus exactement, par une combinaison linéaire de plusieurs fonctions de cette seule variable. Examinons pourquoi ce parti pris de ne point tenir compte du contexte n'est pas forcément déraisonnable, et peut du moins compléter l'analyse multivariée.

Un arbre ne monte pas jusqu'au ciel: il finirait par succomber sous le poids de ses branches ou la pression du vent; et, de ce qu'il était plus petit hier qu'aujourd'hui, ne découle pas forcément qu'il sera plus grand demain. Dans le mouvement quelconque d'un objet à travers l'espace, l'équilibre avec les forces extérieures change à chaque instant. L'environnement de l'arbre est modifié par l'effet de sa croissance. Aussi la connaissance du passé hors du contexte multivarié ne renseigne guère sur le futur.

Dans un mouvement cyclique au contraire, comme la rotation d'une planète sur elle-même, l'équilibre des forces reste stable. Les retours cycliques sont donc plus aisément prévisibles. Par définition, l'onde sinusoïdale du type *cosinus(t)* s'étend implicitement sur tout l'écoulement de la variable t (de moins l'infini à plus l'infini). Elle modélise un *état*.

Or les comportements cycliques, échos, résonances, retours de balanciers, sont extrêmement courants, autant dans la nature physique que dans l'histoire des hommes (par exemple le repos suit le travail, la rigueur suit le laxisme, etc.). On comprend que leur découverte fasse partie des objectifs fondamentaux de l'analyse des données.

3. Réflexions sur des méthodes courantes d'analyse d'une série temporelle

3.1 Méthode de Fourier

L'analyse harmonique, ou décomposition de Fourier est une méthode essentielle dans toutes les branches de la physique. Elle permet d'obtenir de façon précise le spectre en fréquence de la série étudiée. Elle s'applique aux processus "stationnaires de second ordre", par exemple sans tendance (attention: la notion "d'onde stationnaire", en physique, a un sens un peu différent). Faut-il retirer les tendances et autres accidents avant de déterminer les fréquences ? La réponse à cette question dépend des données, et nous verrons qu'il convient le plus souvent d'entremêler les recherches, ce que la méthode de Fourier ne permet pas, car l'ensemble des fréquences est connu en une seule opération.

3.2 Méthodes autoprojectives

Parfois la série, bien que périodique, n'est pas sinusoïdale. On peut cependant calculer la corrélation de la série avec la même série décalée d'une observation, puis de deux observations etc.. Le décalage qui donne la plus forte corrélation permet de connaître la période. Sur ce principe, sont basés des modèles auto-régressifs tels que le modèle ARIMA qui supposent que la série est fonction de ses valeurs passées et d'une perturbation aléatoire. Bien sûr, ces décalages requièrent une cadence régulière d'échantillonnage. Si une structure figée de variation, même compliquée, se reproduit périodiquement, l'autocorrélation est un moyen efficace de prévision. Si au contraire cette structure compliquée est la résultante de plusieurs composantes sinusoïdales de fréquences différentes, elle va se déformer peu à peu, et une prévision basée sur l'autocorrélation sera de plus en plus dégradée.

3.3 Moyennes mobiles

Une façon de déterminer la tendance d'une série est de lui appliquer une transformation qui conserve la tendance et annule les autres composantes. La difficulté réside dans le choix de la transformation. Ce choix n'est pas automatique et doit être fondé sur un examen visuel de la série. Les transformations utilisées le plus souvent sont les moyennes mobiles. La valeur prise par la série à l'instant *t* est remplacée par une somme pondérée des valeurs prises à des dates entourant *t*. C'est une façon de lisser la série. Selon le nombre de dates intervenant dans la transformation, le résultat peut être fort différent. Une cadence régulière d'échantillonnage est recommandée. Une "bonne" moyenne mobile doit satisfaire certains critères: conservation de la tendance, annulation de la saisonnalité, pouvoir de réduction de la variance résiduelle, simplicité des coefficients, élimination de l'effet de phase... [GOU83]. La mise en pratique de la méthode est donc délicate, et pas toujours satisfaisante. A force de transformer les données, on ne sait plus forcément de quoi l'on parle.

4. Le coefficient de corrélation : un instrument de précision

Le coefficient de corrélation qui permet la comparaison des *écarts* autour de la moyenne de deux variables x et y, est l'évaluation mathématique de ce qui *distingue* leurs *variations.*

Soient x' et y' les moyennes de x et y, et soient $X_i = x_i - x'$ et $Y_i = y_i - y'$ les *écarts*, pour i variant de 1 à n. La *variation* de chaque variable peut être représentée par un *vecteur* dans l'espace à n dimensions.

Le coefficient de corrélation est le **cosinus** (dans un espace à n dimensions) de l'angle qui différencie les "vecteurs des écarts" X et Y, dont les coordonnées respectives sont les X_i et les Y_i :

$$r(x,y) = \cos(X,Y) = \frac{X_1Y_1 + X_2Y_2 + ... + X_nY_n}{\sqrt{X_1^2 + X_2^2 + ... + X_n^2}\sqrt{Y_1^2 + Y_2^2 + ... + Y_n^2}}$$

Afin de respecter la discontinuité native des observations, la méthode CORICO repose essentiellement sur le calcul des corrélations totales et partielles. A chaque couple d'observation correspond un instant et un seul. Si les vecteurs sont colinéaires, le cosinus est égal à 1. Si le cosinus est égal à 0, les deux vecteurs sont orthogonaux : les deux séries sont par exemple déphasées de $\pi/2$. La corrélation permet donc de déterminer les déphasages avec une grande précision.

Le coefficient de corrélation est calculé sur l'ensemble des instants; mais, dans la formule ci-dessus, le point X_i de la série x est mis en regard du seul point Y_i de la série y. Jamais deux instants ne sont mélangés comme dans la moyenne, qui n'intervient ici que pour centrer les vecteurs. En tant que cosinus, le coefficient de corrélation possède une précision trigonométrique, fort appréciée en astronomie. D'ailleurs le logiciel CORICO est obligé de travailler avec tous les chiffres significatifs, en double précision, sous peine de résultats dégradés.

Considérons les séries A et B suivantes, de moyennes égales, et de corrélation r(A,B) = 0.6884 :

	Obs1	Obs2	Obs3	Obs4	Obs5	**Moy**
A	*10*	*15*	*17*	*20*	*8*	**14**
B	*5*	*23*	*19*	*15*	*8*	**14**

Changer simultanément l'ordre des observations des deux séries ne change rien aux moyennes ni à la corrélation. Mais changeons seulement l'ordre des observations de la série A (permutons par exemple *17* et *20*) :

	Obs1	Obs2	Obs3	Obs4	Obs5	**Moy**
A	*10*	*15*	*20*	*17*	*8*	**14**

B	*5*	*23*	*19*	*15*	*8*	**14**

Les moyennes ne changent pas, mais la corrélation devient r(A,B) = 0.7694.

Dans la série A, remplaçons *20* par *20.1*, et *17* par *16.9* :

	Obs1	Obs2	Obs3	Obs4	Obs5	**Moy**
A	*10*	*15*	*20,1*	*16,9*	*8*	**14**
B	*5*	*23*	*19*	*15*	*8*	**14**

Ici encore les moyennes ne changent pas, mais la corrélation devient r(A,B) = 0.7697.

La corrélation, en effet, est sensible au moindre changement. Corrélation et moyenne ne dépendent pas de l'ordre des observations (l'ordre des colonnes ou des lignes); mais, contrairement à la moyenne, le coefficient de corrélation dépend de l'ordre relatif des observations de A et de B, et des valeurs précises de ces variables. Supposons que A soit le temps, la corrélation dépend de façon très sensible de ses valeurs, quel que soit l'ordre dans lequel sont rangés les instants d'observation. Cette sensibilité va permettre, une fois retirées les composantes principales, de détecter les autres composantes.

En outre, le coefficient de corrélation ne requiert pas une cadence régulière d'échantillonnage. Pour conserver ces qualités essentielles, on s'interdit ici le recours à l'autocorrélation (bien que le logiciel CORICO permette ce calcul). Plus généralement, on s'interdit toutes les méthodes (tel le calcul différentiel) qui imposent un lien entre des instants successifs, et dépendent de l'ordre dans lequel sont rangés les instants d'observations.

Plusieurs aspects de la méthode ont été décrits ailleurs [LES99]. Voici la démarche utilisée dans la régression multiple : la série à étudier est un vecteur, et, comme il existe une infinité de décompositions possibles de ce vecteur en une somme de vecteurs, nous choisirons les composantes facilitant la prévision. Leur somme est exprimée par un modèle de régression.

L'ensemble des régresseurs peut se composer de facteurs simples, de leurs « interactions » [LES99], de fonctions sinusoïdales des facteurs, de tendances, de variables indicatrices des points aberrants, etc.. Toutes ces variables sont des "vecteurs", et sont strictement traitées de la même façon : on commence par sélectionner celles qui entreront dans le modèle. La première sélectionnée est celle qui corrèle le mieux avec la réponse. La deuxième variable sélectionnée sera la variable la mieux corrélée au résidu de la réponse non expliqué par la première, etc. Les régresseurs sont connus par ordre d'importance décroissante. Il ne reste plus qu'à déterminer les coefficients du modèle selon la méthode classique.

Cette décomposition en cascade ne requiert, à aucun moment, un lissage de la série. On s'épargne ainsi une transformation des données, qui s'accompagne forcément d'une dégradation de la précision, du fait du mélange de plusieurs instants.

Si, pour des besoins de clarté, dans les paragraphes suivants, nous commençons par les composantes sinusoïdales les plus simples (stationnaires), puis nous introduisons progressivement des difficultés : tendances, ruptures, événements atypiques, motifs périodiques, en réalité, l'ordre suivi peut être différent, et dépend beaucoup de l'amplitude relative des différentes composantes. Si l'amplitude saisonnière est plus importante que

l'amplitude de la tendance, c'est la composante saisonnière qui sera trouvée en premier. Si, au contraire, l'amplitude de la tendance est la plus forte, elle sera trouvée d'abord. Mais dans les deux cas, ces composantes seront trouvées. Si deux composantes sinusoïdales ont même amplitude, la précision sur les périodes est d'autant meilleure qu'elle est établie sur une plus longue plage d'échantillonnage. La découverte d'une tendance, ou d'une rupture de tendance, ou d'un pic isolé, peut s'intercaler entre deux composantes sinusoïdales différant par leurs amplitudes, fréquences et phases. Contrairement à la décomposition de Fourier, nous pouvons entremêler, dans notre recherche, les différents types de régresseurs. Aussi, bien que l'analyse de Fourier soit sans doute la meilleure méthode de décomposition en fréquence dans les cas stationnaires, on s'interdit ici son emploi afin de préserver le caractère universel de la méthode.

5. Ajustement

5.1 Décomposition harmonique

Les considérations du paragraphe 2 conduisent à décomposer la série étudiée en ses harmoniques. Et si chaque harmonique représente un cycle réel, comme le cycle journalier, le cycle saisonnier, etc., nous pourrons raisonnablement espérer composer un modèle prédictif (nous supposons, pour le moment, la série "stationnaire", cf. §4.1).

Le principe en est simple: soit une série Y, dont on désire prédire le comportement en fonction d'une variable A (éventuellement le temps). Y comporte *n* mesures $y_1 \ldots y_n$ parallèles aux *n* mesures $a_1 \ldots a_n$ de A. Soient $a_{\min}$ et $a_{\max}$ les minimum et maximum des a_i. On cherche la fonction sinusoïdale de α et φ, notée « α~φ~A » de la forme :

$$\cos inus((\alpha / u)\frac{a_i - a_{\min}}{a_{\max} - a_{\min}}\pi + \varphi)$$

qui est le mieux corrélée à Y. Le nombre α parcourt la suite des entiers positifs; u est un paramètre réglable (≥1) : plus il est grand, plus la précision du balayage en fréquence est grande, mais le temps de calcul augmente. Par exemple : si *u = 10, α = 34 et φ = 49*, alors le nom est *« 34~49~A »*. Mais si *u = 1,* le nom est *« 3~49~A »* et non pas *« 3,4~49~A »*, car α est toujours entier. D'autre part, dans le nom, la phase φ est donnée en degrés. Cette notation propre au logiciel CORICO vise à réduire la taille du nom dans les équations et sur les figures, en n'utilisant que des valeurs entières de α et φ tout en conservant une précision suffisante pour la fréquence et la phase de l'ondulation (dans les exemples illustrant cet article, la valeur u=10 s'est révélée suffisante; la valeur *u = 100* n'apporte pas de changement notable dans la décomposition).

Exprimée dans l'unité de A, la période de l'onde est $T = 2.(a_{\max} - a_{\min}) / (\alpha / u)$

Bien sûr, avec ce système de notation, un même harmonique portera un nom différent selon la plage d'échantillonnage sur laquelle il aura été calculé, comme on le voit par exemple pour le premier harmonique de RT Cygni dans les trois plages d'échantillonnage considérées au §7. Il faut savoir que le nom est simplement destiné à distinguer un harmonique d'un autre dans le modèle, et accessoirement, pour deux harmoniques de même fréquence, à connaître leur différence de phase. En effet, la période et le graphe de chaque harmonique sont systématiquement donnés par le logiciel.

Une fois trouvée la fonction « α~φ~A » qui sera le premier régresseur du futur modèle de Y, on retire cette composante de Y (par les corrélations partielles, [LES99]), afin d'obtenir le résidu non expliqué, et l'on cherche quelle est la fonction sinusoïdale qui approche le mieux ce résidu. Ce sera le second régresseur. Puis l'on extrait le résidu du résidu. Etc.. Les régresseurs sont donc connus par *ordre d'importance*. Lorsqu'on a trouvé un nombre suffisant de régresseurs, il reste à calculer les coefficients du modèle de régression multiple.

Le modèle obtenu n'exige pas d'hypothèse sur ce qui se passe dans les intervalles entre les mesures (d'où sa robustesse, qui sera sans doute encore plus grande si le pas d'échantillonnage n'est pas constant, afin d'éviter d'éventuels effets stroboscopiques). Ce modèle est une fonction du temps (il contiendra bien sûr des termes décrits aux paragraphes suivants si la série n'est pas stationnaire). Nous pouvons donc l'utiliser, mais seulement maintenant, pour interpoler entre les mesures, ou pour extrapoler au-delà du domaine expérimental.

5.2 Ruptures de tendances

Par leur caractère périodique, les composantes harmoniques sont commodes pour la prévision. Mais il est d'autres composantes qui, sans être sinusoïdales, peuvent néanmoins servir à la prévision. Nous abordons ici des termes "non stationnaires" de la série.

Sur la **figure 1 (annexe)**, la variable « ruptur1 » qui d'abord est constante, commence à croître à partir de l'observation 24. La rupture de tendance était imprévisible, mais puisqu'elle se poursuit jusqu'à la dernière observation, on peut penser, avec quelques chances de succès, que la croissance va se poursuivre quelque temps encore. Si donc nous remarquons que la variable A comporte les composantes n5 et « ruptur1 », nous pouvons construire un modèle de quelque utilité pour le proche avenir, toutes choses égales par ailleurs.

De même la variable « ruptur4 », qui d'abord décroît, puis reste constante, se compose avec n5 pour donner D. La variable « ruptur5 », qui décroît, puis croît, se compose avec n5 pour donner E, d'une part, et G, d'autre part. La différence entre E et G réside seulement dans l'amplitude de l'ondulation. Enfin, la variable « marche », qui manifeste un saut brutal vers l'observation 33, se compose avec n5 pour donner C.

Même inexpliquée, une rupture doit être connue pour en dégager la composante sinusoïdale. La prévision suppose qu'un tel événement est rare et ne se reproduira pas dans la suite. Hypothèse peu acceptable si le modèle se compose déjà de plusieurs ruptures.

Soit X une variable « explicative » (éventuellement le temps). CORICO distingue 5 types de ruptures, notées comme suit dans les sorties et les figures:

- « -<*i*.X » : un plat suivi d'une pente positive ou négative (rupture à l'observation n°*i*),
- « >-*i*.X » : une pente suivie d'un plat, (rupture à l'observation n°*i*),
- « *a*V*i*.X » : une pente suivie d'une pente inversée (double pente), où *i* est le numéro de l'observation et *a* caractérise la relation entre les deux pentes (*si a<45*°, la première pente est plus faible que la seconde).
- « *d*>*f*.X » : marche d'escalier = un plat, suivi d'une pente positive ou négative, suivi d'un plat, où *d* et *f* sont respectivement les numéros des observations de début et de fin de la pente.

- « *c{k.*X » : marche d'escalier "émoussée" (courbe logistique de type $\frac{1}{1+e^{-k(x-x_c)}}$), où *c* est le numéro de l'observation au centre de la pente, et *k* est un nombre caractérisant cette pente.

CORICO cherche la fonction *rupture(X)* la mieux corrélée à la série Y, en balayant les valeurs de *i, a, d, f, c et k*. L'ordre d'importance des régresseurs est trouvé en amont du calcul effectif de la régression, de la même façon que pour les composantes harmoniques (voir paragraphe 5.1). Cette liste de variables adaptées à la prévision n'est pas limitative. On peut y ajouter la variable X elle même, pour détecter la tendance simplement croissante ou décroissante, et les "interactions" de X avec elle-même pour détecter divers types de tendances présentant une courbure : X&X, par exemple donne accès aux tendances d'allure plutôt exponentielles [LES99]. Ainsi évite-t-on souvent de recourir au modèle multiplicatif (qui consiste à travailler sur le logarithme de la série).

5.3 Saisonnalités parasites

Une saisonnalité (harmonique) parasite apparaît quand on cherche à expliquer par une onde ce qui relève d'un autre phénomène. La méthode précédente n'exige pas un échantillonnage à pas constant et ne recourt ni au calcul différentiel, ni aux moyennes mobiles qui impliquent le mélange de plusieurs instants d'observations. Ainsi se trouve éliminée une importante source d'erreurs et de saisonnalités parasites. Cependant lorsque les différentes fonctions (sinusoïdes, ruptures,...) par lesquels CORICO essaye d'approcher la réponse ne suffisent pas, il peut encore apparaître des saisonnalités parasites à partir du second régresseur.

Mais, alors que la méthode des moyennes mobiles fonctionne comme une « boîte noire », parce que s'y mélangent les instants précédents et suivants (voir § 4.3), il est plus facile ici, grâce au tracé des différentes composantes trouvées par CORICO, d'apercevoir les causes des saisonnalités parasites, comme le montre l'exemple ci-après, et donc d'y remédier (voir §5.4).

Sur la **figure 2a (annexe)**, la variable M à modéliser forme une marche d'escalier à bords arrondis. Une analyse sans la recherche des fonctions logistiques, donnerait le modèle suivant :

ModelM = 1.64759 + 4.60277 32>40.t + 0.06014 309~90~t **(1)**

Où les coefficients de détermination sont R2 = 0.99841 et R2 ajusté = 0.99839; et le coefficient de Fisher est F = 0.3052E+05.

Le premier régresseur est donc une marche d'escalier, mais le second régresseur : *309~90~t* est une ondulation parasite qui n'existe pas sur l'original M.

Pour expliquer cela, on a représenté sur la **figure 2a**, le résidu *dM* obtenu lorsqu'on retire de la variable M (grâce aux corrélations partielles) la composante liée au premier régresseur *32>40.t.* Certes, le second régresseur *309~90~t* explique bien les variations observées sur dM, mais il s'étend, indûment, de moins l'infini à plus l'infini. En effet, tel le serrurier muni d'un passe partout, CORICO cherche à ouvrir une porte dont la clé est perdue. La clé la moins mauvaise trouvée dans le « trousseau » est *309~90~t.* Heureusement, le coefficient appliqué à ce régresseur dans le modèle (1) est environ 77 fois plus faible que celui du premier régresseur.

En réalité, chaque fois que le logiciel « tombe » sur un régresseur de type « marche d'escalier », il essaye de le remplacer par une fonction logistique. Le modèle trouvé par CORICO est donc le suivant:

ModelM = 1.64759 + 4.60766 35{22.t - 0.03696 27~0~t **(2)**

Où R2 : 0.99974; R2a = 0.99973 ; F= 0.1815E+06.

Le premier régresseur: *35{22.t* est une fonction logistique (**figure 2b, annexe**). Comme les courbures de M ne sont pas exactement celles d'une fonction logistique, le premier régresseur ne suffit pas à expliquer M. Aussi le second régresseur: *27~0~t* est encore une saisonnalité parasite, mais il est affecté d'un coefficient presque deux fois plus faible que celui du modèle (1). Aussi n'est-il pas perceptible à l'oeil nu dans le dessin du modèle (2), **figure 2b**. On pourrait peut-être se contenter du premier régresseur pour expliquer la réponse M, mais il est encore possible de faire mieux.

5.4 Morceaux d'ondes

Une ondulation qui se poursuit indéfiniment permet la prévision; mais elle a le défaut de ses qualités : toute erreur se prolonge indéfiniment! Sur la **figure 2a**, l'ondulation 309~90~t engendre une erreur plus grave que le « défaut » dM qu'elle voulait combattre. Sur la **figure 2b**, la fonction logistique permet d'atténuer cet inconvénient. Mais tous les défauts ne peuvent être rattrapés ainsi, et il peut arriver que, suite à plusieurs « défauts » de type dM, d'autres ondulations parasites, destinées à y remédier, se renforcent au point de devenir gênantes. Si l'on demande un troisième régresseur dans le modèle (1), il va servir à compenser l'ondulation engendrée par le second régresseur. L'erreur originelle de modélisation se propage sur les régresseurs suivant, même si à chaque étape les coefficients deviennent de plus en plus négligeables.

Un remède simple est d'approcher le « défaut » dM par un « morceau d'onde ». Le modèle trouvé par CORICO est alors le suivant:

ModelM = 1.64759 + 4.61530 35{22.t + 0.07240 31|35|252~0~t + 0.03764 37|44|57~0~t **(3)**

R2 = 0.9999673; R2a: 0.9999667; F = 0.1023E+07

Le premier régresseur est encore la fonction logistique *35{22.t*, mais le second est cette fois un morceau d'onde compris entre les observations n°31 et n°35, noté « *31|35|252~0~t* ». Un autre morceau d'onde: « *37|44|57~0~t* » sert à expliquer le résidu non encore expliqué par les deux premiers régresseurs. Comme ces morceaux d'onde sont localisés (cf. **figure 2c, annexe**), ils n'ont pas d'incidence sur la prédiction obtenue lorsqu'on extrapole le modèle (3).

Les « morceaux d'ondes » sont un moyen de « nettoyer » la série à modéliser de ses événements accidentels qui ne peuvent servir à la prévision. Ils confèrent au modèle une plus grande robustesse.

En revanche si la morphologie de la courbe ressemble à une ondulation, à cela près que les maxima sont par exemple plus pointus que ceux d'une sinusoïde, ce « défaut », comme tout défaut régulièrement répété, va engendrer des harmoniques d'ordre supérieur. Ces harmoniques reflètent un phénomène interprétable et sont utiles pour la prévision.

5.5 Points atypiques ou aberrants

La détection d'un pic isolé est aussi inutile que celle d'un morceau d'onde pour la prévision, vu son caractère imprévisible (sauf si l'on peut le relier à une autre variable prévisible). Mais cette détection est nécessaire, car un point aberrant peut gêner l'estimation des coefficients de la régression. Or si la méthode précédente permet d'approcher un « défaut » dM localisé entre deux instants, elle ne permet pas de détecter un point atypique qui par définition apparaît sur un seul instant.

Un remède simple est d'approcher le point atypique par une variable indicatrice partout nulle, sauf à l'instant considéré. Considérons la courbe C de la **figure 2d (annexe)**. Elle comprend 100 observations de janvier1963 à avril 1971, avec deux points aberrants. S'il ne détectait pas ces points, ni les morceaux d'onde, CORICO donnerait:

ModelC1 = 1.755 + 7.227 202~60~t + 4.164 33>34.t + 1.509 209~274~t + 0.988 187~0~t
- 1.043 124~0~t + 1.033 79>81.t

R2 = 0.92678; R2a: 0.92288; F = 196.2

Mais comme CORICO détecte les points atypiques et morceaux d'onde, il donne (**figure 2d**) :

ModelC2 = 1.755 + 7.411 202~60~t + 4.593 33>34.t - 2.169 85|90|153~271~t + 1.936 |obs24 + 1.351 195~0~t + 0.909 |obs10

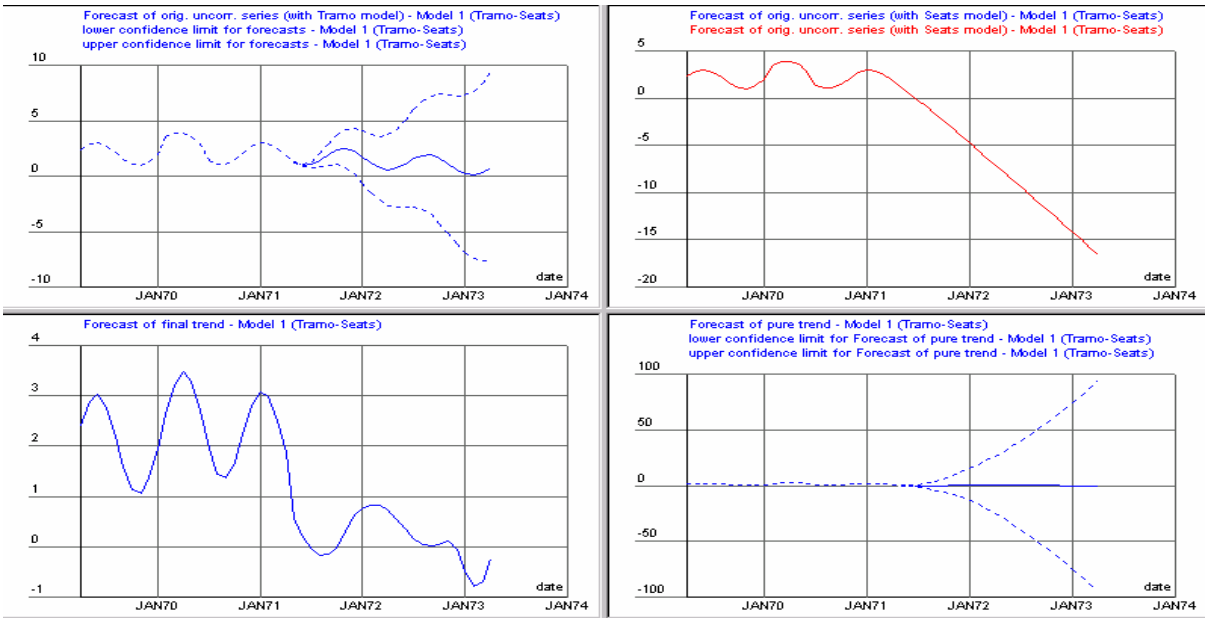
R2 = 0.98402; R2a: 0.98317; F = 954.8

Outre une marche d'escalier, des ondes sinusoïdales et un morceau d'onde, le modèle prend en compte les deux points atypiques, représentés par des variables indicatrices notées |obs24 et |obs10.

Sur la **figure 2e (annexe)**, ModelC2 est utilisé pour extrapoler le modèle ci-dessus, de mai 71 à 1975 (observations 101 à 150). L'extrapolation, qui suppose l'absence, à court terme, d'événement aléatoire, est facile car chaque composante est elle même extrapolable de façon simple. Ce modèle fournit aussi les valeurs des observations 10 et 24 qu'on aurait obtenues en l'absence des points aberrants. A titre de comparaison, ModelC1 montre le modèle incorrect obtenu quand on ne cherche pas à détecter les points atypiques et les morceaux d'onde.

La série C a été traitée ensuite avec le logiciel Tramo/Seats développé par Agustín Maravall et disponible sur le site Internet de la Banque d'Espagne : la **figure 2f** ci-dessous montre la difficulté d'obtenir, dans ce cas d'école, une prévision fiable de la tendance avec une méthode ARIMA : en haut à gauche (modèle Tramo), la courbe en trait plein montre la partie prédite (au delà d'avril 71). C'est à tort que l'amplitude décroît et que l'ondulation "descend". En haut à droite (modèle Seats), le graphe descend, à tort, au delà d'avril 71. En bas à gauche (modèle Tramo-Seats) la tendance descend, à tort, au delà d'avril 1974. En bas à droite on aperçoit aussi, mais plus difficilement à cause de l'échelle, cette tendance à "descendre".

Fig. 2f



5.6 Processus autorégressif d'ordre 2

L'exemple précédent était peu adapté aux méthodes traditionnelles. A contrario, on pourrait penser que la méthode CORICO ne s'applique pas à un processus fonction du passé, tel le processus autorégressif d'ordre 2 : $X_t = 0.9 X_{t-1} + (0.9)^2 X_{t-2} + \varepsilon_t$, où ε_t est un "bruit blanc" (variable aléatoire de moyenne nulle non corrélée à ses valeurs passées). En effet, l'algorithme de CORICO, purement géométrique, exclu l'influence du passé (cf. §2, §3 et §4.4). Or les performances d'un sportif, par exemple, prennent en compte l'effet des semaines précédant la mesure (entraînement, prise d'aliment énergétique, fatigue, etc..). En réalité, si ces effets perdurent, ils sont encore "présents". Il est donc légitime, si la cadence d'échantillonnage est régulière, de créer (sous Excel, ou avec CORICO) les séries X_{t-1}, X_{t-2}, etc., par simple
décalage des observations de la série initiale. Après quoi, la méthode géométrique peut s'appliquer. La sélection des régresseurs (§4.4) tient compte de ces nouvelles variables, au même titre que des saisonnalités, ruptures, interactions, morceaux d'onde et points atypiques. Les variables qui ont un effet sont intégrés au modèle, les autres non.

Sur la **figure 2g (annexe)**, la variable X_t, tirée de l'annexe VII.A de [GOU83], est formée de valeurs simulées du processus autorégressif ci-dessus. CORICO donne le modèle suivant.

$$ModelXt = -0.037 + 11.986\,(X_{t-1} - X_{t-2}) - 2.329\,|e112 + 2.285\,645{\sim}0{\sim}t + 1.851\,35\{100.t - 1.586\,|e114$$

R2 = 0.75839; R2a: 0.75344; F = 121.8

A cause du bruit blanc, la valeur R2 est ici plus faible que dans les précédents exemples. Le régresseur principal : $(X_{t-1} - X_{t-2})$ qu'il faut considérer comme un tout, voir [LES99], fait intervenir les valeurs passées. Les autres termes de l'équation, d'une part, et, d'autre part les résidus (les différences entre la série et son modèle), correspondent au "bruit blanc". Si SCE est la somme des carrés des résidus et n le nombre de termes de la série, l'Erreur Standard de Prédiction dans l'unité de X est ici $\sqrt{SCE/n} = 0.53$

Bien sûr, le logiciel Tramo/Seats donne ici un modèle ARIMA exact. Il détecte, comme CORICO, le point atypique *e112*. En revanche, le deuxième point atypique détecté est *e114* pour CORICO, et *e127* pour Tramo/Seats. La différence vient du fait que CORICO explique préalablement une partie du "bruit blanc" par les composantes *645~0~t et 35{100.t.*

Sur ces données simulées, le "bruit blanc" est aléatoire par construction. Seul le premier régresseur est utile pour la prévision. Les termes suivants peuvent au moins nous aider à estimer la barre d'erreur sur cette prévision. Si, au contraire, il s'agissait de données physiques, les composantes *645~0~t et 35{100.t* pourraient correspondre à des effets interprétables, et contribuer à la prévision. Mais un terme atypique contribue toujours à la barre d'erreur, sauf s'il est possible de le relier à une variable prévisible. Par exemple, si *e112* correspond au seul jour où il a plu, et si, par ailleurs, nous avons de bonnes raisons de suspecter la pluie de jouer un rôle sur les performances sportives, nous pouvons, dans le modèle de prévision, remplacer la variable indicatrice |*e112* par une variable indicatrice de la pluie.

5.7 Motifs périodiques non sinusoïdaux

L'autocorrélation d'une série sur elle-même, qui est un moyen de détecter la répétition périodique d'un motif, même compliqué (§4.2), requiert cependant une cadence régulière d'échantillonnage. On a vu, §5.1, un autre moyen de trouver la période, quelle que soit la cadence. Le logiciel peut donc récupérer les valeurs de la série pour une période, et les recopier telles quelles dans chacune des périodes (il interpole éventuellement le motif si la cadence est irrégulière). Si la nouvelle série ainsi créée corrèle bien avec la série originelle, elle pourra faire partie des régresseurs.

Considérons par exemple la série H de la **figure 2h (annexe)**, fabriqué à partir d'une onde de période 50.77, à laquelle on a ajouté deux motifs périodiques quelconques, de période 12 et 20 respectivement. Un modèle trouvé, sans la recherche des motifs périodiques, est le *ModelH1* :

```
 0           3.066118955612            Constante
 1          35.967613220215 39~98~t              , période :     50.77
 2           7.820130825043 164~244~t            , période :     12.07
 3           3.722646236420 333~105~t            , période :      5.95
 4           2.123522043228 493~266~t            , période :      4.02
 5          -1.542114019394 |obs29
 6          -1.423042178154 |obs89
 7          -1.616799116135 165~0~t              , période :     12.00
 8          -1.114941120148 |obs91
```

R2 = 0.9826028; R2a: 0.9812791; F = 642.4

```
Erreur Standard de Prediction =  0.4896530    , dans l'unité de H.
```

Tandis que si l'on recherche les motifs périodiques, on obtient le *ModelH2* :

```
 0           3.066118955612            Constante
 1          35.403858184814 39~98~t              , période :     50.77
 2          10.360790252686 165~~t               , période :     12.00
 3           2.364821910858 99~~t                , période :     20.00
```

```
  4          -0.300035387278 32~0~t              , période :     61.87
  5          -0.276497274637 48~0~t              , période :     41.25
  6           0.183029130101 332~99~t            , période :      5.96
  7          -0.111364223063 162~0~t             , période :     12.22
  8          -0.083276063204 |obs56
```

R2 =0.9999310; R2a: 0.9999257; F = 0.1636E+06

Erreur Standard de Prediction = 3.0953284E-02, dans l'unité de H.

Le premier régresseur est le même pour les deux modèles. En revanche le deuxième régresseur 164~244~t, de période 12.07, est remplacé dans ModelH2 par le motif périodique (ou, si l'on préfère, "l'onde sans phase") noté "165~~t", de période 12, et dessiné sur la **figure 2h**. Comme il s'agit d'un cas d'école, fabriqué exprès, nous savons que la période est exactement 12. Si *ModelH1* a donné 12.07, c'est qu'il est difficile parfois d'approcher, par une sinusoïde, un motif périodique compliqué. Dès lors, les régresseurs suivants servent, d'une part, à compenser cette inexactitude de période (grâce aux ondes de période moitié et un tiers de la première onde), et, d'autre part, à compenser (grâce aux points atypiques) le fait que, en réalité, l'onde n'est pas sinusoïdale.

Par contre, *ModelH2* ne devrait comporter que les trois premiers régresseurs, puisque par construction, nous n'avions mis dans H que ces trois composantes. De fait, les cinq régresseurs suivants ont des coefficients très faibles, comparés aux deux premiers. Il s'agit donc de composantes destinées à compenser les erreurs numériques d'arrondis, inévitables dans tout calcul d'ordinateur. Le modèle comportant seulement les trois premiers régresseurs de *ModelH2* donne R2 =0.9997956; R2a: 0.9997913; F = 0.1565E+06.

Quoi qu'il en soit, même *ModelH1* donne d'excellentes valeurs pour R2, R2a et F, ce qui montre la robustesse de la méthode. Dans *ModelH2*, on remplace plusieurs composantes sinusoïdales par des motifs périodiques compliqués, moulés sur l'original. C'est un moyen radical d'obtenir de bons R2; à utiliser toutefois avec modération, car il signifie que, dans un souci d'efficacité de la prévision (du moins à cours terme), on renonce à expliquer la série. En effet, si la série observée est une somme de composantes sinusoïdales de fréquence différentes, le motif va se modifier peu à peu au cours du temps, comme les vagues de la mer. Dans ce cas de figure, un modèle du type *ModelH1* serait préférable à long terme. Par exemple, pour prédire la survenue d'un effet indésirable comme la vague solitaire "géante" qui découle de la composition de plusieurs pics au même moment.

Le modèle ARIMA de la série H distingue aussi la saisonnalité, de période 50.77, et le premier motif, de période 12. En revanche, ARIMA ne distingue pas le second motif, de période 20, qui se retrouve donc dans le résidu, alors que CORICO distingue la saisonnalité et les deux motifs; il pourrait encore fonctionner si la cadence d'échantillonnage n'était pas régulière, bien qu'il soit moins facile alors de déceler les motifs (le logiciel ne peut inventer l'information qui n'existe pas). La décomposition sous forme d'ondes sinusoïdale, elle, est encore plus robuste en cadence irrégulière qu'en cadence régulière, car peu de points suffisent alors à définir une onde de façon unique, tandis que plusieurs ondes de fréquences différentes peuvent s'ajuster à un échantillonnage régulier (effet stroboscopique).

6. Exemple d'utilisation du logiciel : une extrapolation du Trafic SNCF

Le **tableau 1** contient deux variables : le temps exprimé en mois et en années (18 ans x 12 mois = 216 mois), et le trafic des trains.

Tableau 1: Trafic de la SNCF en 2ème classe, en million de voyageurs/km

[GOU83]

	J	F	M	A	M	J	J	A	S	O	N	D
1963	1750	1560	1820	2090	1910	2410	3140	2850	2090	1850	1630	2420
1964	1710	1600	1800	2120	2100	2460	3200	2960	2190	1870	1770	2270
1965	1670	1640	1770	2190	2020	2610	3190	2860	2140	1870	1760	2360
1966	1810	1640	1860	1990	2110	2500	3030	2900	2160	1940	1750	2330
1967	1850	1590	1880	2210	2110	2480	2880	2670	2100	1920	1670	2520
1968	1834	1792	1860	2138	2115	2485	2581	2639	2038	1936	1784	2391
1969	1798	1850	1981	2085	2120	2491	2834	2725	1932	2058	1856	2553
1970	1854	1823	2005	2418	2219	2722	2912	2771	2153	2136	1910	2537
1971	2008	1835	2120	2304	2264	2175	2928	2738	2178	2137	2009	2546
1972	2084	2034	2152	2522	2318	2684	2971	2759	2267	2152	1978	2723
1973	2081	2112	2279	2661	2281	2929	3089	2803	2296	2210	2135	2862
1974	2223	2248	2421	2710	2505	3020	3327	3044	2607	2525	2160	2876
1975	2481	2428	2596	2923	2795	3287	3598	3118	2875	2754	2588	3266
1976	2667	2668	2804	2806	2976	3430	3705	3053	2764	2802	2707	3307
1977	2706	2586	2796	2978	3053	3463	3649	3095	2839	2966	2863	3375
1978	2820	2857	3306	3333	3141	3512	3744	3179	2984	2950	2896	3611
1979	3313	2644	2872	3267	3391	3682	3937	3284	2849	3085	3043	3541
1980	2848	2913	3248	3250	3375	3640	3771	3259	3206	3269	3181	4008

Cliquez au menu *Outil...Régression multiple* du logiciel CORICO.

Sélectionnez le fichier **sncf192.cor.** Il contient les 192 premières valeurs du **tableau 1**, soit les années 1963 à 1978.

Donnez un nom au fichier résultat, par exemple **aaa.cor**. Il contiendra les données d'origine plus les composantes trouvées, la variable prédite par le modèle et le résidu. Dans la liste « *Variable dépendante* » sélectionnez la variable à expliquer *Trafic*, et donnez un nom à son futur modèle dans la zone « *Nom du modèle* », par exemple *ModelTrafic.*

Cochez *Facteurs*, *Interactions*, *Résonances*, *Ruptures*, *Double pentes et Points atypiques.* Cochez *Morceaux d'ondes.* Ne cochez pas, pour le moment, *Motifs périodiques*.

Cliquez sur le *bouton Choix des régresseurs* et indiquez par exemple « 9 » régresseurs dans la zone en dessous du bouton *Aide au choix,* puis cliquez sur ce bouton. CORICO propose les 9 meilleurs régresseurs (vous pouvez connaître leur ordre décroissant d'importance en cliquant sur le bouton Ordre des régresseurs).

Cliquez OK, puis OK. Alors les coefficients sont calculés. Ils sont disponibles dans le fichier **sortie,** ainsi que la période des différents harmoniques, la valeur R2 du coefficient de détermination (carré du coefficient de corrélation entre *Trafic* et sa valeur prédite *ModelTrafic*) et celle du R2 ajusté et du coefficient F de Fisher.

ModelTrafic = 2454.6 + 4949.9 137{3.t - 3473.0 319~0~t + 2170.6 639~0~t + 1980.0 957~100~t - 887.9 307~0~t - 580.4 |nov.74 - 432.4 |nov.73 - 427.4 |nov.72+ 388.5 |dec.78

R2 = 0.9143, R2a = 0.9106 et F = 216.0

Dans le menu *Fichier...Ouvrir,* choisissez le fichier **aaa.cor.**

Dans le menu *Voir...Allure des variations*, sélectionnez la variable *t*, les neuf régresseurs harmoniques du modèle, la variable *ModelTrafic* et la variable *Trafic*. Puis Ok.

Sélectionnez le menu *Voir...Aperçu PostScript*. Agrandissez la fenêtre noire qui apparaît.

Vous obtenez la **figure 3 (annexe)**. Nous constatons que la variable *ModelTrafic* reflète assez bien la variable *Trafic.* Ici le logiciel n'a pas détecté de « morceaux d'ondes ». Le terme le plus important est une « rupture logistique » notée « 137{3.t » (dont la pente maximum est centrée sur le mois n° 137: mai 1974), c'est à dire une partie constante qui s'infléchit vers le haut puis redevient constante : dans les années 70 le trafic voyageur n'a cessé d'augmenter, pour se stabiliser au milieu de 1978. Là dessus se surajoutent les variations saisonnières et cycliques (vacances...) représentées par quatre régresseurs harmoniques. Les harmoniques de haute fréquence, comme 957~100~t, ne semblent pas avoir une ondulation régulière. Il n'en est rien: cet effet apparaît quand on relie par un trait des points relevés seulement tous les mois. Quatre point atypiques sont détectés : valeurs anormalement faibles (coefficient négatifs), en novembre 72, 73,74; valeur anormalement forte en décembre 78.

Tentons une prédiction en extrapolant le modèle aux années 1979-80 (mois 193 à 216) :

Sélectionnez le menu *Outil...Optimiser la réponse en fonction du modèle*. Et répondez comme ci-dessous en **gras**. Vous obtenez :

O P T I M I S E la R E P O N S E !

Fichier f:\essai\corient\sncf192.cor

M doit etre > 2

Coefficients du modèle de Trafic

2454.562500000000 Constante

4949.948730468750 137{3.t

-3473.027343750000 319~0~t

2170.648437500000 639~0~t

```
 1980.034179687500  957~100~t
 -887.858825683594  307~0~t
 -580.417907714844  |nov.74
 -432.370788574219  |nov.73
 -427.444000244141  |nov.72
  388.521057128906  |dec.78

S'agit-il d'un mélange ?
 Oui / Non
n

Fréquence    319 => période t          =  11.9749
Fréquence    639 => période t          =  5.97809
Fréquence    957 => période t          =  3.99164
Fréquence    307 => période t          =  12.4430


Modèle fonction de la seule variable t
 MINIMISATION
          Unité du calcul   Unité d'origine
t            -0.76286E-01        38.06594

 MAXIMISATION
          Unité du calcul   Unité d'origine
t             0.11820E+00        187.0360

Réponse minimum prévue par le modèle :   1648.878
Réponse maximum prévue par le modèle :   3705.819

Voulez-vous :
 1 - la réponse prévue par le modèle pour une
     combinaison particulière des facteurs ?
 2 - appliquer le modèle à un autre fichier ?
 3 - extrapoler hors du domaine ?
```

(le modèle dépend d'une seule variable)

4 - les facteurs prévus pour une réponse donnée ?

q - Quitter ?

3

VALEUR MINIMUM de la plage de t

Donnez une valeur de t

(domaine initial: 1.000000 192.0000)

1

VALEUR MAXIMUM de la plage de t

Donnez une valeur de t

(domaine initial: 1.000000 192.0000)

216

Donnez le PAS (0.307143E-01 < PAS < 71.6667)

1

plage : 1.000000 à 216.0000 , PAS = 1.000000

EXTRAPOLATION

Données préparées pour CORICO dans le fichier :

E:\corrette\MDL.COR

Consultez-les au menu Fichier...Voir, modifier

Voulez-vous :

1 - la réponse prévue par le modèle pour une
combinaison particulière des facteurs ?

2 - appliquer le modèle à un autre fichier ?

3 - extrapoler hors du domaine ?
(le modèle dépend d'une seule variable)

4 - les facteurs prévus pour une réponse donnée ?

q - Quitter ?

q

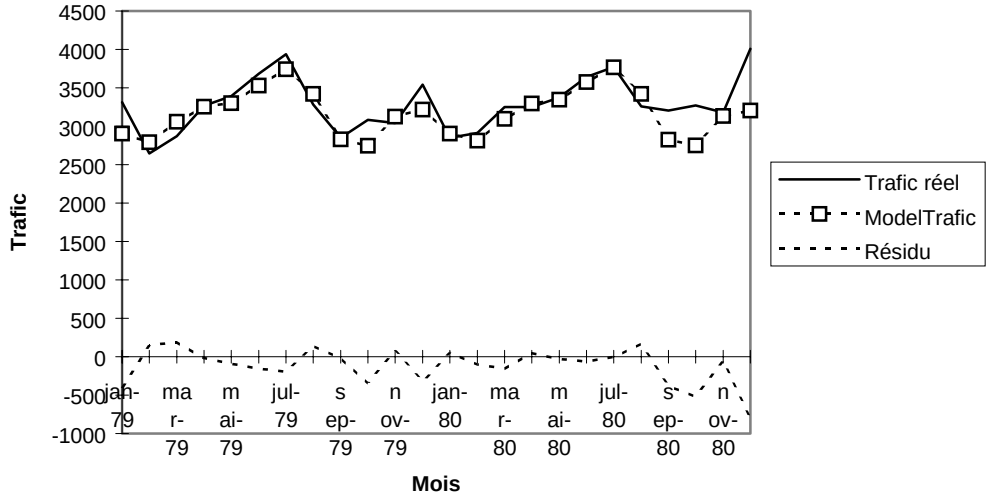
Pour plus de détails, voir fichier Sortie.

Vous constatez qu'après avoir rappelé les coefficients du modèle par ordre d'importance décroissante, le programme donne la période de chacun des harmoniques dans

l'unité de la variable explicative, ici le mois. La première est égale à 11.9749 mois, c'est-à-dire en gros annuelle. Elle est suivie par une onde de période 6 mois environ, une autre d'environ 4 mois; puis vient une onde de période supérieure à 1 an. Légèrement différente de la première onde annuelle, elle se décale donc peu à peu au cours du temps sur la **figure 3** (il peut s'agir d'un effet stroboscopique si l'échantillonnage par pas de un mois n'est pas assez fin, mais il peut aussi s'agir (de même que les périodes 6 et 4) du rattrapage de l'imperfection d'ajustement par le régresseur de période 11.9749. En effet ce dernier présente des formes sinusoïdales assez rondes, tandis que les pics du trafic, certes cyclique, sont plus pointus). Rappelons que la phase est notée en degré dans le nom. Le programme indique que c'est au 38ème mois que se trouve la valeur minimum du Modèle de *Trafic*, et au 187ème mois que se trouve la valeur maximum.

Quand vous choisissez d'extrapoler hors du domaine initial (de 1 à 192 mois), le programme vous demande le début et la fin de la période choisie, ainsi que le pas de variation. Ici, en choisissant entre les mois 1 et 216, par pas de 1 mois, vous avez extrapolé sur les 24 mois à venir (1979-80). Le listing indique que le résultat de l'extrapolation se trouve dans le fichier **mdl.cor.** La **figure 4a** ci-dessous montre la comparaison entre les valeurs prédites sur cette période et les valeurs observées tirées des deux dernières lignes du **tableau 1**. La prédiction est assez bonne pour 1979. C'est seulement à la fin de 1980 que la prédiction commence à s'écarter vraiment de l'observation.

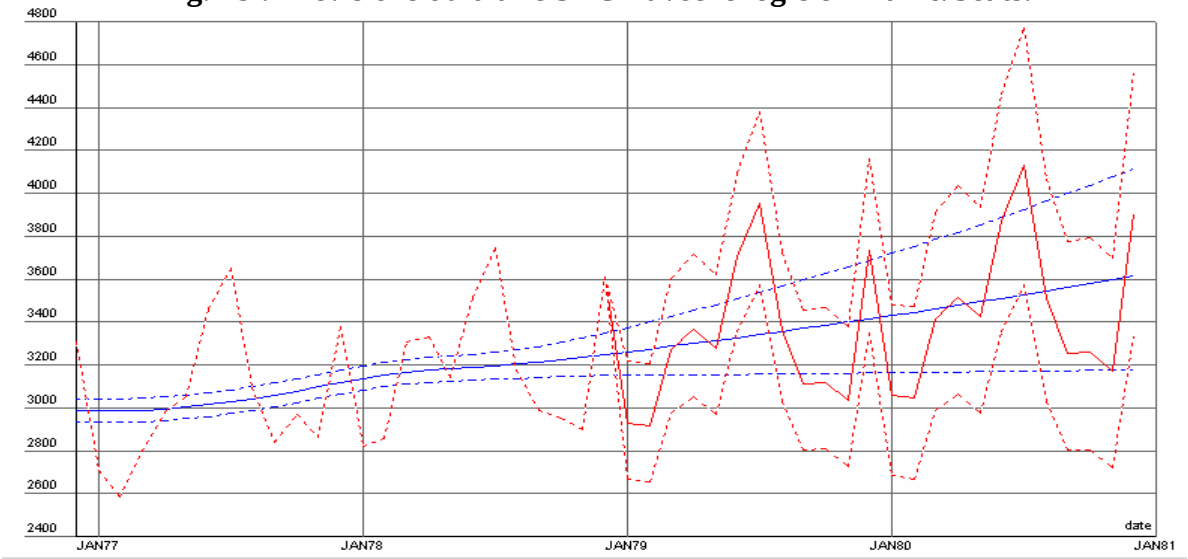
Fig. 4a: comparaison des valeurs prédites aux valeurs observées, d'après CORICO.



La série SNCF se prête bien aux modèles ARIMA classiques. De fait le modèle obtenu avec le logiciel Tramo/Seats sur la période 1963-1978 donne un meilleur R2 que celui obtenu par CORICO. En effet, par autocorrélation il est plus facile de reproduire une structure périodique d'allure tourmentée. Ce qui n'entre pas dans la structure périodique se retrouve dans la tendance. Mais cette dernière, assez irrégulière n'est pas si aisément prolongeable que les composantes décrites au §5.2. Ici composantes déterministes et aléatoires sont mélangées contrairement au principe énoncé au paragraphe 3. Quand le logiciel extrapole le modèle aux années 1979-1980 (**figure 4b** ci-dessous), la tendance continue sa croissance vers le haut, alors que les données du **tableau 1** marquent au contraire un palier (la courbe logistique de CORICO). On constate donc ici des pics un peu trop élevés. La méthode classique, excellente

pour l'interpolation, dans le cas d'une cadence régulière d'échantillonnage, se prête moins à l'extrapolation.

Fig. 4b : Prévisions du trafic SNCF avec le logiciel Tramo/Seats.



Maintenant relançons l'analyse précédente de CORICO, en demandant, en plus, les "motifs périodiques" (voir §5.7) :

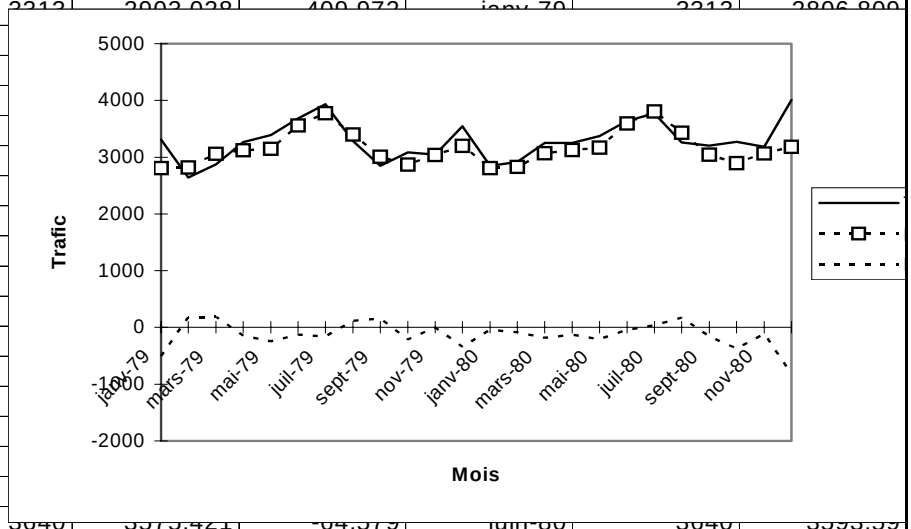
0	2454.562500000000	Constante		
1	4864.319335937500	137{3.t		
2	4889.003417968750	319~~t	, période :	11.97
3	-688.676635742187	329~0~t	, période :	11.61
4	-491.528594970703	304~0~t	, période :	12.57
5	-472.451782226562	|jun.71		
6	-443.661743164062	|jul.68		
7	-411.692016601562	|nov.74		
8	407.008544921875	|dec.78		
9	370.519317626953	646~0~t	, période :	5.91

R2 = 0.9831, R2a = 0.9665 et F = 585.2.

Le motif 319~~t a été trouvé. Au vu des coefficients R2, La qualité du modèle équivaut maintenant à celle du modèle ARIMA. Par rapport au premier modèle de CORICO, les points atypiques ont changé, et c'est normal, car la notion de point "atypique" est relative au modèle. En revanche, la meilleure qualité du modèle sur les données d'apprentissage, n'entraîne pas une meilleure qualité de la prévision. De fait, la **figure 4c** ci-dessous équivaut en gros à la

figure 4a. Ceci confirme la robustesse du premier modèle, et rejoint les remarques faites à la fin du §5.7.

Fig. 4c : prévision, d'après CORICO, avec prise en compte du motif périodique.



Trafic réel	ModelTrafic	Résidu		Trafic réel	ModelTrafic
3771	3768,256	-2,744	juil-80	3771	3807,629

7. Interpolation d'une étoile variable : RT Cygni[1]

Dans l'exemple précédent, une extrapolation du modèle a permis la prévision. Une autre application du modèle [LES01] vise à reconstituer les valeurs manquantes, par interpolation.

L'étoile RT Cygni de la constellation du Cygne fait partie des étoiles variables de type Mira. Les données sur sa *courbe de lumière* nous ont été aimablement fournies par l'AFOEV (Association Française des Observateurs d'Etoiles Variables). La **figure 5 (annexe)** représente en abscisse le temps, en jour julien du 8-12-1988 au 11-4-1993, et en ordonnées les variations de la *magnitude apparente* (d'autant plus faible que l'éclat est grand). C'est un exemple de cadence irrégulière d'échantillonnage: les observations à la jumelle, impossibles par temps nuageux, sont d'autant plus difficiles que l'éclat apparent est faible. D'où un moindre échantillonnage dans les fortes magnitudes et surtout l'hiver (un pic sur deux) où les maxima sont donc incertains.

NOTE : *L'échelle dite des* magnitudes apparentes *varie comme l'opposé de 2,5 fois le logarithme décimal de* l'éclat apparent, *le zéro de cette échelle étant fixé arbitrairement par un choix d'étoiles standard. Dans cette échelle, les astres très brillants ont des magnitudes très petites qui peuvent être négatives. Les étoiles de magnitude apparente visuelle supérieure à 6 ne sont pas visible à l'œil nu. Un observateur entraîné peut atteindre, à la jumelle, une*

[1] NDLR La constellation du Cygne inspire décidément les statisticiens : voir dans le Journal de la Société de Statistique de Paris un article la Voie Lactée en 1983 (4ème trimestre).

précision de 0,1. La Magnitude absolue *d'une étoile correspond à la valeur qu'aurait la magnitude apparente si l'étoile se trouvait à une distance de 32,5 années lumière*.

On a proposé deux explications de la variabilité des étoiles:

- une activité interne: pulsation de l'enveloppe de gaz, dont la période est en général extrêmement précise (*Céphéides)*, explosion (*novae ou supernovae)*, développement de taches étendues.
- une cause externe: étoiles doubles, étoiles à éclipse qui se recouvrent périodiquement... Il s'agit alors d'une variabilité simplement optique. Mais l'on peut envisager aussi des perturbations venant d'un compagnon éloigné, ou d'un changement d'orbite causé par une grosse planète.

Pour certaines étoiles variables, dont la période est peu régulière, il n'est pas aussi évident d'attribuer toutes les variations à des processus physiques se déroulant au sein de l'astre ou dans son proche voisinage. Des corps célestes en mouvements, nuages de poussières, etc. peuvent s'interposer entre l'étoile et nous, et pourraient expliquer des variations d'éclat apparent surajoutées à la pulsation naturelle de l'étoile. Ceci pourrait se manifester par des ruptures de tendances telles que « marche d'escalier », etc.

Une analyse harmonique de type CORICO est bien adaptée à l'échantillonnage irrégulier de la **figure 5** (1352 observations sur 1586 jours). On trouve le modèle suivant :

$$ModèleCygrt = 9.1044 + 52.9125\ 168\sim275\sim Tjulien + 10.2723\ 78\sim98\sim Tjulien - 7.0223\ 335\sim0\sim Tjulien - 6.2243\ 876>914.Tjulien - 5.6083\ 41\sim0\sim Tjulien + 4.5866\ 162\sim0\sim Tjulien + 3.3061\ 406\sim90\sim Tjulien + 3.1324\ 212\sim89\sim Tjulien + 3.0637\ 60>66.Tjulien$$

R2 = 0.9639, R2a = 0.9637 et F = 3978

Avec pour les harmoniques les périodes suivantes :

168~275~Tjulien,	période : 188.81
78~98~Tjulien,	période : 406.67
335~0~Tjulien,	période : 94.69
41~0~Tjulien,	période : 773.66
162~0~Tjulien,	période : 195.80
406~90~Tjulien,	période : 78.13
212~89~Tjulien,	période : 149.62

Ce modèle calculé sur 9 maxima et 8 minima de magnitude (jours juliens 2447502.3 à 2449088.3) donne la période 188.81 jours pour le premier harmonique. Une étude portant sur 82 maxima et 76 minima a montré que la période du premier harmonique de RT Cygni varie au cours du temps. En effet, [SCH93], [AND94] et [MAR96] ont calculé avec un programme d'ajustage multi harmonique une période de 191.86 jours sur l'intervalle 2423003 et 2424340, et une période de 189.4 jours sur l'intervalle 2446406 - 2450089.

En plus du premier harmonique, le modèle montre d'autres harmoniques et deux ruptures de tendance en « marche d'escalier », **figure 6 (annexe)**. Sur cette figure l'abscisse n'est pas le temps mais la suite des observations. Aussi la courbe de *Tjulien* n'est pas

rectiligne, mais présente des décrochements correspondant aux périodes hivernales moins riches en observations. La courbe des résidus du modèle (fortement amplifiée pour plus de clarté), suggère que les incertitudes d'estimation de la magnitude sont plus grandes durant les périodes hivernales. On peut supposer que le modèle qui s'appuie sur l'ensemble des observations, corrige ces incertitudes.

Par interpolation du modèle, nous pouvons reconstituer les valeurs manquantes. On obtient la **figure 7 (annexe)** où, maintenant, le temps est en abscisse. La rupture *60>66.Tjulien* est nettement visible sur la courbe *Modelcygrt* et on la retrouve en effet sur la **figure 5**. L'effet de la rupture *876>914* est deux fois plus important (puisque son coefficient dans le modèle est deux fois plus grand que celui de *60>66.Tjulien).* Cependant il est moins visible sur la courbe de *Modelcygrt,* car la pente de la marche d'escalier est plus douce, et, puisque le coefficient est négatif, elle est justement dans le même sens que celle de l'ondulation à cet endroit. Supposée la justesse du modèle, il appartient aux astrophysiciens d'essayer d'expliquer ces ruptures. Peut-être sont-elles en rapport avec les variations de période. Remarquons encore que la période du troisième harmonique est à peu près la moitié de celle du premier. A titre de comparaison voici les modèles obtenus, avec leurs coefficients et périodes,

- pour les 1228 observations du 7-4-75 au 28-10-84 (2442509.7 à 2445998.3):

8.94690	Constante		
-51.49132	368~0~Tjulien,	période :	189.60
7.92970	732~0~Tjulien,	période :	95.32
7.84573	197~0~Tjulien,	période :	354.17
-5.33046	66~0~Tjulien,	période :	1057.15
4.80206	358~274~Tjulien,	période :	194.89
3.84509	232~0~Tjulien,	période :	300.74
4.75360	144~0~Tjulien,	période :	484.53
4.39851	173~90~Tjulien,	période :	403.31
2.98583	596~90~Tjulien,	période :	117.07

R2 = 0.9479, R2a = 0.9476 et F = 2468

- et pour les 879 observations du 25-10-84 au 8-12-88 (2445998 à 2447502):

8.94209	Constante			
45.59826	157~315~Tjulien,	période :	190.69	
8.30712	26~0~Tjulien,	période :	1151.46	
6.69170	312~267~Tjulien,		période :	95.96
-5.13753	82~85~Tjulien,	période :	365.10	
4.89831	346{100.Tjulien,			
4.56045	42~270~Tjulien,	période :	712.81	

2.20510	156~0~Tjulien,	période :	191.91
2.89418	376~273~Tjulien,	période :	79.62
2.03045	475~269~Tjulien,	période :	63.03

R2 = 0.9659, R2a = 0.9656 et F = 2745

Ces résultats confirment la variation de période des harmoniques. Ces modèles peuvent donc seulement être utilisés pour l'interpolation mais non pour la prévision, sinon à court terme. Cependant, dans les trois cas on observe une harmonique de période à peu près moitié de l'harmonique principale (ce qui, soulignons-le, n'arrive pas pour d'autres étoiles Mira que nous avons testés, comme U Cygni, S Carinae de la Carène, ou R Vulpeculae du Petit Renard). On note aussi une « rupture logistique »: *346{100.Tjulien,* dans l'intervalle du 25-10-84 au 8 -12-88.

Note: Une fonctionnalité de CORICO (non décrite dans l'article) permet de tracer la courbe des variations de période au cours du temps.

8. Conclusion

[LES99] avait montré la recherche de modèles de régression fonction de plusieurs variables et d'interactions logiques de ces variables. Cet article montre des modèles fonction d'une seule variable et de fonctions cycliques ou de « ruptures » de cette variable. Rien n'empêche de combiner ces approches et de construire un modèle fonction de plusieurs variables, d'interactions, de fonctions cycliques et de « rupture » de certaines de ces variables, sans postuler de modèle a priori. La force de la méthode vient de ce que les variables ou fonctions (interactions, sinusoïdes, ruptures, morceaux d'ondes, points atypiques, motifs périodiques) introduites dans le modèle sont choisies par le programme en amont du calcul de la régression, et donc puisées dans un « vivier » qui peut être bien plus grand que le nombre d'observations, même si à la fin le modèle ne peut avoir plus de régresseurs que d'observations, puisqu'il ne peut y avoir plus d'équations que d'inconnues.

Dans la décomposition traditionnelle des séries temporelles en saisonnalité, tendance et résidu, la prévision est possible si la série désaisonnalisée (la tendance) présente une forme aisément prolongeable vers le futur. Ce qui n'est pas toujours le cas lorsque la tendance est irrégulière. La méthode CORICO conduit au contraire à décomposer la série en :

1. des composantes qui se prêtent à la prévision (saisonnalités, rupture de tendances, interactions, motifs périodiques...),
2. des composantes qui n'empêchent pas la prévision, car localisées dans le temps (morceaux d'onde, points atypiques). Ces composantes, dont la survenue est aléatoire, peuvent entrer dans l'évaluation de l'erreur de prévision, à condition d'estimer leur fréquence d'apparition.

Remerciements

L'auteur remercie Claude Lesty, et le comité de lecture de la revue, pour leurs précieuses suggestions.

RÉFÉRENCES

[AND94] Andronov, I.L., 1994, Odessa Astron. Publ., 7, 49.

[GOU83] Gourieroux C., Monfort A. Cours de Séries Temporelles, *Ed. Economica, Paris 1983.*

[LES99] Lesty M., *1999.* Une nouvelle approche dans le choix des régresseurs de la régression multiple en présence d'interactions et de colinéarités. *La revue de Modulad, n°22, pp.41-77.*

[LES01] Lesty M., 2001, Bull. AFOEV, n°96, pp. 8-13.

[MAR96] Marsakova, V.I., Andronov, I.L., 1996, Odessa Astron. Publ., 9.

[SCH93] Schweitzer E., 1993, Bull. AFOEV, n°64, 14.

ANNEXE

Fig.1 : Séries diverses avec ruptures de tendances

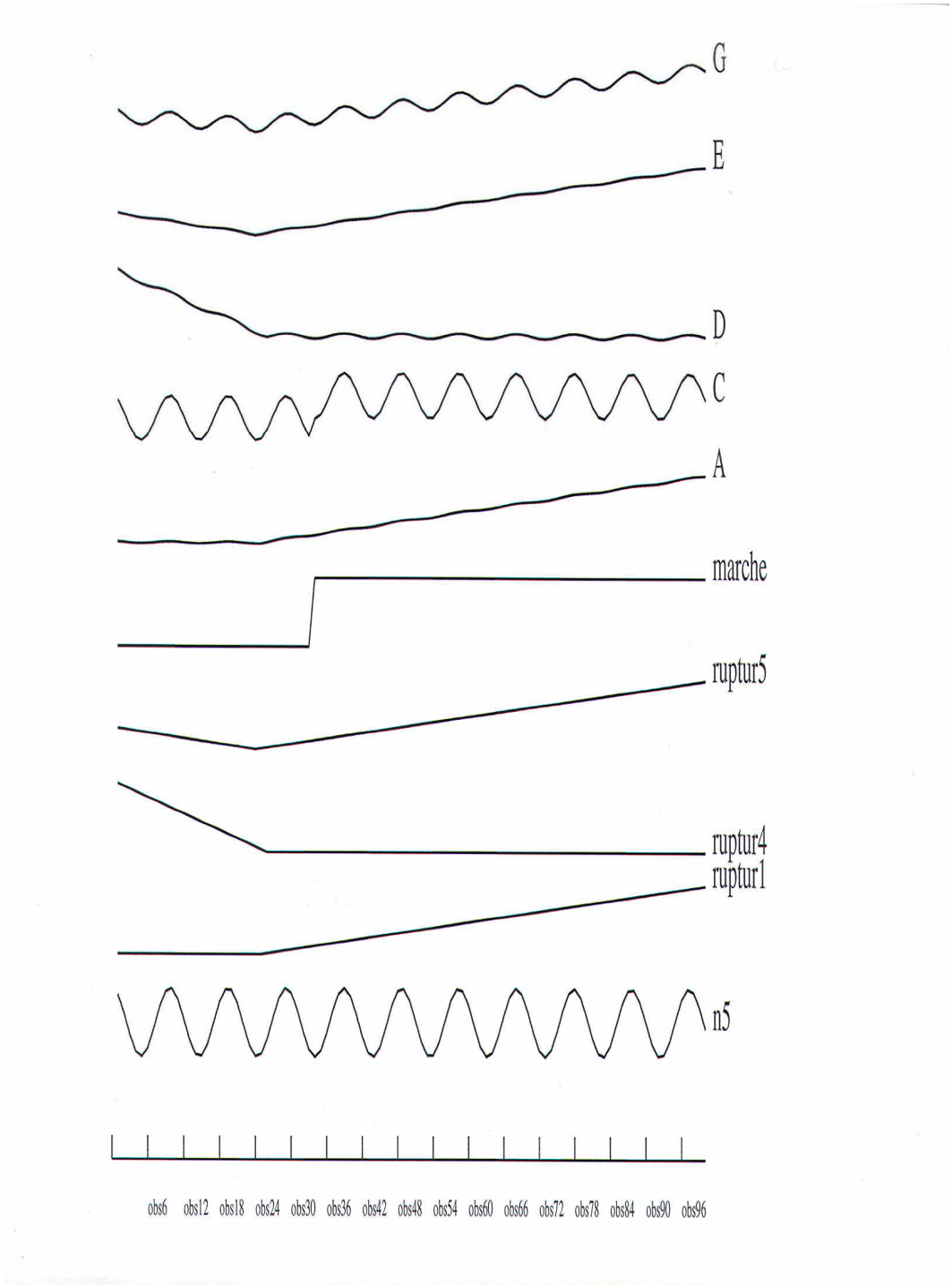


Fig.2a : Une saisonnalité parasite

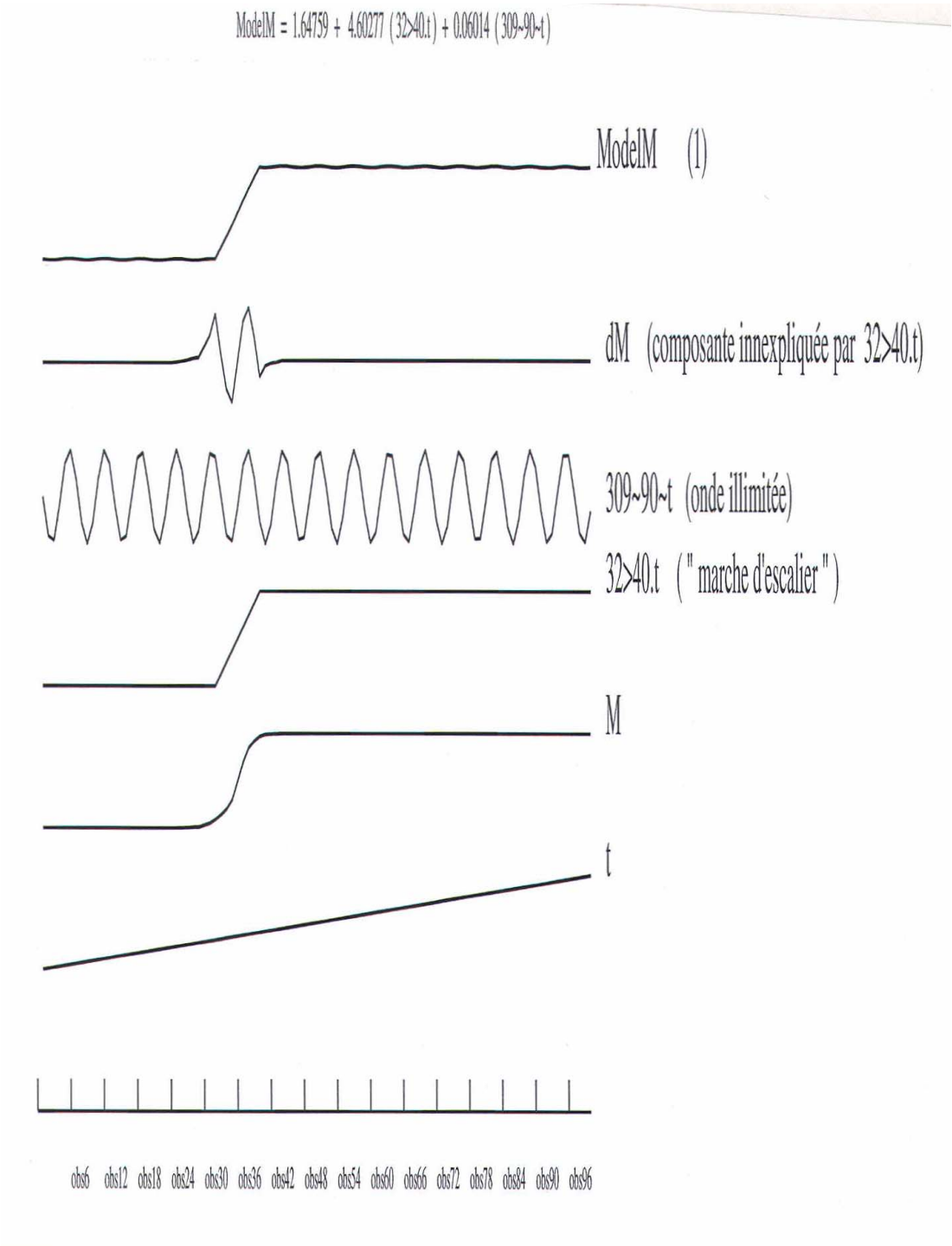


Fig.2b : Une saisonnalité parasite presque imperceptible

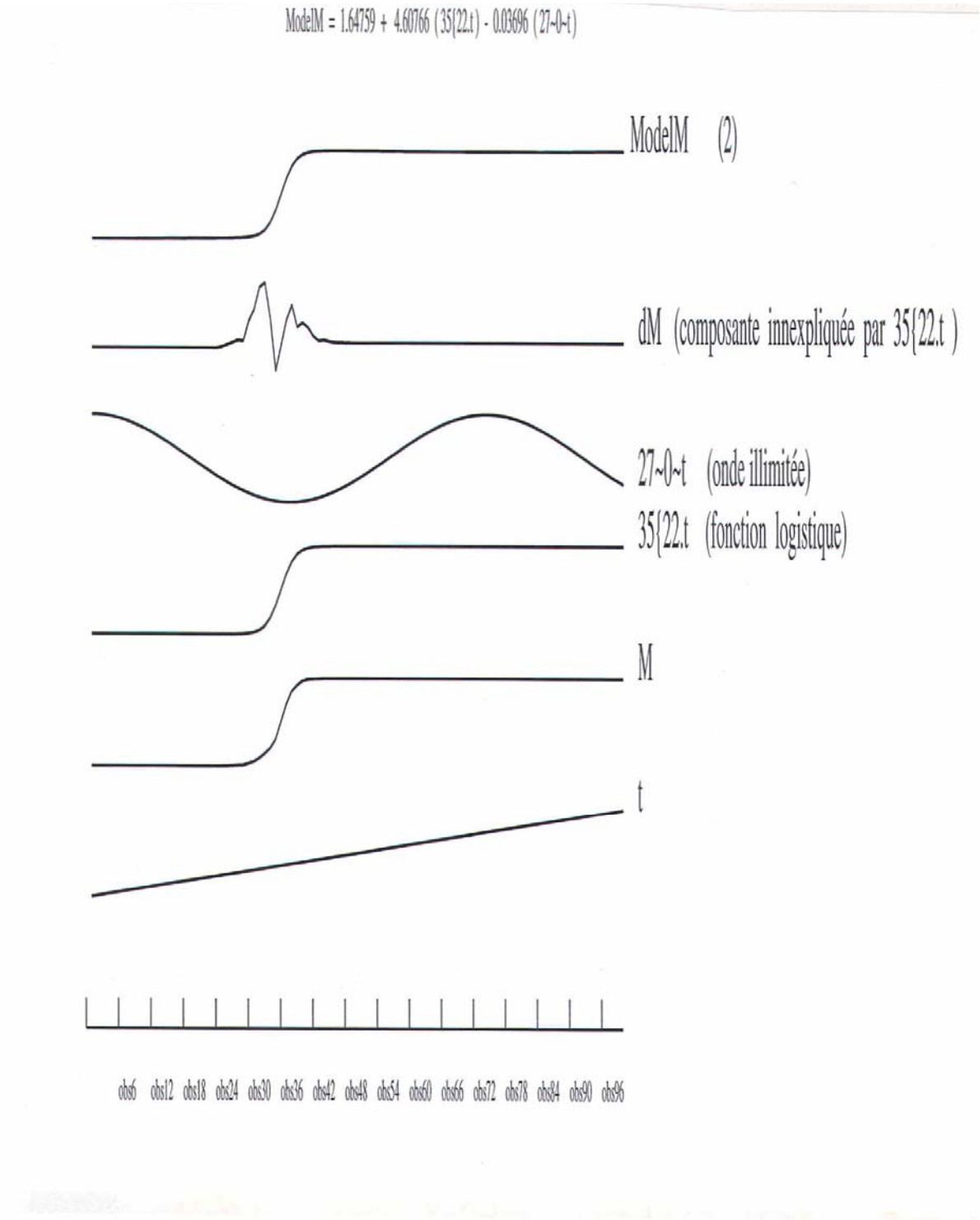


Fig.2c : Modèle sans saisonnalité parasite

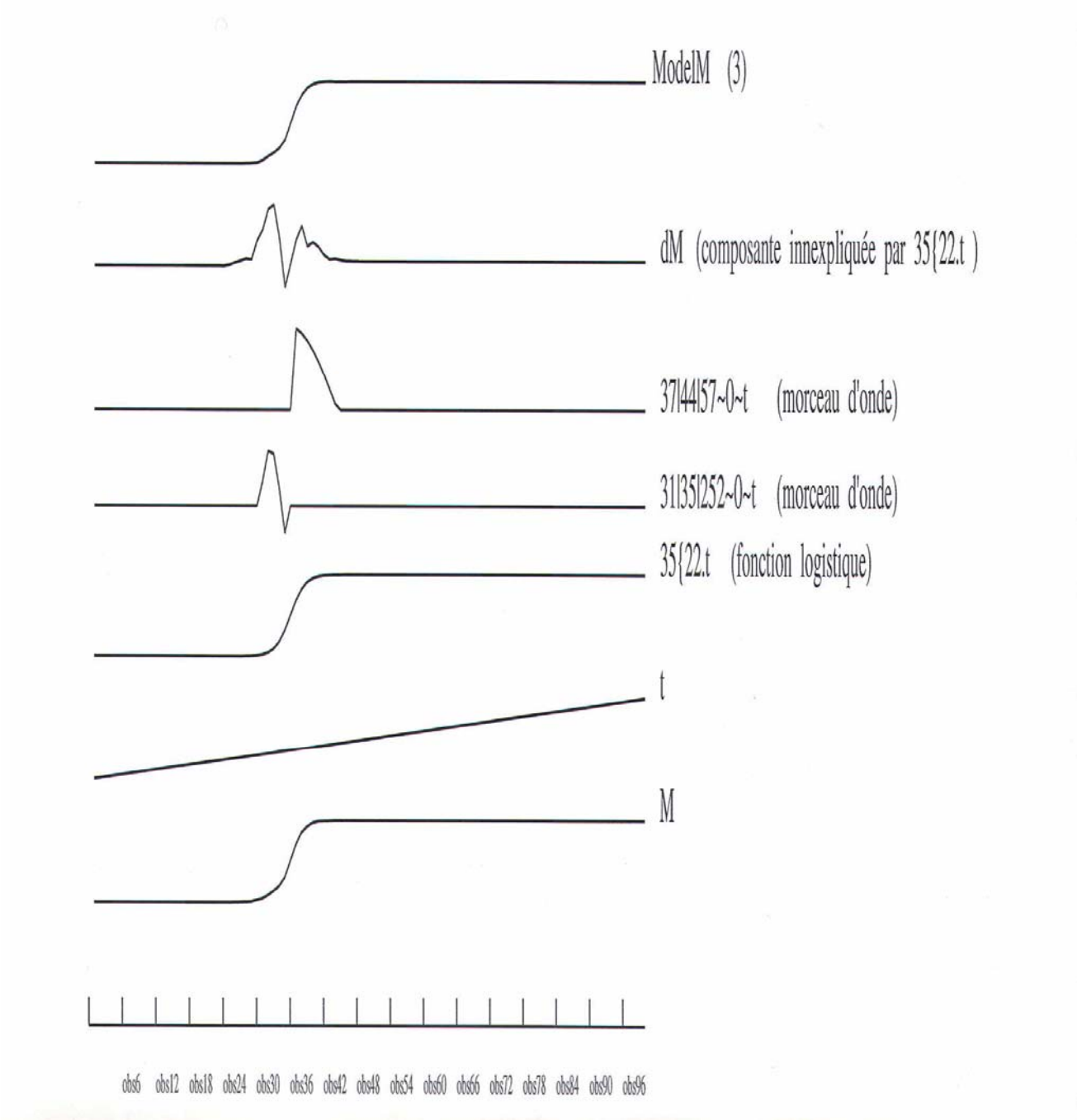


Fig.2d : Modèle avec points atypiques

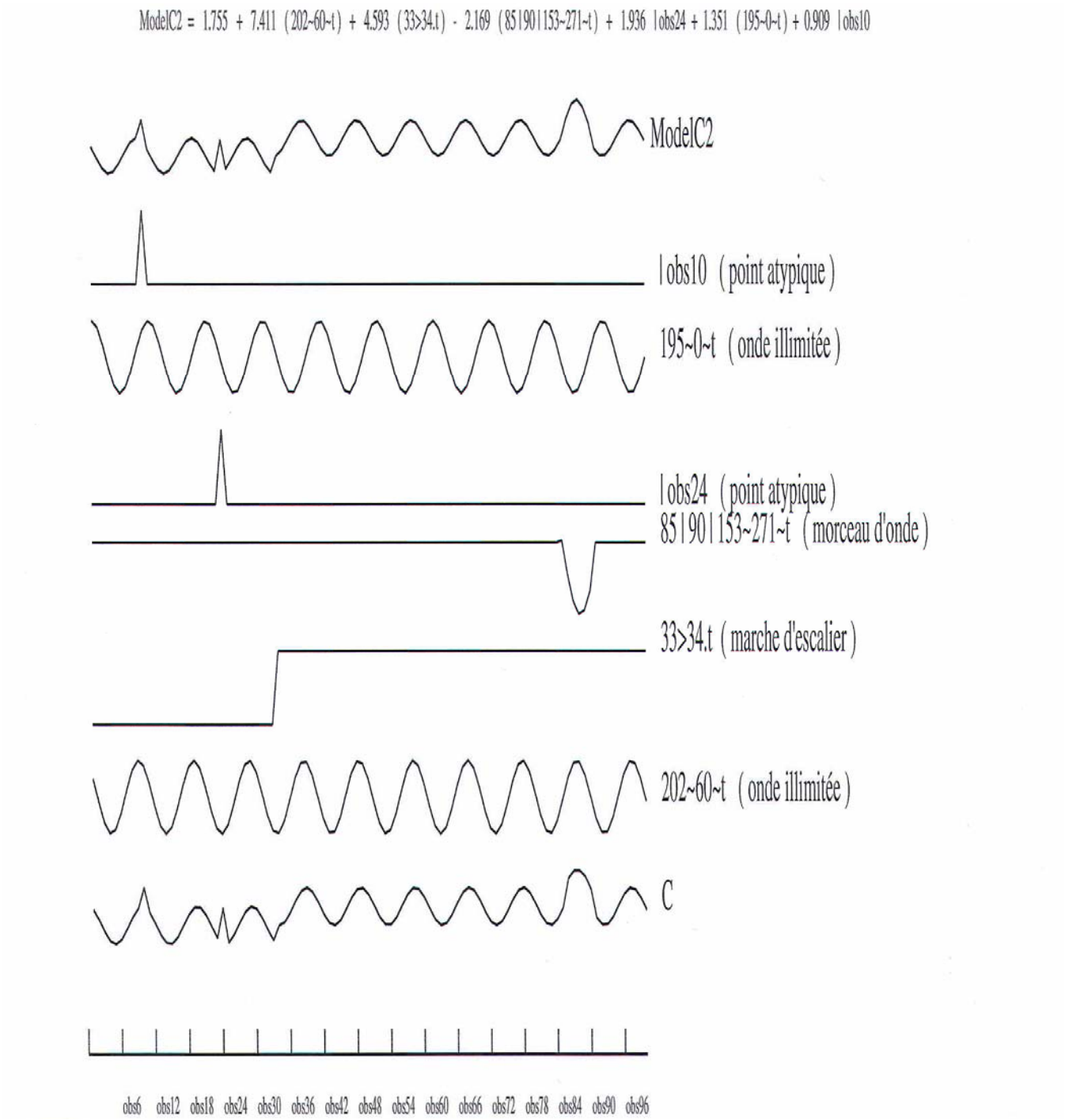


Fig.2e : Extrapolation vers l'avenir

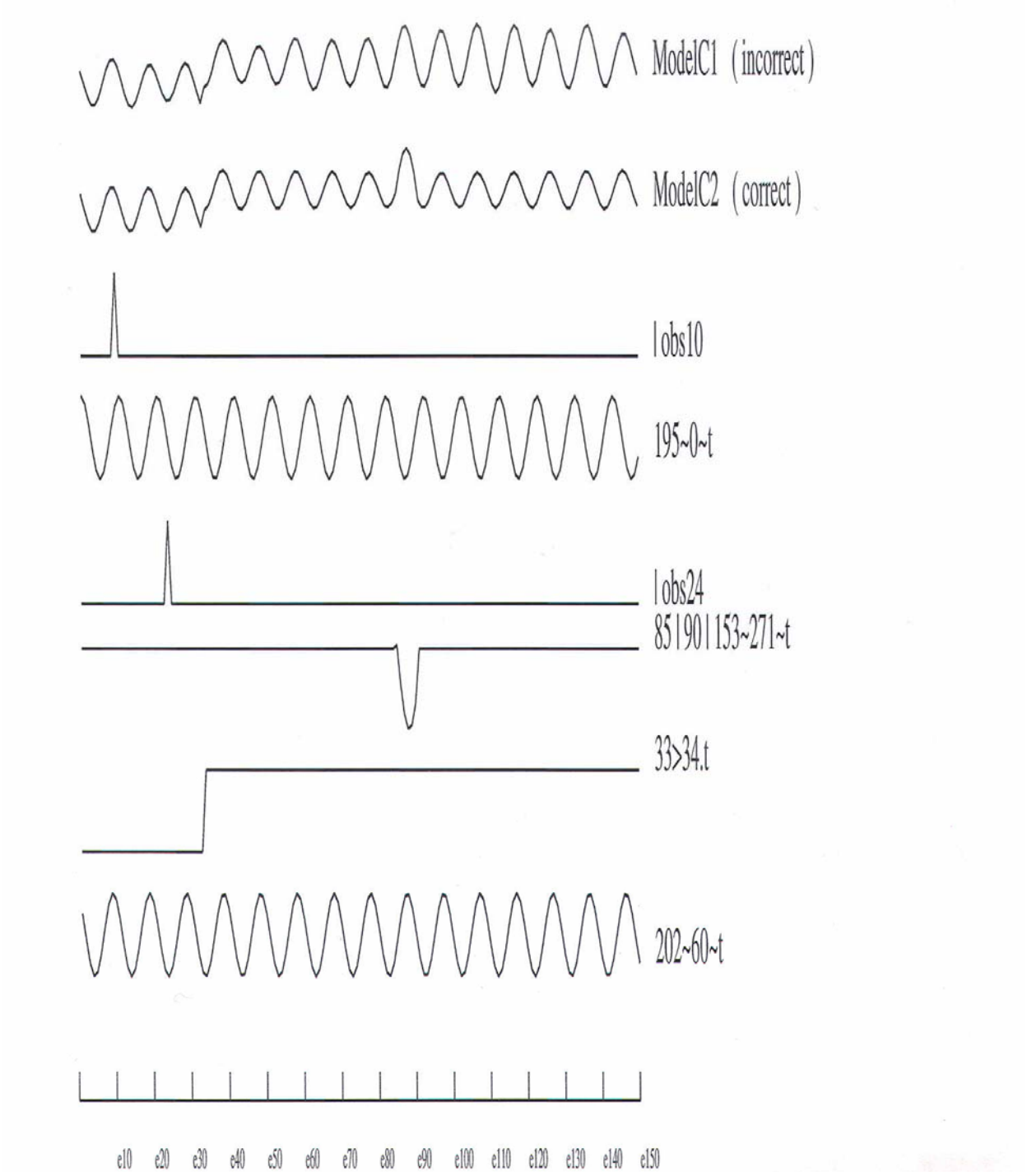


Fig.2g : Processus autorégressif d'ordre 2, et bruit blanc

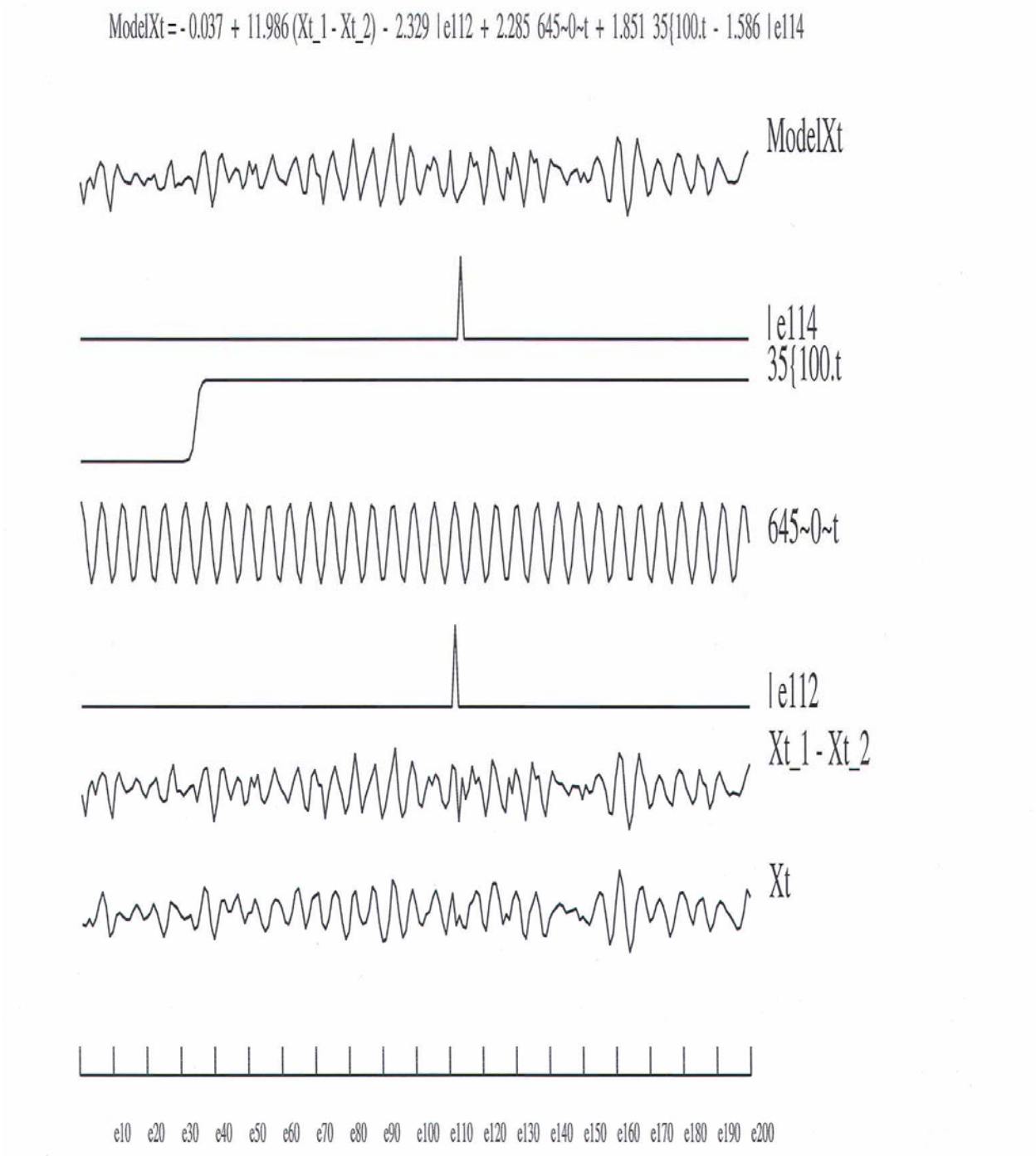


Fig.2h : Motifs périodiques

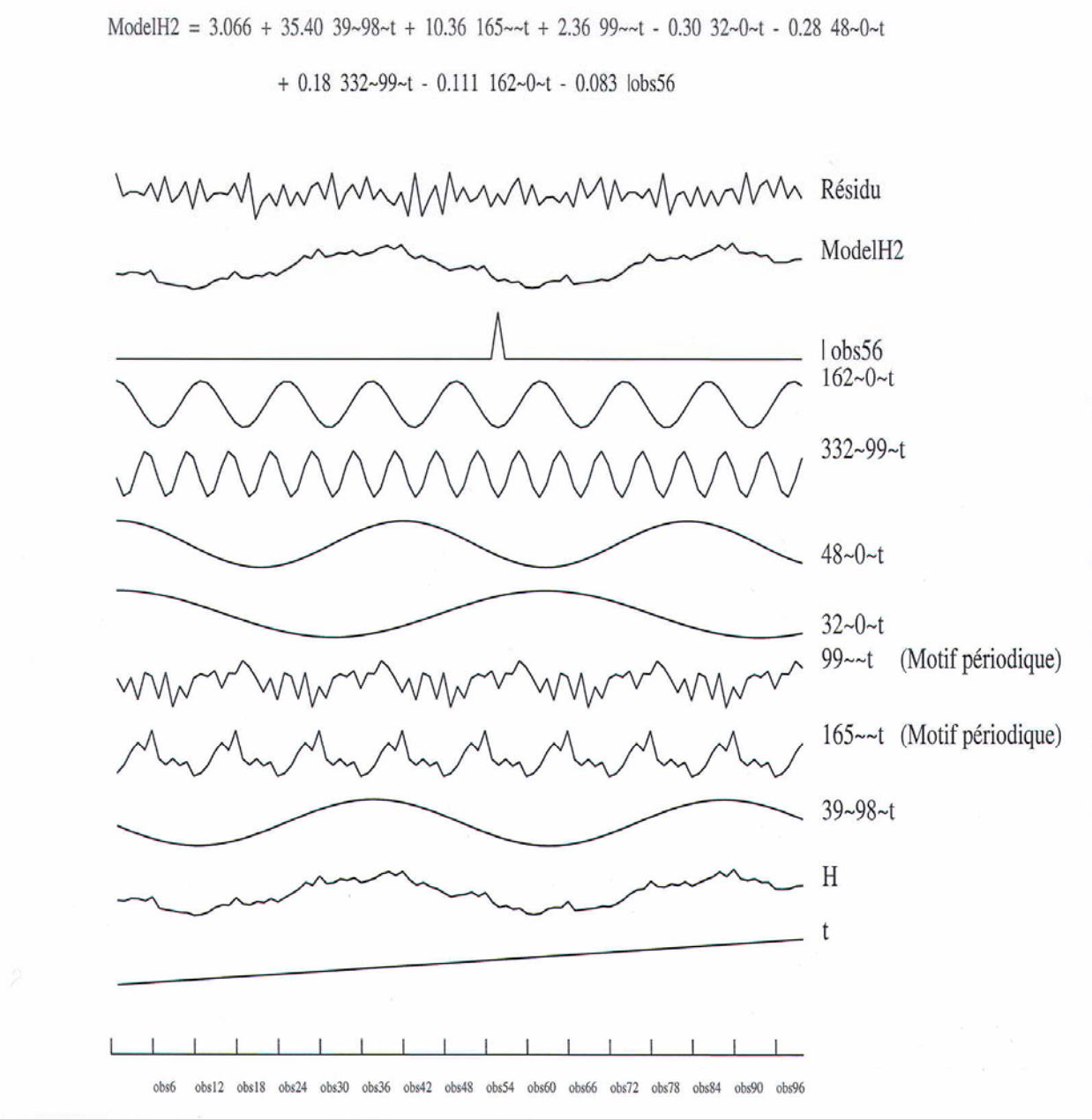


Fig.3 : Trafic SNCF, décomposition harmonique

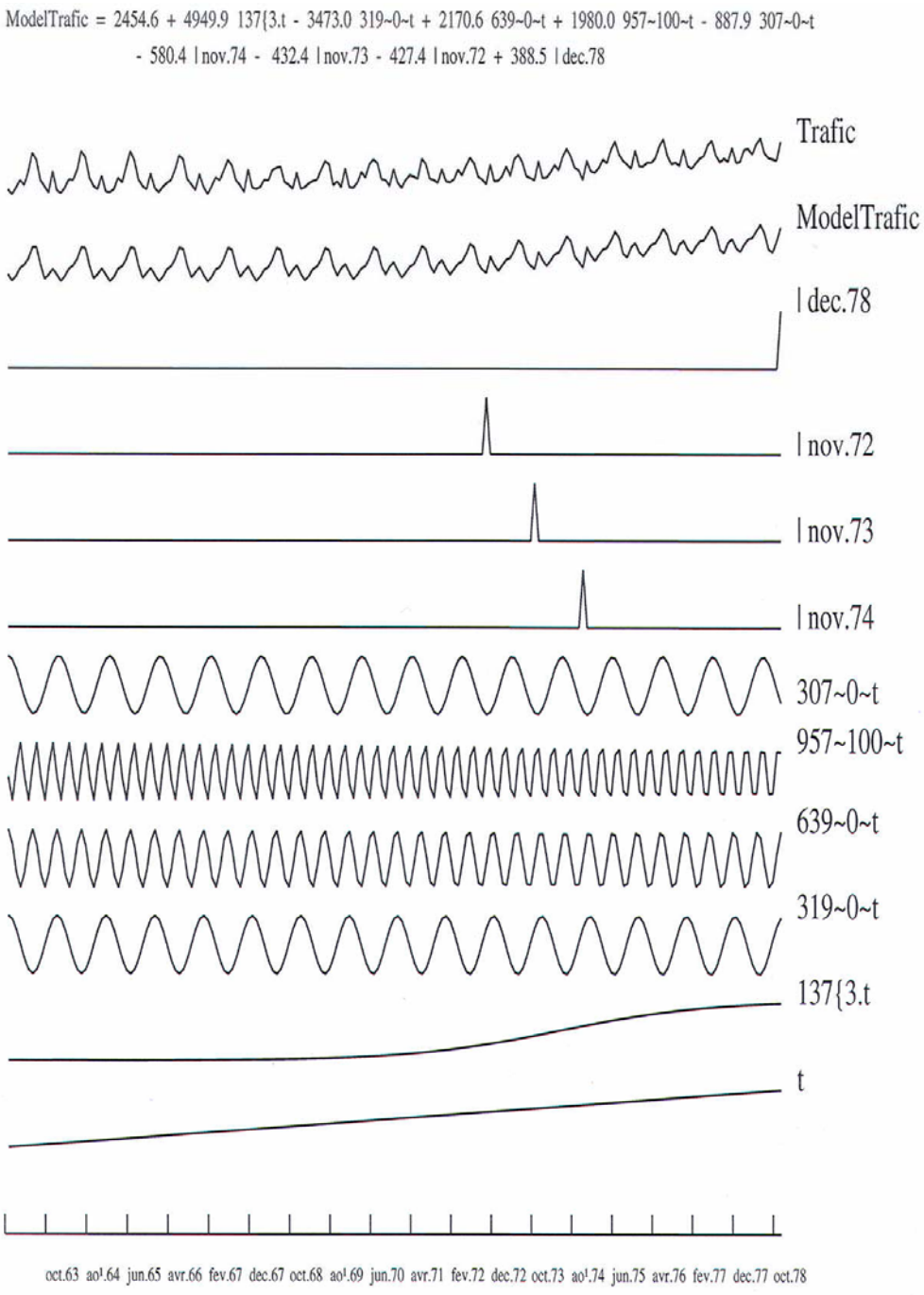


Fig.5 : Etoile RT Cygni du 8/12/88 au 11/4/93

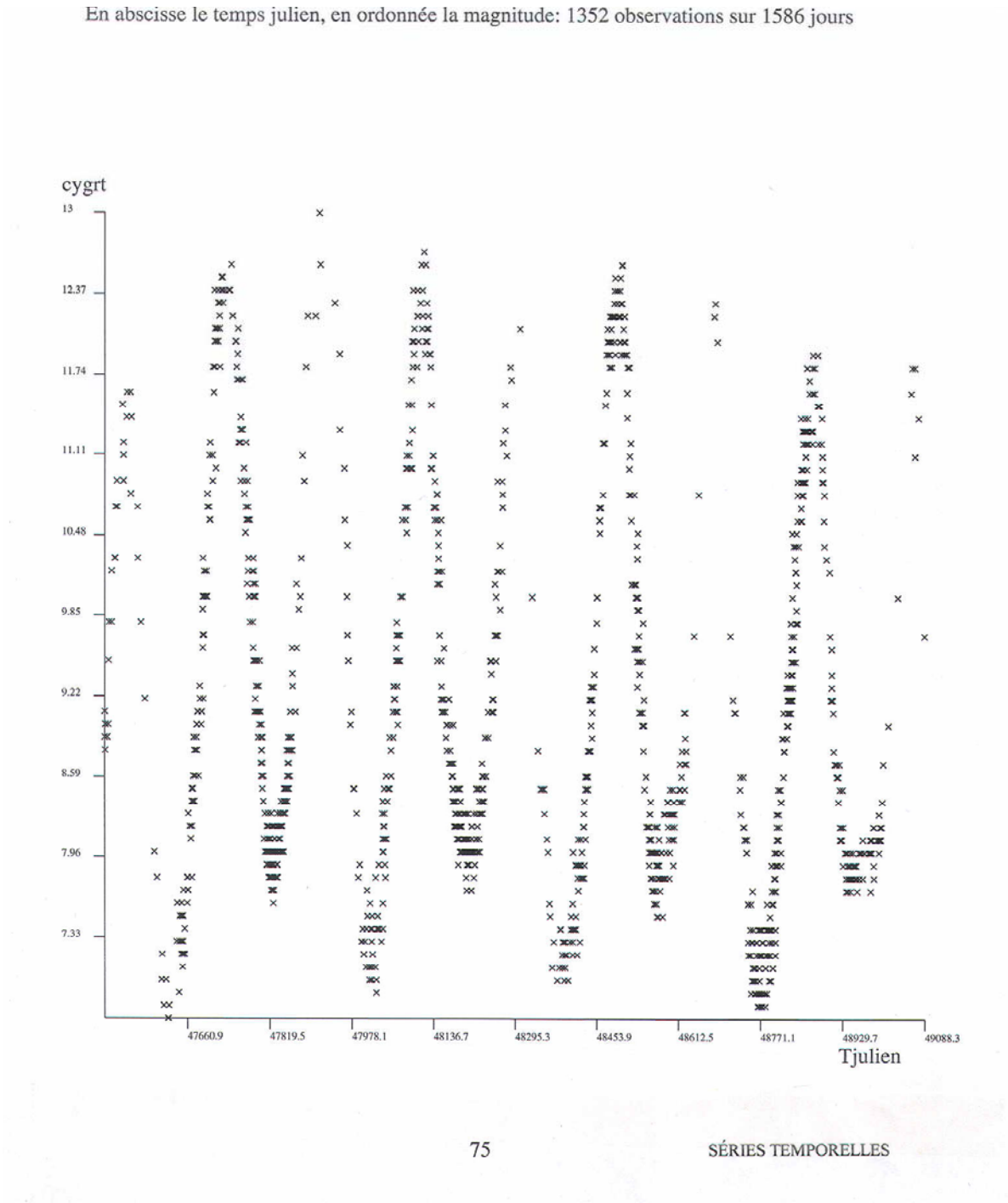


Fig.6 : Etoile RT Cygni du 8/12/88 au 11/4/93

Modèle et composantes à partir des observations

(intervalles de temps irréguliers).

ModelCygrt = 9.10 + 52.91 168~275~Tjulien + 10.27 78~98~Tjulien - 7.02 335~0~Tjulien
- 6.22 876>914.Tjulien - 5.61 41~0~Tjulien + 4.59 162~0~Tjulien
+ 3.31 406~90~Tjulien + 3.13 212~89~Tjulien + 3.06 60>66.Tjulien

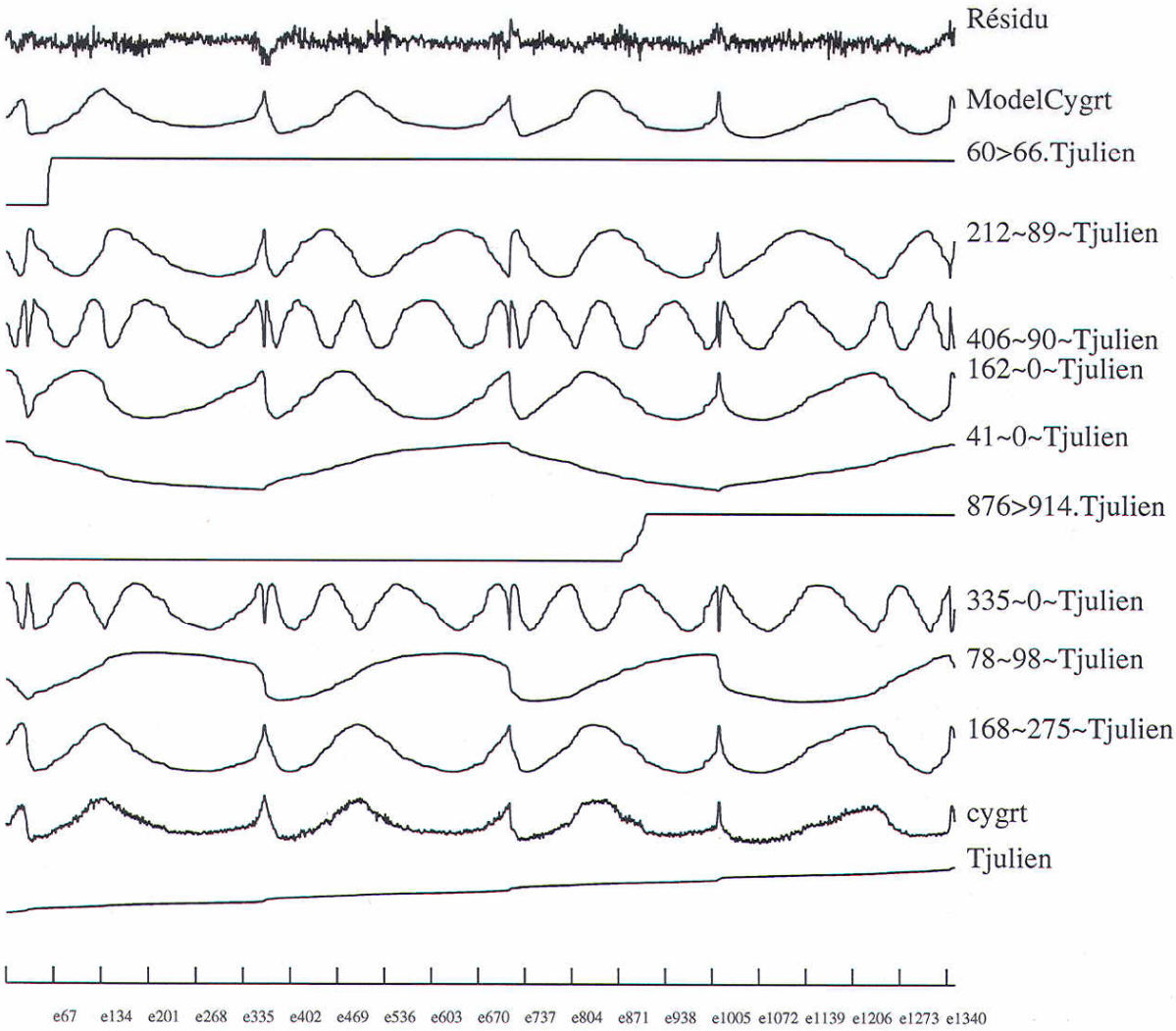
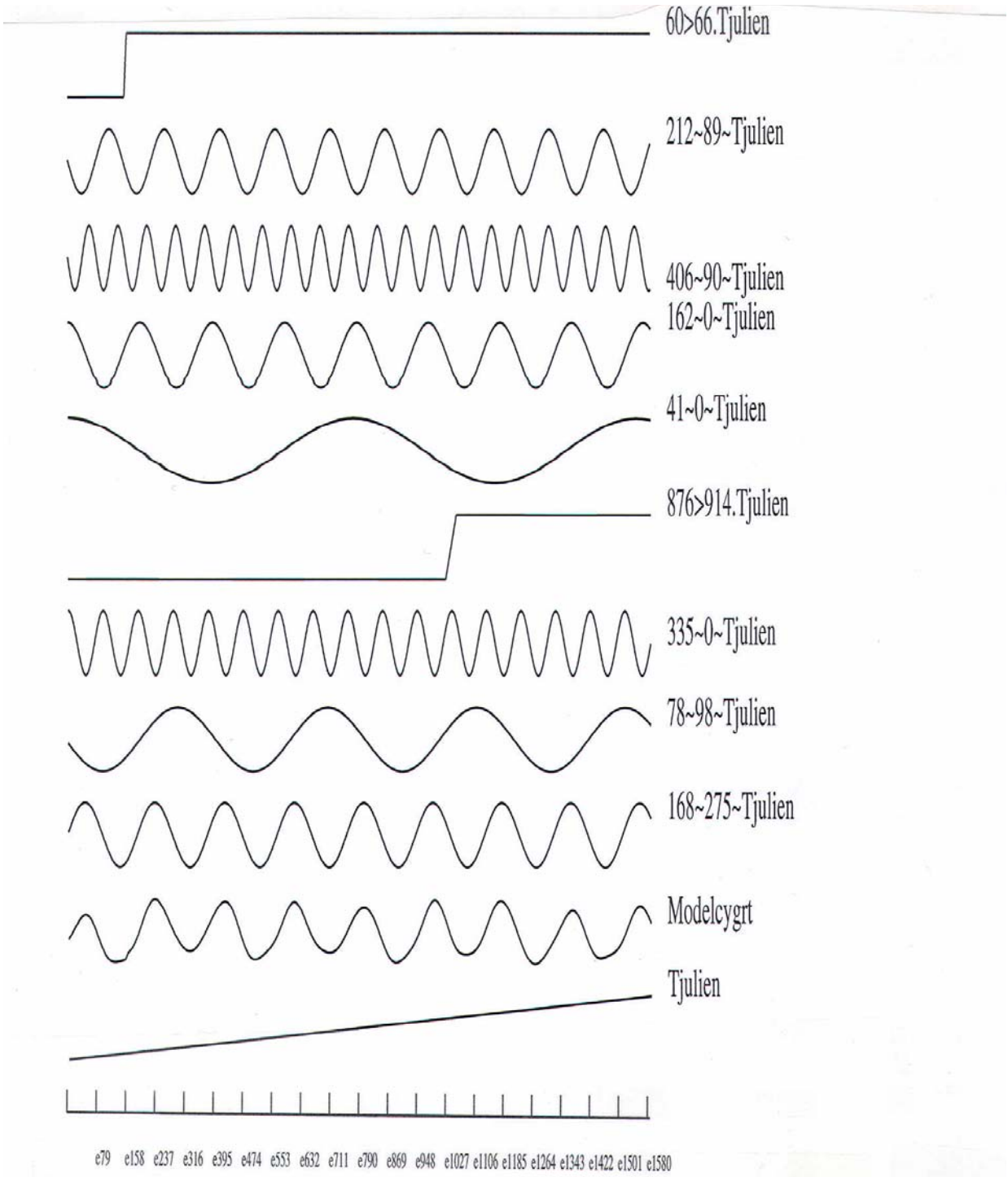


Fig.7 : Interpolation, étoile RT Cygni, modèle et composantes du 8/12/88 au



11/4/93