Séries temporelles chronobiologiques : qualité des données, modélisation

Laurent Gouthière [1], Benoît Mauvieux [2]

1- Laboratoire de Statistiques Appliquées et d'Informatique Biomédicale, Expert Soft Tech. , 7 chemin de la Birotte, F-37320 Esvres, France

2 - Laboratoire du Centre de Recherches en Activités Physiques et Sportives (CRAPS) - UPRES EA 2131 - Université de Caen, 2, Boulevard Maréchal Juin, F-14032 Caen Cedex, France

Correspondance : Laurent Gouthière, Tél. +33 247 582 641, l.gouthiere@euroestech.net, http://www.euroestech.fr

Résumé :

La problématique qui se pose au sujet de l'analyse des rythmes a fait l'objet de réflexions chez les méthodologistes. Seulement beaucoup de questions sont restées sans réponses. Nous proposons à travers cet exposé une méthodologie qui permet de définir les étapes qui nous paraissent essentielles dans l'analyse scientifique des rythmes.

La qualité des données est une notion nouvelle qui est présente depuis longtemps dans l'industrie et qui semble essentielle dans toute expérimentation scientifique. Ainsi l'expérimentateur peut juger du degré d'exploitabilité de ses échantillons de données et par delà du degré de validité des résultats d'exploitation. Nous proposons de donner quelques méthodes.

La recherche des périodes est aussi l'objet d'une grande problématique. On dispose pour cela de différentes méthodes mais il faut pouvoir déterminer celle qui est la plus adaptée et la plus fiable comme nous le montre la diversité des résultats dans les publications scientifiques. Nous proposons des méthodes spectrales dont deux nouvelles issues de la régression et complémentaires à la méthodologie « Cosinor ».

La modélisation par contre utilise différents modèles. Nous nous intéresserons aux modèles issus de la régression (le modèle cosinus) et plus particulièrement aux tests complémentaires qui permettent de juger d'une bonne modélisation.

Introduction :

Dans le domaine de l'analyse statistique des rythmes différentes écoles se côtoient et on notera plus particulièrement les travaux de l'école Belge et Américaine avec des auteurs comme Jean de Prins (1, 2) ou Germaine Cornélissen (1, 3), les travaux de l'école Allemande dans ce domaine sont aussi remarquables.

La problématique dans l'analyse statistique des rythmes porte à ce jour principalement sur les points suivants.

- L'absence d'une méthodologie nous permet d'observer une grande diversité dans les publications scientifiques.
- Le choix du modèle périodique. Nous rappellerons que le modèle sinusoïdal (cosinusoïdal) a été adopté à cause de sa simplicité et parce que qu'il semble toujours être le mieux adapté. Mais cependant il peut présenter différentes variations suivant que le modèle est mono-rythmique ou non et possède une amplitude ou une phase complexe fonction du temps.
- La recherche de la période est toujours la deuxième grande problématique et jusque là peu de méthodes spectrales ou non semblent dignes d'intérêt. Nous proposons ici des méthodes spectrales dont deux, particulièrement fiables, sont issues de la régression.

Notre méthodologie est une alternative à l'analyse statistique Bayesienne et classique : l'Exploratory Data Analysis (EDA). Un grand nombre des graphiques que nous présenterons ici sont issus de l'EDA (4)

1. Méthodologie d'analyse des données dans le cadre de l'étude de rythmicité

Le premier point essentiel de cette méthodologie est la normalisation. En effet une méthodologie normalisée permet d'adopter un langage commun. L'EDA est une méthode Normalisée par le US National Institute of Standard et Technology (Il est difficile pour le moment de trouver un organisme Européen)

Nous avons choisi des méthodes issues de l'EDA principalement pour les raisons suivantes : L'EDA parmi ces concepts n'impose pas un type particulier de modélisation, l'analyse des données précède la modélisation, les méthodes d'analyses sont principalement graphiques et donc beaucoup plus « parlantes » pour le Chronobiologiste.

Des tests complémentaires performants viennent compléter les études graphiques.

2. Etapes essentielles dans l'analyse des données

(Suite page 100)

(Suite de la page 99)

2.1 Qualité des données

La notion de qualité est une notion issue de l'industrie. Elle a été créée afin d'améliorer la productivité. Elle est déjà présente dans la recherche scientifique sous d'autres formes et nous avons cherché à introduire cette notion au niveau de l'échantillonnage dans le cadre de l'analyse de données en Chronobiologie.

La qualité permet ainsi d'apporter suffisamment de critères pour juger si un échantillon de données d'un phénomène étudié est susceptible d'être exploitable et de présumer de même des résultats.

Pour cela on dispose de différents outils graphiques comme :

- Le diagramme de décalage (« Lag Plot »)
- Le diagramme d'autocorrélation (5)
- Le graphe de probabilité Normale (6)

Les conditions nécessaires à une bonne qualité des données sont les suivantes :

- Absence de répartition aléatoire des données qui traduirait effectivement l'étude d'un phénomène sans intérêt ou même l'absence de phénomène. Le diagramme d'autocorrélation (5) permet de mettre en évidence un caractère aléatoire (voir figures 2.1-a, 2.1-b)
- Indépendance des données. En effet le caractère dépendant des données introduit une relation supplémentaire dans les données et risque de fausser la modélisation et de nombreux tests statistiques. On peut employer dans ce cas le diagramme de décalage (voir figures 2.1-c, 2.1-d) complété par un test de Q Ljung Box (7) Cependant certains auteurs nous font remarquer

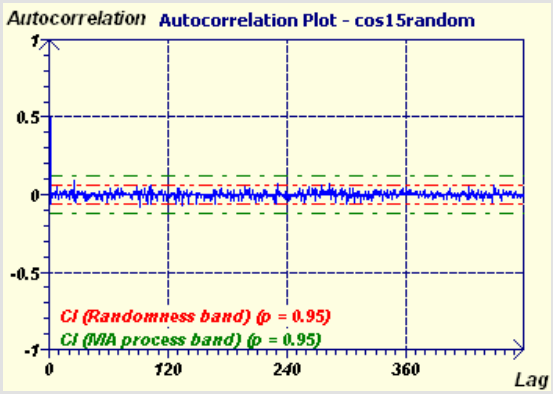


Figure 2.1-a : *Courbe d'autocorrélation : Le caractère aléatoire est marqué lorsque la courbe se rapproche de zéro (Fonction cosinus aléatoire de période 15)*

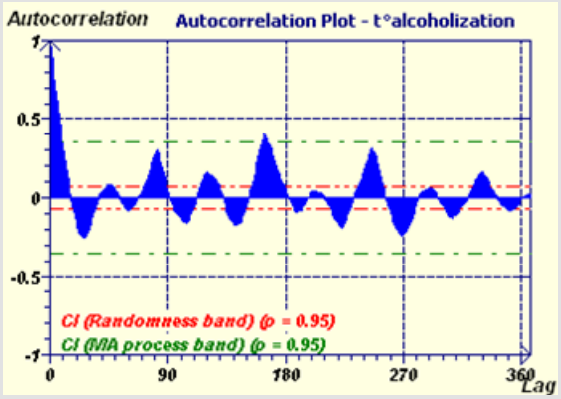


Figure 2.1-b : *Courbe d'autocorrélation : Caractère aléatoire non marqué. On pourra noter l'aspect sinusoïdal de la courbe (Etude de la température de sujets alcoolisés)*

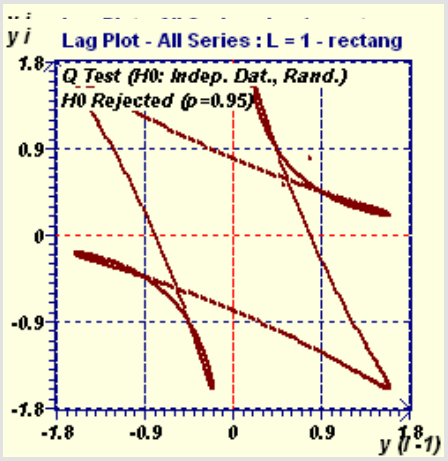


Figure 2.1-c : *« Lag plot » et Q test : Données fortement dépendantes (Etude d'un signal rectangulaire)*

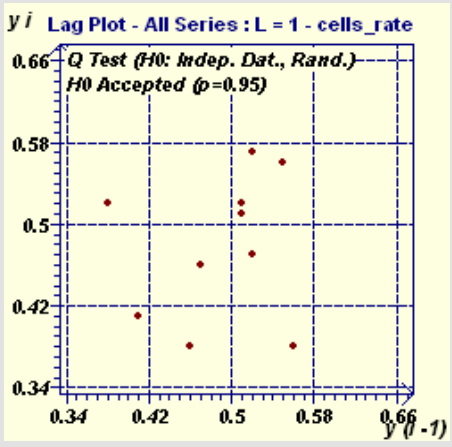


Figure 2.1-d : *« Lag plot » et Q test. Données indépendantes (Etude de la vitesse de prolifération de fibroblastes)*

(Suite page 101)

(Suite de la page 100)

que les données provenant de séries temporelles sont souvent fortement corrélées (5)

- La non existence d'un caractère stationnaire limitant d'une manière marquée l'intérêt d'étude de périodicité. Un calcul complémentaire du PACF (Partial AutoCorrelation Function) permet de tester cette hypothèse. L'emploi du diagramme d'autocorrélation et un certain type d'analyse spectrale (selon Blochner) permet de détecter ce type de caractère (voir figures 2.1-e, 2.1-f)

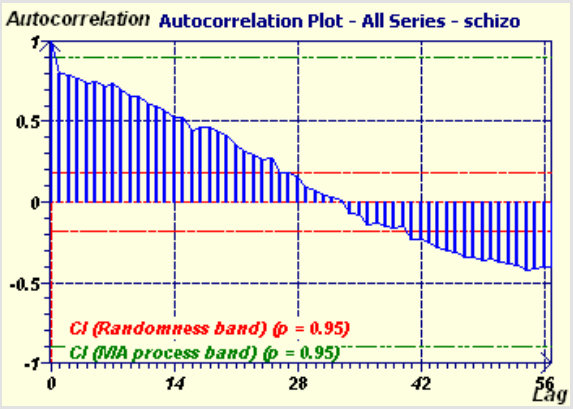


Figure 2.1-e : *Courbe d'autocorrélation. La présence d'un phénomène de type stationnaire est probable (Etude de l'activité intellectuelle d'un patient schizophrène : L'activité a été ralentie au bout de 60 jours par la Chlopromazine)*

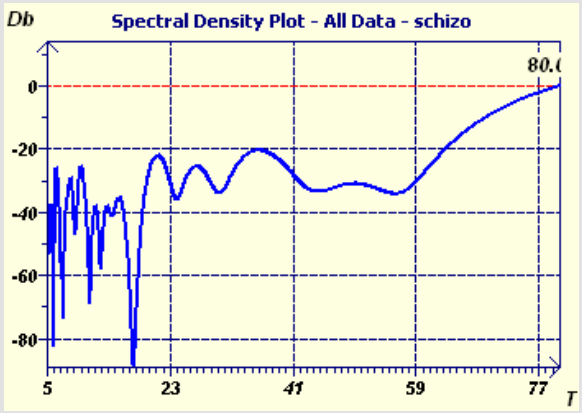


Figure 2.1-f : *Analyse spectrale selon Blochner. On peut observer plusieurs pics avant 60 jours. Ce type de spectre est un complément à l'analyse des processus « auto-régressifs » et stationnaires (Même étude que pour la figure 2.1-e)*

- La répartition normale des données est souhaitable (voir figures 2.1-g, 2.1-h) On emploie le graphe de probabilité Normale (6) Un test complémentaire de Kolmogorov-Smirnov (K-S test) permet de confirmer ou d'infirmer cette hypothèse. Les données d'un phénomène biologique présentent souvent ce caractère de Normalité. D'autre part de nombreux tests statistiques et d'analyses (Par exemple le Périodogramme de Fisher) nécessitent cette hypothèse.

Les conditions présentées précédemment présument de l'échantillonnage et sont des critères de mesure de la qualité de celui-ci. Si la plupart des conditions précédentes sont réunies l'investigation sur les données permettra d'aboutir à des résultats caractérisant le phénomène étudié. Dans le cas contraire il est nécessaire de procéder à un nouvel échantillonnage.

2.2. Recherche de la périodicité

Le choix de la méthode spectrale de recherche de périodicité peut être effectué selon le critère équi-réparti ou non équi-réparti dans le temps des données de l'échantillon.

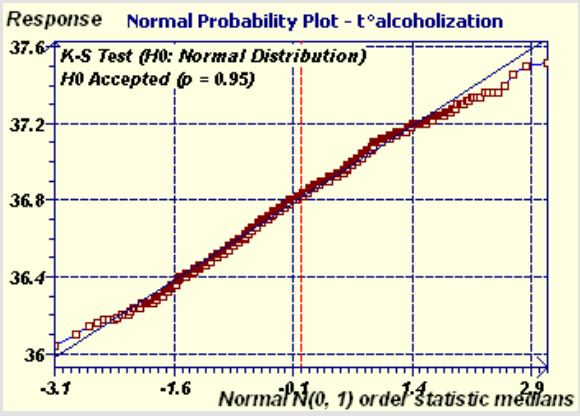


Figure 2.1-g : *Graphe de Normalité. Répartition des données selon une distribution « Normale » avec K-S test (Etude de la température de sujets alcoolisés)*

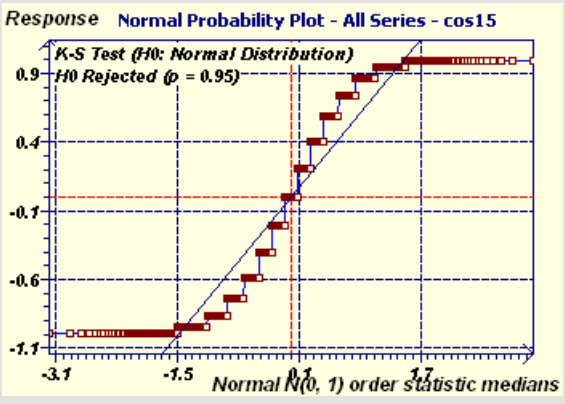


Figure 2.1-h : *Graphe de Normalité. Répartition des données selon une distribution non « Normale » avec K-S test (Etude d'une fonction étalon cosinus de période 15)*

(Suite page 102)

(Suite de la page 101)

2.2.1 Données non équi-réparties :

Si les données sont non équi-réparties, seules trois méthodes sont envisageables. Nous préconisons les méthodes suivantes qui parmi l'ensemble des méthodes utilisables présentent la plus grande fiabilité :

- **Le spectre du « percent rhythm » :** Le spectre du « Percent Rhythm » a pour principe le test d'«Amplitude nulle » (3, 8) (voir figures 2.2.1-a, 2.2.1-b)
- **Le spectre Elliptique Inverse**. : Le spectre Elliptique Inverse a pour principe le test de l' « Ellipse » (3, 8) (voir figures 2.2.1-c, 2.2.1-d)

Ces analyses spectrales existent aussi dans le cadre de recherche de périodes de population .

- **Le périodogramme de Lomb and Scargle**.: Le périodogramme de Lomb et Scargle (9) a pour principe à la fois la régression et l'analyse de Fourier (voir figures 2.2.1-e, 2.2.1-f)

2.2.2 Données équi-réparties :

Les méthodes spectrales dérivées de l'analyse de Fourier sont seulement applicables dans le cas de données équi-réparties, les méthodes dérivées de la régression restent aussi applicables dans ce cas (voir paragraphe précédent)

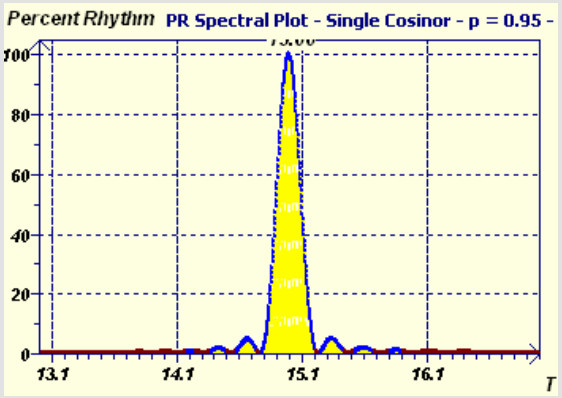


Figure 2.2.1-a : *Spectre du « Percent Rhythm » permet de détecter la période de 15. Ce spectre permet de détecter une période suivant le test d' « Amplitude nulle » (Etude d'une fonction cosinus étalon de période 15)*

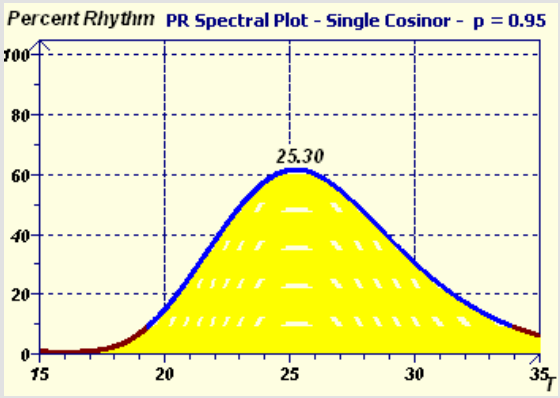


Figure 2.2.1-b : *Spectre du « Percent Rhythm » permet de détecter la période de 25,3 heures. (Etude de l'activité de Hamsters sous l'action d'antidépresseurs)*

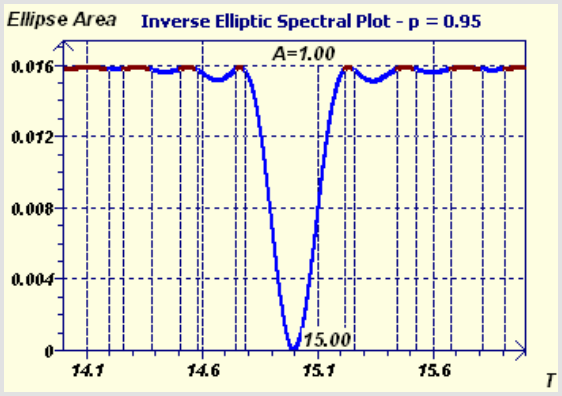


Figure 2.2.1-c : *Spectre elliptique inverse (normalisé) permet de détecter la période de 15. Ce spectre permet de détecter une période suivant le test de l' «Ellipse »*

(Etude de la fonction cosinus étalon de période 15)

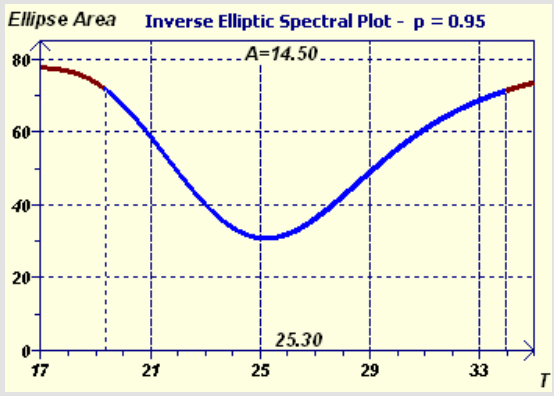


Figure 2.2.1-d : *Spectre elliptique inverse (normalisé) permet de détecter la période de 25,3 heures. En bleu l'intervalle de confiance sur la détermination de la période à la probabilité donnée (Etude de l'activité de Hamsters sous l'action d'antidépresseurs)*

Nous présenterons ici les plus performantes et leur intérêt. Elles sont issues la plupart de méthodes employées en Astronomie. Leur inconvénient est leur fiabilité moyenne c'est à dire la répétabilité moyenne dans la détermination de résultats exacts. On présentera les analyses spectrales suivantes :

- **Spectre Autospectral selon Jenkins et al** (10) : C'est un spectre provenant de la transformée de Fourier de la fonction d'autocorrélation

(Suite page 103)

(Suite de la page 102)

avec quelques modifications (voir figures 2.2.2-a, 2.2.2-b)

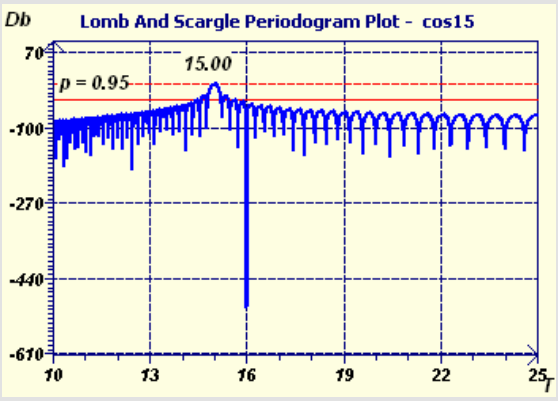


Figure 2.2.1-e : *Le périodogramme de Lomb et Scargle (normalisé) permet une détection proche des méthodes précédentes, mais avec moins de fiabilité (Etude d'une fonction cosinus étalon de période 15)*

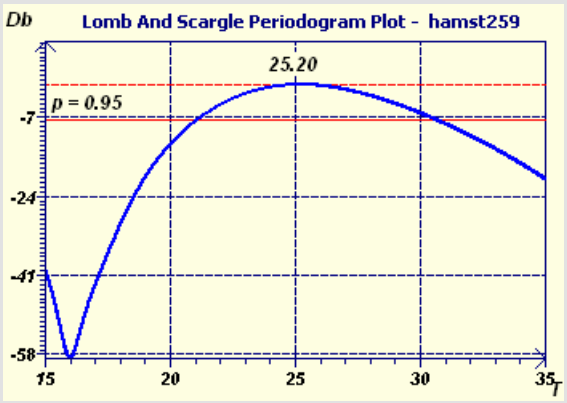


Figure 2.2.1-f : *Le périodogramme de Lomb et Scargle présente un pic principal à 25,2 heures (Etude de l'activité de Hamsters sous l'action d'antidépresseurs)*

- **Autopériodogramme selon Jenkins et al** (10) : Le principe de ce spectre est le même que le précédent avec d'autres modifications (voir figures 2.2.2-c, 2.2.2-d)

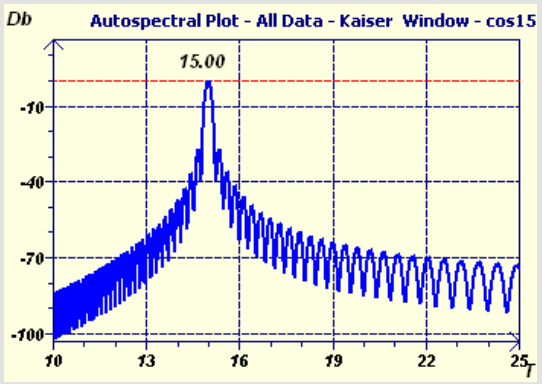


Figure 2.2.2-a : *Spectre Autospectral (normalisé) selon Jenkins et al (Fenêtrage de Kaiser) On arrive à une approche de la période exacte (Etude d'une fonction cosinus étalon de période 15)*

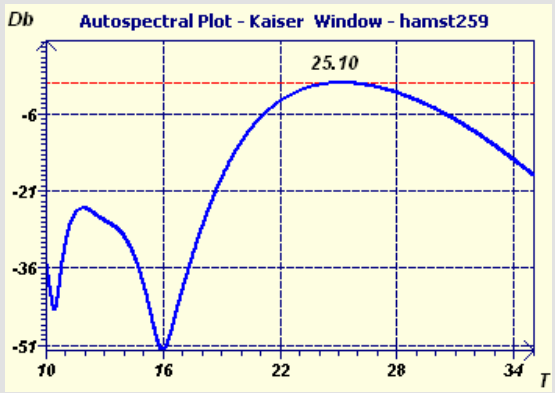


Figure 2.2.2-b : *Spectre Autospectral (normalisé) selon Jenkins et al. On observe un pic remarquable à 25,1 heures (Etude de l'activité de Hamsters soumis à des antidépresseurs)*

- **Périodogramme de Fisher** : Le périodogramme de Fisher est employé en Chronobiologie. C'est une méthode qui n'est pas contestée, mais pourtant qui présente, à notre avis, beaucoup moins d'intérêt que les méthodes déjà précédemment citées (voir figures 2.2.2-k, 2.2.2-l).

2.3 Modèle et validation statistique du modèle

2.3.1 Modélisation par régression

Dans le cadre de l'étude d'un modèle mono-rythmique, l'emploi de la fonction cosinus classique calculée par régression semble le mieux adapté (figures 2.3.1-a, 2.3.1-b) Cependant il faut vérifier que l'amplitude et la phase restent constantes. En effet les phénomènes biologiques mono-rythmiques ne semblent pas entièrement suivre ce type de modèle. On observe souvent une expression de la phase plus complexe (ou de l'amplitude) L'emploi de spectres de démodulation complexes permet de vérifier si le modèle classique (y(t) = aCos(2Pit/T + Phi) + M n'est pas une fonction complexe du temps (figures 2.3.1-c, 2.3.1-d)

2.3.2 Validation statistique du modèle

Le modèle calculé par régression doit absolument vérifier les hypothèses suivantes. Dans le cas contraire l'étude statistique peut être fortement remise en cause.

- Indépendance des résidus (voir figure 2.3.2-e)

(Suite page 104)

(Suite de la page 103)

la période, les méthodes spectrales issues de la

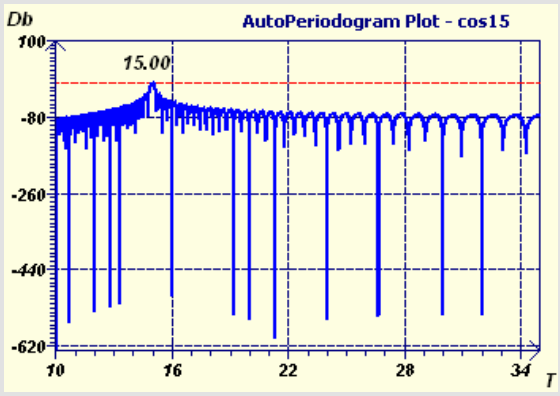


Figure 2.2.2-c : *Autopériodogramme (normalisé) selon Jenkins et al. Un pic remarquable à 15 (Etude d'une fonction cosinus étalon de période 15)*

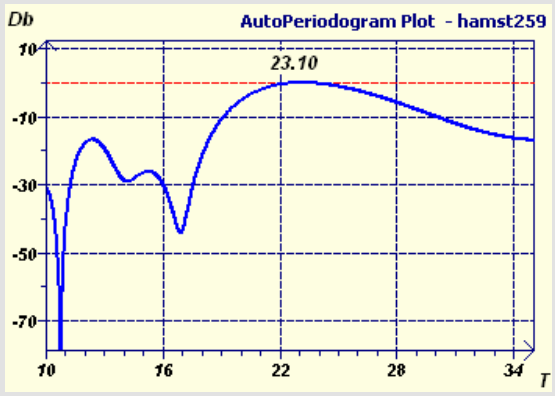


Figure 2.2.2-d : *Autopériodogramme (normalisé) selon Jenkins et al. On observe un pic remarquable à 23,1 heures (Etude de l'activité de Hamsters soumis à l'action d'antidépresseurs)*

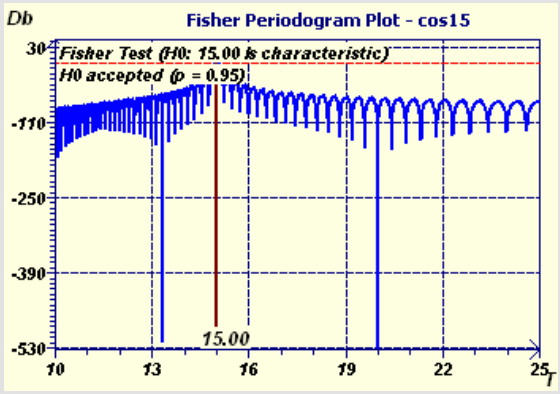


Figure 2.2.2-e : *Périodogramme de Fisher (normalisé) C'est une méthode directement issue des DFT (Discrete Fourier Transforms) avec un test statistique sur la fondamentale à 15 dans cet exemp*le (Etude d'une fonction cosinus étalon de période 15)

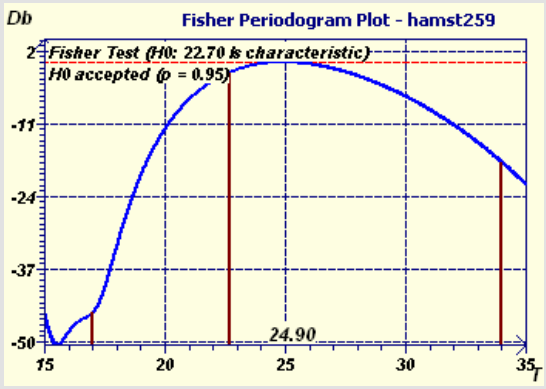


Figure 2.2.2-f : *Périodogramme de Fisher (normalisé) La fondamentale testée est à 22,7 heures (Etude de l'Activité de Hamsters soumis à l'action d'antidépresseurs)*

- Moyenne des résidus nulle (voir figure 2.3.2-f)
- Normalité des résidus (voir figure 2.3.2-g)
- Homogénéité de la variance des résidus, lors de l'étude de différents groupes (voir figure 2.3.2-h).

Conclusion :

La méthodologie exposée ici à travers des graphiques d'EDA introduit la notion de qualité dans l'échantillonnage. Celle-ci pourra être complétée ultérieurement par d'autres tests complémentaires. Nous pensons plus particulièrement à la détermination de la taille minimale de l'échantillon. Ceci a déjà été abordé par certains auteurs (1) mais son application dans le cadre de la Chronobiologie semble difficile. On peut aussi essayer de déterminer la taille minimale de l'échantillon après modélisation, en étudiant le comportement des résidus.

Pour répondre à la problématique de recherche de régression nous ont montré leur plus grande fiabilité par rapport à celles provenant de l'analyse de Fourier. Il existe par exemple d'autres méthodes comme plus particulièrement le MESA (Maximum Entropy Spectral Analysis) (11), la méthode des « Concordances » (12) qui présentent aussi un intérêt certain.

Nous avons cherché à présenter l'utilisation de ces méthodes d'une manière la plus abordable possible pour le Chronobiologiste. Le lecteur trouvera dans la consultation de la bibliographie le complément théorique nécessaire.

Note : *Les graphiques sont issus du logiciel Time Series Analysis Seriel Cosinor du Laboratoire de Statistiques Appliquées (http://www.euroestech.fr) Les différents échantillons de données proviennent d'études que nous avons effectuées et qui ont donné lieu pour certaines à des publications (13) et d'autres qui sont en cours.*

(Suite page 105)

(Suite de la page 104)

Références :

(1) de Prins J, Cornélissen G et Malbecq W, Statistical Procedures in Chronobiology and Chronopharmacology, Annual Review of Chronopharmocology 1986; Vol. 2: pp. 27-141.

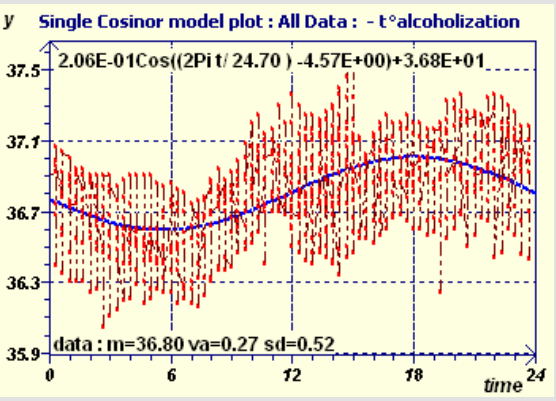


Figure 2.3.1-a : *Modèle de période 24,7 heures calculé par régression « cosinus » y(t) = aCos(2Pit/T + Phi) + M (Etude de la température de sujets alcoolisés)*

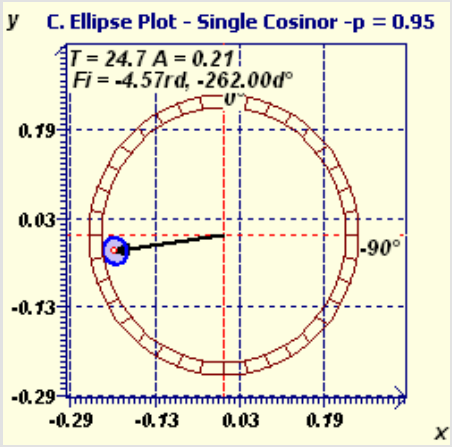


Figure 2.3.1-b : *Ellipse de confiance selon le « Single Cosinor » (14) La période est de 24,7 heures. Plus la surface de l'ellipse est faible plus la précision sur la détermination de la période est importante (Etude de la température de sujets alcoolisés)*

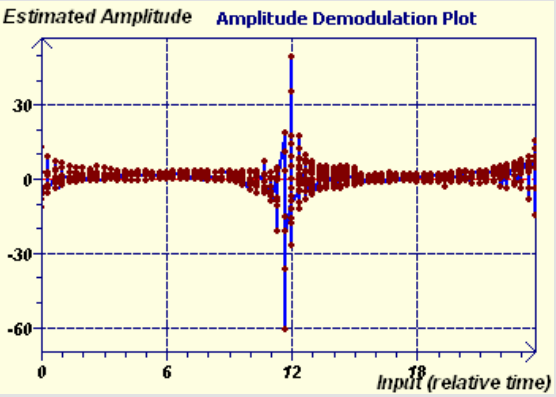


Figure 2.3.1-c : *Spectre de démodulation complexe en amplitude. L'amplitude pour une période de 24,7 heures reste suffisamment constante pour que l'on puisse affirmer qu'elle ne soit pas une fonction complexe du temps (Etude de la température de sujets alcoolisés)*

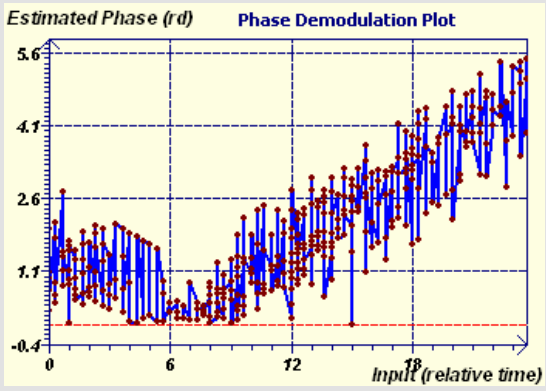


Figure 2.3.1-d : *Spectre de démodulation complexe en phase. La phase ne reste pas constante au cours du temps. Elle croit durant la phase d'éveil, décroît pendant la phase de sommeil et se stabilise lors du réveil. Dans ce cas la phase peut être une fonction complexe du temps et le modèle s'écrire y(t) = aCos(2Pit/T + Phi(t)) + M (Etude de la température de sujets alcoolisés)*

(2) de Prins J, Comment Définir un Rythme par une Méthodologie Scientifique, Les Rythmes Lectures et théories, sous la direction de J.J. Wunenburger, Centre culturel international de Cerisy, "Conversciences", L'Harmattan, ISBN 2-7384-1355-2, 1991, pp. 57-65,.

(3) Cornélissen G, Halberg F, Stebbings J, Halberg E, Carandente F, Hsi B, Chronobiometry: with pocket calculators and computer systems, La Ricerca Clin. Lab., 1980, 10: pp. 333-385.

(4) Tukey J, Exploratory Data Analysis, Addison-Wesley, 1977.

(5) Box GEP, Jenkins GM, and Reinsel GC, Time Series Analysis: Forecasting and Control, Third edition, Prentice Hall, 1994.

(6) Chambers J, Cleveland W, Kleiner B, and Tukey P, Graphical Methods for Data Analysis, Wadsworth, 1983.

(7) Ljung, GM and Box GEP. , On a measure of lack of fit in time series models. Biometrika 1978, 65: pp. 553-564.

(8) Gaudeau C, Gouthière L, Méthodologie d´analyse des rythmes dans les systèmes non linéaires, Les rythmes: Lectures et Théories, Centre Culturel International de Cerisy, "Conversciences", L´Harmattan Paris ISBN 2-7384-1355-2, 1991, pp. 31-56.

(9) Scargle JD, Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data, AA (NASA, Ames Research Center, Space Science Div., Moffett Field, CA), Astrophysical Journal 1982, Part 1, vol. 263, Dec. 15 pp. 835-853.

(Suite page 106)

(Suite de la page 105)

(10) Jenkins GM and Watts, Spectral Analysis and Its Applications, Holden-Day, 1968.

(11) Burg JP, Maximum entropy spectral analysis. Ph. D. thesis, Dep. Geophysics Stanford Univ., Stanford, CA, 1975.

(12) Hassnaoui M, Pupier R and M. Rehailia M, A Concordance Method For Analyzing Categorical Time Series. An Application For The Search of Periodicities, Biological Rhythm Research 2000, Vol.31, No.2, pp. 177-201.

(13) Mauvieux B, Gouthière L, Sesboüe B et Davenne D, Etudes Comparées Des Rythmes Circadiens De La Température et Reflet Actimétrique Du Sommeil De Sportifs Et Sédentaires En Poste Régulier De Nuit, Canadian Journal Applied of Physiology, 28(6): 831-887, 6 Décembre 2003.

(14) Nelson W, Tong YL, Lee JK, Halberg F, Methods for cosinor-rhythmometry, Chronobiologia;6:305-323, 1979.

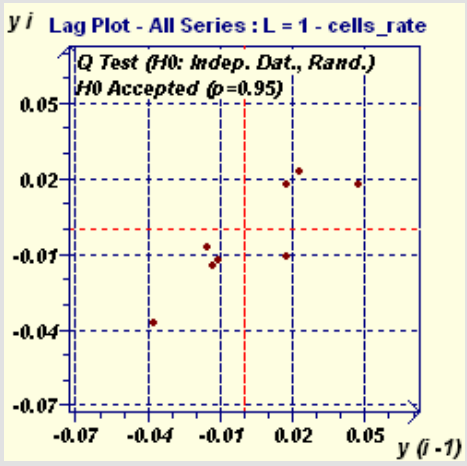
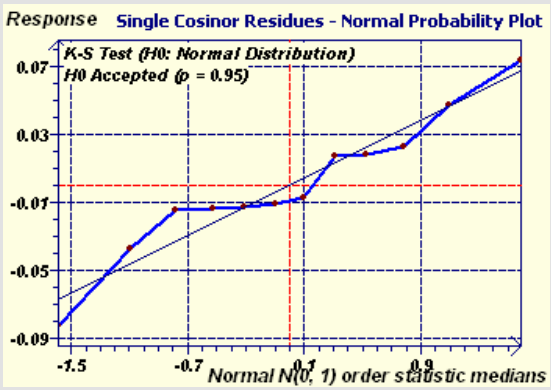


Figure 2.3.2-e : *« Lag plot » avec Q test sur les résidus provenant d'un modèle de période d'ordre 1 de 43,5 jours (Etude de la vitesse de prolifération de fibroblastes)*

2.4. Residuals distribution - Goodness of fit :

- **Ajusted r²** : -3.75E-01
- **Residual Sums of Squares** : 1.79E-02
- **κ² Test (H0: Normal residuals distribution) H0 accepted** : 0.9500
- **K-S Test (H0: Normal residuals distribution) H0 accepted** : 0.9500
- **Average Test (H0: RS Average = 0) H0 accepted** : 0.9500
- **Q Test (H0: Independent Residues) H0 accepted** : 0.9500
- K-S Test is the Kolmogorov and Smirnov test
- Average Test (H0: RS Average = 0) is a test of average on the average sum of residues
- Q Test is the Ljung-Box Q-statistic lack-of-fit hypothesis test

Figure 2.3.2-f : *Tests de « goodness of fit » sur un modèle de période 43,5 jours (Etude de la vitesse de prolifération de fibroblastes)*



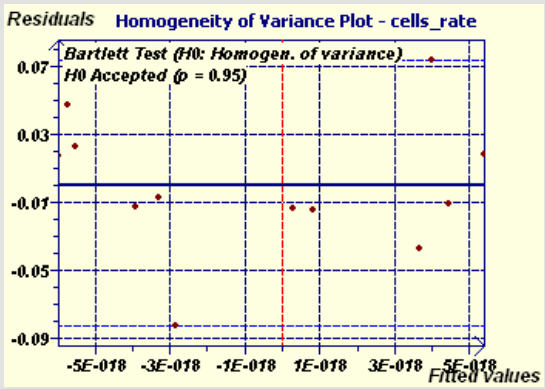


Figure 2.3.2-g : *Graphe de probabilité Normale et K-S test (p =0,95, alpha =1-p) sur les résidus d'un modèle de période 43,5 jours. Le test démontre ici la Normalité de la répartition des résidus pour p = 0,95. (Etude de la vitesse de prolifération de fibroblastes)*

Figure 2.3.2-h : *Graphe d'homogénéité de la variance et test de Bartlett des résidus (p =0,95, alpha =1-p) Le test démontre ici une bonne homogénéité de la variance pour p = 0,95 (Etude de la vitesse de prolifération de fibroblastes)*

Commémoration

Commémoration du centenaire de la naissance d'Erwin Bünning le 23 janvier 2006, Université de Tübingen D-72070 Tübingen. Programme non encore précisé, comportant notamment l'exposé de l'hypothèse de Bünning pour l'interprétation du photopériodisme.

Contact : Wolfgang Engelmann Email : engelmann@uni-tuebingen.de