Avril 2003 **Numéro 31**

Cahiers de l'IMA

Nombre de données des séries temporelles : Réduire pour mieux analyser ?

André Berchtold

Institut de Mathématiques Appliquées
Faculté des S.S.P.
Université de Lausanne
BFSH 2
CH-1015 Lausanne

Nombre de données des séries temporelles : Réduire pour mieux analyser ?

André Berchtold

Université de Lausanne
Institut de Mathématiques Appliquées
SSP, BFSH 2
CH-1015 LAUSANNE
Suisse

Andre.Berchtold@imaa.unil.ch
Andre@AndreBerchtold.com
http ://www.AndreBerchtold.com/

RÉSUMÉ. *Dans de nombreux domaines, il est maintenant possible de travailler avec de très longues séries temporelles pouvant aller jusqu'à plusieurs centaines de milliers d'observations. Si cet état de fait peut paraître a priori réjouissant, il comporte cependant un revers. Des algorithmes de plus en plus complexes étant utilisés en fouille de données, le coût en terme de temps de calcul devient rapidement prohibitif. La question se pose dès lors de savoir s'il n'est pas possible d'appliquer les mêmes méthodes et d'obtenir des résultats similaires en n'utilisant qu'une partie des données à disposition. Nous apportons ici des éléments de réponse par le biais d'expériences numériques réalisées sur des données réelles.*

ABSTRACT. *In various fields, it is now possible to work with very long time series up to several hundreds of thousand observations. If this situation is a priori great, there is a major drawback. More and more complex algorithms being used in data mining, the computing costs become quickly prohibitive. The question then is to know whether it would be possible to apply the same methods and to obtain similar results using only a subset of the available data. We answer to this question by the mean of numerical experiments based on real data sets.*

MOTS-CLÉS : *Série temporelle, nombre de données, réduction, échantillonnage, distribution de fréquences, dépendance temporelle.*

KEYWORDS: *Time series, number of data, reduction, sampling, frequency distribution, time dependence.*

1. Introduction

Le concept de fouille de données s'est généralisé notamment suite à la création de bases de données de plus en plus grandes pour lesquelles des méthodes d'extraction et d'analyse statistique adéquates ont dû être développées. Cet accroissement quasi-exponentiel de la masse d'information à traiter est particulièrement visible dans un domaine tel que la finance où le cours des titres cotés en bourse peut être obtenu avec une précision de l'ordre de la minute, voire mieux. La sismologie, pour laquelle les données concernant une éruption volcanique sont fréquemment enregistrées avec une précision d'une seconde, ou encore l'étude des vitesses et directions des vents constituent d'autres exemples. De façon plus générale, avec les dévelopements des techniques de détection et d'enregistrement, tout phénomène continu peut être suivi et analysé avec une précision aussi grande que désirée. Cette question de la multiplication des observations évoquée ici ne concerne pas uniquement les variables temporelles, mais tous les types de données. Le recensement d'une population conduit aussi à une très grande quantité d'information. De plus, l'utilisation de variables qualitatives plutôt que quantitatives peut aussi amener des problèmes propres. Cependant, dans un souci de clarté de l'exposé, nous nous concentrerons ici sur les problèmes particuliers liés à l'analyse des séries temporelles numériques.

D'un point de vue pratique, l'augmentation constante de la taille des bases de données implique que le temps de calcul nécessaire soit pour une simple extraction des données présentant certaines caractéristiques particulières, soit pour des analyses statistiques plus sophistiquées, devient de plus en plus long. Un bon exemple est constitué par l'utilisation de modèles de mélange gaussien pour la prédiction de données financières pour lesquels l'utilisation combinée d'un algorithme EM et d'un algorithme génétique implique des calculs extrêmement lourds [BER 03]. De même, le temps nécessaire à l'exécution d'une requête SQL pour l'extraction d'un ensemble d'enregistrements s'accroît dans la même proportion. Même s'il est utile de disposer des données les plus complètes et précises possible, la question se pose de savoir si une telle quantité d'information est réellement indispensable dans tous les cas. Supposons par exemple que l'on veuille étudier la distribution d'un actif financier coté en bourse durant l'année 2002. Est-il indispensable alors d'utiliser des données établissant le cours de cet actif minute par minute, ou ne serait-il pas possible d'obtenir à moindre coût la même distribution à partir de données horaires, voire journalières ?

Dans cet article, nous étudierons l'influence de différentes procédures de réduction du nombre de données sur le comportement de séries temporelles. Nous prendrons comme point de départ plusieurs séries de données réelles présentant des comportements variés, puis nous appliquerons plusieurs principes de réduction des données et nous comparerons les nouvelles séries ainsi formées à la série d'origine aux moyens de tests statistiques.

La suite de cet article est organisée comme suit. Différentes méthodes permettant de réduire le nombre d'observations d'une série temporelle sont introduites dans la section 2 et les tests statistiques utilisés sont présentés dans la section 3. Les résultats

détaillés de nos analyses apparaissent dans la section 4 et une conclusion résume nos résultats et donne des recommandations générales.

2. Réduction des données

Nous considérons une variable aléatoire X_t prenant valeur dans l'ensemble des réels et observée durant N périodes successives. L'ensemble des observations de cette variable nous donne une série temporelle de longueur N que nous noterons S_1. Notre problème consiste à remplacer cette série temporelle par une autre série plus courte, de longueur $n < N$, conservant les principales caractéristiques de la distribution de la série d'origine. Une remarque préalable ayant trait à la notion même de série temporelle s'impose tout d'abord. Si nous travaillions avec des données transversales, comme celles résultant d'un recensement de population par exemple, le problème ci-dessus serait résolu en sélectionnant aléatoirement un échantillon représentatif de la taille désirée dans l'ensemble des données (se référer par exemple à [KIS 65]). Dans le cas de séries temporelles, cette méthode n'est pas adéquate car elle brise la relation existant éventuellement entre observations successives. La procédure suivante est alors utilisée :

1) La série originale d'observations S_1 est tout d'abord découpée en intervalles de longueur ℓ, ℓ étant un entier strictement positif représentant le nombre d'observations successives de S_1 inclues dans chaque intervalle. Le nombre d'intervalles ainsi créé est égal à la partie entière du rapport N/ℓ et est noté n. Dans le cas où N/ℓ n'est pas entier, les dernières observations de la série d'origine ne sont pas utilisées pour la création de la nouvelle série.

2) Les ℓ observations de chaque intervalle sont ensuite remplacées par une nouvelle valeur unique et la suite de ces valeurs constitue une série de données réduites de longueur n. Cette série ayant été créée à partir d'intervalles de longueur ℓ, elle est notée S_ℓ.

La procédure ci-dessus amène deux questions, à savoir le choix de la longueur ℓ de chaque intervalle et la méthode utilisée pour remplacer chaque groupe de ℓ valeurs par une nouvelle valeur unique. Le choix de ℓ, dépend de deux considérations antagonistes :

– Plus ℓ est grand, plus la réduction du nombre de données sera importante. Le choix de ℓ pourrait donc être effectué de façon à obtenir exactement le nombre de données désiré. Par exemple, si la série d'origine S_1 contient 10'000 observations et que l'on désire n'en avoir que 1'000, le choix de $\ell = 10$ serait judicieux.

– Ainsi que nous l'observerons dans la section 4, plus ℓ est grand, plus le risque que la nouvelle série S_ℓ se comporte de façon différente de S_1 augmente. De plus, il ne faut pas tomber dans l'excès consistant à choisir un trop petit nombre de données, auquel cas la fiabilité des méthodes statistiques appliquées ensuite sur ces données serait remise en cause.

Le choix de ℓ influe donc directement sur la qualité des résultats et dépend fortement du contexte, c'est-à-dire des données d'origine. En ce qui concerne la méthode utilisée pour transformer un groupe de ℓ observations en une nouvelle valeur unique, deux principes peuvent être utilisés :

– Tout d'abord, il est possible de remplacer les ℓ observations d'origine par l'une de ces ℓ observations. Cela se justifie aisément si l'on considère l'exemple suivant : Supposons que le cours d'une action soit enregistré dix fois à intervalles réguliers chaque jour d'ouverture de la bourse. Si nous désirons diminuer par 10 la quantité de données ainsi accumulées, un choix logique consiste alors à considérer les dix observations journalières comme un intervalle et à les remplacer soit par le cours d'ouverture de l'action (la première des 10 observations), soit par son cours de clôture (la dernière des 10 observations). Dans le cadre de cet article, la première méthode utilisée consistera à attribuer à chaque intervalle la valeur de la première observation de ce dernier.

– Une seconde possibilité consiste à remplacer les valeurs observées dans un intervalle par une fonction traduisant le comportement moyen de ces observations. Dans ce contexte, la moyenne et la médiane constituent deux choix logiques, chacune d'elles traduisant la notion de centre des données, avec un accent mis sur l'utilisation de toute l'information dans le cas de la moyenne et de la robustesse dans le cas de la médiane. Ces deux posibilités seront également utilisées dans la partie numérique de l'article.

3. Tests statistiques

Nous nous intéressons ici plus particulièrement à deux caractéristiques des séries temporelles, à savoir la distribution de fréquence des données et la relation entre deux observations successives. Le principe général des expériences que nous avons effectuées consiste à comparer à l'aide de tests statistiques les caractéristiques d'une série réellement observée S_1 avec celles de toutes les séries réduites S_ℓ, $\ell > 1$, calculées à partir de S_1.

Dans le cas des distributions de fréquences inconditionnelles, deux tests sont considérés, soit le test de Kolmogorov-Smirnov et le test du rapport de vraisemblance. Le test de Kolmogorov-Smirnov présente l'avantage de s'appliquer directement sur les données, sans qu'il soit nécessaire d'effectuer des transformations supplémentaires. En revanche, le test du rapport de vraisemblance nécessite la répartition préalable des observations en un nombre fini de classes. En pratique, nous déterminons celles-ci en fonction de la série originale S_1. Un nombre de classe k est choisi (par exemple 10 classes ou 20 classes, selon la précision recherchée), puis les bornes de ces classes sont calculées de façon à ce que chaque classe contienne approximativement la même proportion des données originales. Les données de la série S_ℓ à tester sont ensuite réparties dans ces mêmes classes et le test du rapport de vraisemblance est effectué. Le lecteur peut se référer notamment à [ZAR 99] pour de plus amples détails sur les caractéristiques et procédures d'utilisation de chaque test et à [HAM 94] pour des compléments sur les méthodes d'analyse des séries temporelles.

Lorsque l'on s'intéresse à la relation existant entre deux données successives, les observations des deux séries temporelles S_1 et S_ℓ sont tout d'abord réparties en classes selon le principe exposé ci-dessus. Les données catégorisées sont ensuite utilisées pour créer des tables de contingence mettant en relation deux observations successives. Le test du rapport de vraisemblance est alors appliqué afin de comparer les fréquences observées dans la table de contingence correspondant à la série d'origine S_1 avec celles calculées à partir de la série réduite.

La relation entre les distributions de fréquence de deux séries différentes observées durant la même période peut également être étudiée. Nous pensons par exemple à la relation entre deux indices boursiers. Dans ce cas, les deux séries de données à comparer sont tout d'abord catégorisées selon le principe énoncé ci-dessus, puis des tables de contingence sont calculées mettant en relation chaque observation de la première série avec l'observation correspondante de la seconde série. Le test du rapport de vraisemblance est ensuite appliqué comme précédemment pour comparer les tables de contingence résultant des séries réduites avec celle calculée à partir des séries d'origine. Il est à noter que lors de la création de tables de contingence mettant en relation deux séries différentes, le nombre de catégories utilisé pour chaque variable peut aussi être calculé de façon à maximiser l'intensité de la relation entre les deux variables [RIT 01].

4. Analyses numériques

Nous présentons dans cette section les données réelles utilisées pour nos expériences numériques, puis nous détaillons les résultats obtenus à partir de ces dernières.

4.1. *Données*

Les séries temporelles peuvent présenter une très grande variété de comportements, différant de par l'espérance ou la variance, présentant ou non des phénomènes d'hétéroscédasticité ou d'hétérogénéité, etc ... Dans le cadre d'une approche empirique, il serait bien entendu illusoire de vouloir traiter tous les cas possibles. Aussi, nous avons décidé de nous restreindre dans le cadre de cet article à un petit nombre de séries soigneusement choisies, mais présentant des comportements suffisamment représentatifs pour que des enseignements généraux puissent en être déduits.

Les quatre premières séries apparaissent sur la figure 1. La première série utilisée donne le prix de l'once d'or en dollars du 3 janvier 1968 au 31 mars 1998 pour un total de 7890 observations. Ainsi que cela apparaît sur la figure, cette série temporelle varie considérablement au fil du temps, mais cette évolution est relativement lente, les différences d'un jour à l'autre n'étant généralement que de faible amplitude.

La deuxième série est un enregistrement de mesures sismiques (accélérations verticales exprimées en nanomètres par seconde au carré) effectué lors du tremblement de terre de Kobe (Japon) le 16 janvier 1995. Les données ont été récoltées sur une

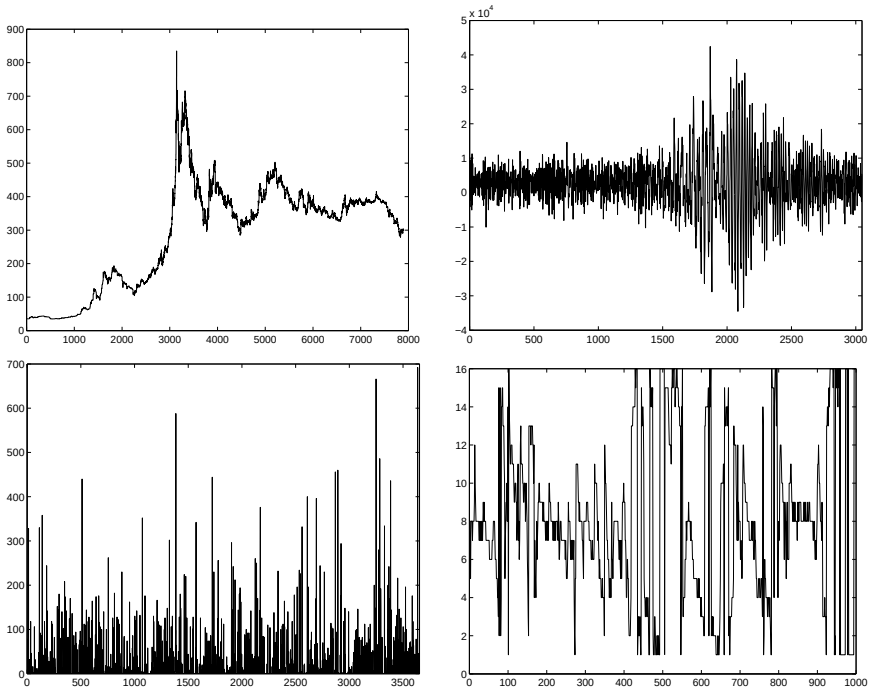

Figure 1. *Quatre séries temporelles. En haut à gauche : Prix de l'once d'or en dollars entre le 3 janvier 1968 et le 31 mars 1998. En haut à droite : Données sismiques enregistrées durant le tremblement de terre de Kobe le 16 janvier 1995. En bas à gauche : Précipitations quotidiennes de pluie mesurées en millimètres à Melbourne de 1981 à 1990. En bas à droite : Direction moyenne du vent mesurée chaque heure à Koeberg entre le premier mai 1985 et le 30 avril 1989 et codée sur 16 valeurs. Seules les 1000 premières observations de cette série sont présentées ici.*

période de 51 minutes à une seconde d'intervalle, pour un total de 3048 observations. Au contraire de la série précédente, les données de celle-ci oscillent très rapidement autour d'une valeur centrale proche de zéro. Un phénomène d'hétéroscédasticité est également présent. Ce type de comportement se rencontre dans de nombreux autres domaines, notamment en finance lorsque l'on s'intéresse aux rendements composés d'un titre ou d'un indice, c'est-à-dire à une série calculée comme la différence des logarithmes de deux observations successives.

La troisième série présente les précipitations quotidiennes de pluie mesurées en millimètres à Melbourne (Australie) de 1981 à 1990 pour un total de 3653 observations. Ainsi que le montre la figure 1, ces données sont très variables et il n'est pas rare de passer directement d'un jour sec à un jour durant lequel il est tombé plusieurs centimètres de pluie.

La quatrième série, dont le début est présenté sur la figure 1, donne la direction horaire moyenne du vent relevée à Koeberg (Afrique du Sud) pour la période allant du premier mai 1985 au 30 avril 1989 (35064 observations). Les directions sont codées dans le sens des aiguilles d'une montre en 16 catégories allant du nord (code 1) au nord-nord-ouest (code 16). Par rapport aux autres séries utilisées dans cet article, celle-ci présente deux différences majeures : Tout d'abord les valeurs de la série ne sont pas continues, mais discrètes. D'autre part, les données sont circulaires, ce qui signifie que le code 16 est lui-même suivi du code 1. Cette dernière caractéristique aura une influence certaine sur les résultats.

Finalement, deux fichiers de données multivariées sont utilisés pour tester l'influence des recodages sur les relations intervariables. Le premier de ces fichiers donne la vitesse journalière moyenne du vent relevée dans 12 stations météorologiques d'Irlande durant la période 1961-1978 et exprimée en noeuds par heure. Chaque variable comporte 6574 observations. Ces variables ont un comportement similaire à celui observé pour les relevés sismiques du tremblement de terre de Kobe, excepté le fait qu'elles ne peuvent pas prendre de valeurs négatives et qu'elles sont pour la plupart centrées sur une valeur proche de 10 noeuds par heure. Le second fichier donne la valeur de 7 indices boursiers relevée entre le premier juin 1986 et le 31 décembre 1997 pour un total de 3128 observations par variable. Ces indices présentent tous des caractéristiques similaires à celles du prix de l'once d'or.

4.2. *Résultats*

Le principal élément que nous avons dû fixer au début de l'étude est l'ensemble de valeurs considérées pour le paramètre ℓ, c'est-à-dire les différentes longueurs des intervalles utilisés pour découper les séries d'origine S_1. Notre choix s'est porté sur 6 valeurs, soit 2, 3, 4, 5, 10 et 20. Pour tous les tests statistiques effectués, le risque de première espèce a été fixé à 5%. Le nombre de catégories utilisé pour les tests du rapport de vraisemblance est fixé en fonction des données. Si le nombre de catégories est trop élevé, les fréquences observées dans tout ou partie des catégories seront faibles, spécialement dans le cas des données recodées, ce qui risque de biaiser les tests en faveur d'une trop grande acceptation de l'égalité des distributions.

Nous avons tout d'abord comparé les distributions de fréquences obtenues après recodage avec la distribution des données originales. Dans le cas du prix de l'or, aucun des recodages n'affecte significativement la distribution des données. Quel que soit le type de recodage (remplacement par la première valeur du sous-intervalle, par la moyenne ou par la médiane), la longueur du sous-intervalle ou le nombre de classes retenues (dans le cas du test du rapport de vraisemblance), la distribution des données recodées est toujours jugée similaire à celle des données originales. Les mêmes tests effectués sur les 7 indices boursiers amènent aux même conclusions, ce qui n'est pas surprenant étant donné le comportement similaire de ces différentes variables. Ces résultats s'expliquent de par la nature des données utilisées, les valeurs évoluant lentement et les changements brusques restant l'exception.

La série de données sismiques amène à des conclusions extrêmement différentes. Si le recodage effectué selon le principe de la première observation de l'intervalle ne modifie en rien les distributions de fréquences, l'utilisation de la moyenne ou de la médiane amène au rejet de tous les tests, exception faite du cas ℓ=2 pour le test du rapport de vraisemblance. Un phénomène similaire s'observe pour les 12 séries de vitesse du vent. La raison de ce comportement est parfaitement illustrée par la figure 2 présentant des histogrammes à 20 classes. En haut à gauche, nous avons l'histogramme des données sismiques originales S_1. En haut à droite, l'histogramme présente le résultat obtenu en regroupant les données en intervalles de longueur $\ell = 5$ et en attribuant à chaque intervalle la valeur de la première observation. Cet histogramme reste assez proche de l'histogramme original. En bas à gauche, l'histogramme présente le résultat obtenu en utilisant la moyenne des 5 observations de l'intervalle et le dernier histogramme utilise la médiane. Etant donné que les données originales oscillent très rapidement autour d'une valeur centrale proche de zéro, la probabilité d'obtenir un résultat proche de zéro lorsque l'on prend la moyenne ou la médiane de plusieurs observations successives (5 dans notre exemple), est grande. Cela explique que sur les deux derniers histogrammes, les fréquences correspondant aux valeurs proches de zéro soient sur-représentées au détriment des autres classes.

Les données pluviométriques de Melbourne donnent des résultats similaires à ceux obtenus à partir des données sismiques, à savoir un rejet quasi-systématique des tests d'égalité des distributions de fréquence dès lors que la moyenne ou la médiane sont utilisées. Cependant, la raison de ces rejets est différente. Les données originales ont ici deux caractéristiques qui les différencient de la série précédente : La valeur zéro est fréquemment présente plusieurs fois consécutivement et de très grandes valeurs apparaissent périodiquement. Ainsi, lorsque l'on calcule la moyenne de plusieurs observations consécutives, deux phénomènent peuvent se produire : i) Certains intervalles se voient attribuer une valeur supérieure à zéro alors que presque toutes les valeurs de l'intervalle sont nulles. ii) Certains intervalles se voient attribuer une très forte valeur uniquement parce qu'une seule des observations est très élevée. Ainsi que l'on peut le constater à partir des histogrammes de la figure 3, ces phénomènent accentuent très fortement la fréquence des grandes valeurs au détriment des plus petites et de zéro en particulier. L'utilisation de la médiane conduit à un phénomène inverse, nettement moins marqué, mais suffisant toutefois pour amener au rejet des tests.

Les tests effectués sur la direction du vent à Koeberg aboutissent à des conclusions similaires, à savoir une acceptation de l'égalité des distributions de fréquences des données originales et recodées lorsque l'on attribue à chaque intervalle sa première valeur, et un rejet dans les deux autres cas. Ici, l'explication est due en grande partie au problème de la circularité des données. Supposons que durant un intervalle de longueur 5, la direction du vent passe progressivement du nord-ouest au nord-est. Les 5 observations sont alors codées respectivement 15, 16, 1, 2 et 3, leur moyenne valant 7.4 et leur médiane 3. Aucune de ces deux valeurs ne peut donc prétendre résumer valablement les 5 valeurs originales, ce qui explique le rejet des tests.

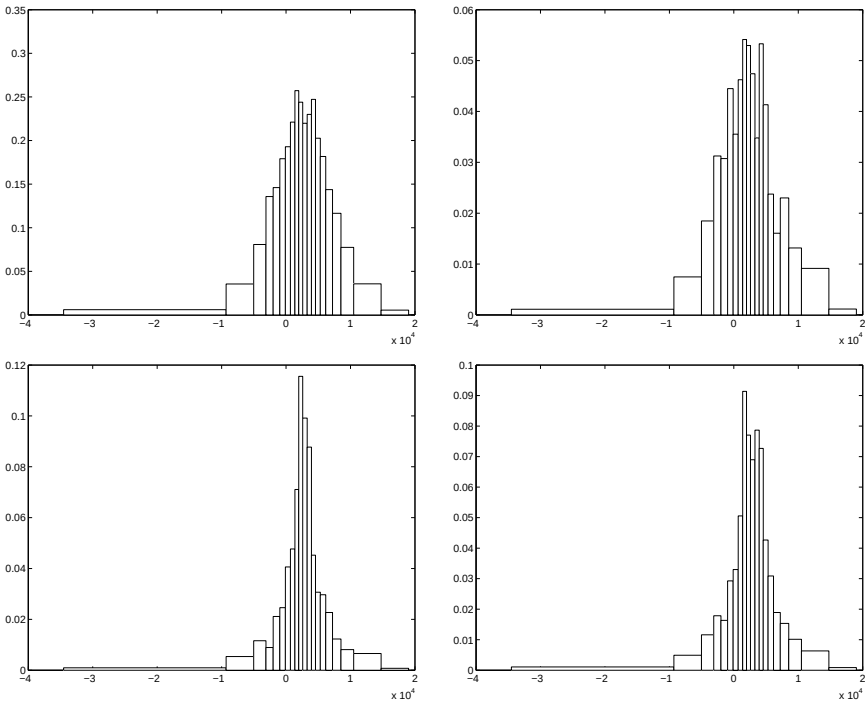

Figure 2. *Quatre histogrammes à 20 classes pour les données sismiques de Kobe. En haut à gauche, nous avons les données observées. Pour les trois autres histogrammes, les données originales ont été préalablement réparties en intervalles de longueur 5 et les principes suivants ont été utilisés : première observation de l'intervalle (en haut à droite), moyenne (en bas à gauche) et médiane (en bas à droite).*

Dans un second temps, nous avons étudié la relation liant deux observations successives d'une série temporelle. Les données ont été catégorisées, puis nous avons construit des tables de contingence mettant en relation deux observations successives et ces tables ont finalement été comparées à l'aide du test du rapport de vraisemblance.

En ce qui concerne le prix de l'or, les trois méthodes de recodage se révèlent toujours adéquates lorsque les intervalles sont de longueur ℓ=2 ou 3. Pour ℓ=4, la moyenne et la médiane conviennent encore et seule la moyenne convient pour ℓ=5. Au-delà, tous les tests sont rejetés. Dans le cas des 7 indices boursiers, les résultats sont similaires, mais les tests sont cependants acceptés, pour certaines séries, pour des valeurs de ℓ allant jusqu'à 20, y compris en utilisant la méthode de la première valeur. Cette dégradation des résultats par rapport à ceux enregistrés pour les tests portant sur les distributions de fréquences inconditionnelles s'explique notamment par le plus grand nombre de situations qui sont testées, c'est-à-dire ici de couples

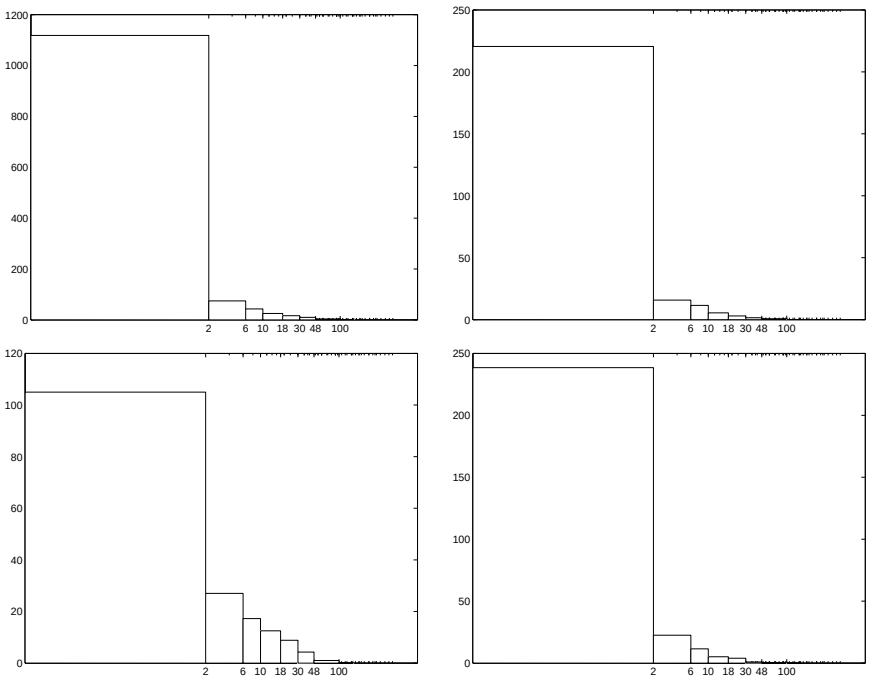

Figure 3. *Quatre histogrammes pour les précipitations à Melbourne. En haut à gauche, nous avons les données observées. Pour les trois autres histogrammes les données originales ont été réparties en intervalles de longueur 5 et les principes suivants ont été utilisés : première observation de l'intervalle (en haut à droite), moyenne (en bas à gauche) et médiane (en bas à droite). En raison du très grand nombre de valeurs égales à zéro dans les données originales, la première classe de l'histogramme contient beaucoup plus d'observations que les autres classes. De plus, une échelle logarithmique est utilisée pour l'axe horizontal.*

d'événements successifs. Il peut suffire que le recodage reproduise incorrectement une petite proportion de ces couples pour que le test soit rejeté.

Dans le cas des données sismiques de Kobe et dans celui des 12 séries de vitesse du vent, tous les tests rejettent systématiquement les recodages, quelles que soient la méthode de recodage et la longueur des intervalles. Ce résultat traduit une très forte relation entre observations successives dans la série originale. Toute tentative de recodage casse cette relation pour la remplacer par une autre.

Pour les précipitations à Melbourne, les résultats sont similaires à ceux obtenus dans le cas des fréquences inconditionnelles, à savoir que la méthode de la première valeur conduit toujours à l'acceptation du test, alors que les deux autres méthodes mènent à un rejet systématique. Au-delà des explications fournies précédemment, il est possible de faire l'hypothèse que ces données ont un comportement proche des

fractales, à savoir que la relation observée entre deux observations successives des données originales se reproduit à l'identique lorsque l'on ne considère par exemple qu'une observation sur 2 ou sur 5. La première méthode de codage des intervalles est alors parfaitement appropriée.

La direction du vent à Koeberg amène à un rejet systématique de tous les tests. Cela est notamment dû au fait que dans les données originales, la direction du vent évolue lentement. En ne considérant plus qu'une observation sur 2 ou sur 5, les fréquences correspondant à de grands changements de direction augmentent et celles traduisant une absence de changement diminuent en proportion.

Finalement, nous avons étudié les effets de la réduction des données sur les relations entre séries. A cet effet, nous avons utilisés les deux fichiers de données contenant respectivement 7 séries d'indices boursiers et 12 séries de vitesses du vent. Dans chacun des deux cas, nous avons catégorisé les séries et nous avons construit des tables de contingence mettant deux séries en relation. Les 7 séries d'indices boursiers ont permis un total de 21 comparaisons deux à deux et les 12 séries de vitesse du vent ont permis 66 comparaisons.

Dans le cas des indices boursiers, tous les tests d'égalité des tables de contingence ont été systématiquement accepté, quelle que soit la méthode de codage employée. Dans le cas des vitesses du vent, la méthode de la première valeur est systématiquement acceptée, alors que les deux autres méthodes sont systématiquement rejetées. Cette divergence par rapport aux séries boursières s'explique par la même raison invoquée dans le cas des fréquences inconditionnelles pour ces mêmes données, à savoir une tendance des méthodes de la moyenne et de la médiane à surestimer les fréquences des valeurs se trouvant proches du centre de la distribution. Ce phénomène se répercute sur les relations entre séries.

5. Conclusion

Nous sommes fermement convaincus que dans le futur le nombre de données disponibles, quel que soit le domaine de recherche, ne fera que s'accroître. Par ailleurs, les outils statistiques utilisés pour explorer de telles bases de connaissances deviendront de plus en plus complexes et requéreront encore plus de puissance et de temps de calcul qu'actuellement [DEB 94]. Dans ce contexte, le besoin de méthodes permettant de travailler de manière efficace à partir d'un sous-ensemble des données originales se fera de plus en plus sentir. Nous avons exploré ici certaines pistes dans le cas des séries temporelles. Nous espérons que ce travail pourra servir de point de départ pour étudier d'autres possibilités et généraliser plus avant nos résultats.

La possibilité de réduire le nombre de données dépend non seulement de l'utilisation finale que l'on a prévu de faire de ces données (type de modèle statistique utilisé), mais aussi du comportement des données de départ. Nous avons ainsi pu observer que la variabilité des observations est un facteur très important. Si cette variabilité est trop

grande, la réduction du nombre de données s'accompagne d'un changement significatif du comportement des variables et influe donc directement sur leur interprétation.

Les mauvais résultats obtenus sur les tables de contingence construites entre deux observations successives d'une même variable n'indiquent pas forcément que le recodage est une mauvaise chose. En réalité, seules les méthodes de recodage envisagées ici sont peut-être en cause. Il est possible que l'utilisation de modèles statistiques plus performants pour associer à chaque intervalle une valeur unique puisse résoudre le problème. On pourra rétorquer que le fait de réduire le nombre de données selon les méthodes proposées dans cet article doit de toute façon aboutir à une modification fondamentale de la relation entre observations. En pratique, cela n'est pourtant pas automatiquement vrai. Un bon exemple est proposé par le modèle MTD [RAF 85] développé pour l'approximation des chaînes de Markov d'ordre élevé et qui fonctionne parfaitement tout en supprimant la relation entre les différents retards de la chaîne de Markov. Il est encore à noter que dans le cas de la méthode de la première valeur, le recodage obtenu représente toujours de façon correcte les relations inter-données, mais à une échelle différente, ce qui peut mener à de graves difficultés lors de l'interprétation des résultats.

En ce qui concerne les relations inter-séries, nous observons que si deux variables ayant un lien fort (parallélisme, relation inverse, ...) sont recodées de la même façon, la même dégradation s'opère sur les deux séries. Cela implique que la relation inter-série n'est pas affectée par le recodage, alors que le comportement des deux séries prises individuellement peut éventuellement l'être. En revanche, deux séries observées sur la même période mais sans qu'une relation forte n'existe entre-elles seront beaucoup plus difficiles à recoder.

6. Bibliographie

[BER 03] BERCHTOLD A., « Mixture Transition Distribution (MTD) Modeling of Heteroscedastic Time Series », *Computational Statistics and Data Analysis*, vol. 41, 2003, p. 399-411.

[DEB 94] DEBOECK G., *Trading on the Edge : Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets*, Deboeck Editor, Wiley, 1994.

[HAM 94] HAMILTON J., *Time Series Analysis*, Princeton University Press, 1994.

[KIS 65] KISH L., *Survey Sampling*, John Wiley & Sons, 1965.

[RAF 85] RAFTERY A., « A Model for High-Order Markov Chains », *Journal of the Royal Statistical Society B*, vol. 47, 1985, p. 528-539.

[RIT 01] RITSCHARD G., ZIGHED D., NICOLOYANNIS N., « Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé », *Mathématiques et Sciences Humaines*, vol. 154-155, 2001, p. 81-97.

[ZAR 99] ZAR J., *Biostatistical Analysis*, Prentice Hall, 4ème édition, 1999.