L'ENSEIGNEMENT DE LA STATISTIQUE AVEC TABLEUR. MISE À JOUR

Guy Mélard[1]

[1] *Université libre de Bruxelles, ECARES CP114/4, Avenue Franklin Roosevelt 50, B-1050 Bruxelles, Belgique et ITSE sprl, Bruxelles, Belgique. gmelard@ulb.ac.be*

Résumé. Mélard (2010) a présenté une communication sous le titre "Utiliser un tableur dans l'enseignement de statistique. Pourquoi et comment ?". Prenant la suite de McCullough (2008), l'exposé commençait par une mise à jour des critiques relatives aux possibilités statistiques d'Excel 2010 de Microsoft. Mélard (2014) a ajouté l'examen de l'outil Solver et a considéré également OpenOffice Calc et Gnumeric. On commence par actualiser les conclusions pour Excel 2013 et Excel 2016, ainsi que pour les autres tableurs. On discute ensuite des avantages et inconvénients de l'emploi d'Excel pour l'enseignement de la statistique et on examine un échantillon d'ouvrages qui mettent en œuvre cette approche. La présentation de 2010 contenait déjà une esquisse d'étude de cas relative à l'enseignement de la statistique à des fins de prévision (Mélard, 2007) mais elle est ici approfondie avec un examen particulier de la nécessité de mettre à jour les classeurs pour les besoins des versions récentes d'Excel et des autres tableurs.

Mots-clés. Microsoft Excel, OpenOffice, LibreOffice, Gnumeric, séries chronologiques, méthodes de prévision.

Abstract. Mélard (2010) presented a talk on the subject "Using a spreadsheet in teaching statistics. Why and how?". Following McCullough (2008), the presentation started with an update of the criticisms relative to the statistical capabilities of Microsoft Excel 2010. Mélard (2014) added the Solver add-in to the analysis and extended it also to OpenOffice Calc and Gnumeric. We begin with an update of the conclusions for Excel 2013 and Excel 2016, as well as for other spreadsheets. Then, we discuss the advantages and inconveniences of using Excel for teaching statistics. We examine a sample of textbooks which implement that approach. The 2010 presentation already contained a sketch a case study relative to teaching statistics with a focus on forecasting methods (Mélard, 2007) but it is extended here by examining in particular the need to update the workbooks for recent versions of Excel and the other spreadsheets.

Keywords. Microsoft Excel, OpenOffice, LibreOffice, Gnumeric, time series, forecasting methods.

1 Introduction

Mélard (2010) a présenté une communication sous le titre "Utiliser un tableur dans l'enseignement de statistique. Pourquoi et comment ?". Prenant la suite de McCullough (2008), l'exposé commençait par une synthèse des critiques relatives aux possibilités statistiques des différentes versions d'Excel de Microsoft, et en particulier la version 2010. Ces critiques portaient sur la précision des fonctions statistiques, les outils complémentaires, le générateur de nombres aléatoires, et différents graphiques, et s'étendait à OpenOffice.org 3.2. Un article (Mélard, 2014) a poursuivi cette étude en l'approfondissant et en l'étendant à d'autres tableurs et au module Solver. L'auteur reçoit régulièrement des demandes pour actualiser ses conclusions à des versions plus récentes. La présentation de 2010 contenait déjà une esquisse d'étude de cas relative à l'enseignement de la statistique à des fins de prévision (Mélard, 2007) qui est ici approfondie.

Les conclusions de Mélard (2014) au sujet d'Excel 2010 étaient les suivantes.

1. Pour les fonctions statistiques, la plus grande partie des problèmes d'imprécision d'Excel 2007 ont été corrigés dans la version 2010, presque aussi bonne qu'OpenOffice.org Calc 3.3. Notons l'ajout de nouvelles fonctions avec des noms plus explicites.
2. Concernant les générateurs de nombres pseudo-aléatoires, celui du Visual Basic for Application (VBA) n'a pas changé depuis longtemps et est connu pour être de mauvaise qualité. Celui fourni par l'outil complémentaire a les mêmes défauts qu'auparavant. Enfin, le générateur invoqué par la fonction ALEA() d'Excel 2007 a été remplacé par un algorithme Mersenne Twister dont la période est $2^{19937} - 1$. Effectivement il passe la plupart des tests.
3. Les outils complémentaires d'Excel 2010 sont inchangés par rapport à la version 2003. Ils présentent les mêmes problèmes que ceux signalés depuis de nombreuses études.
4. Le Solver d'Excel permet notamment de réaliser de l'optimisation linéaire ou non linéaire, sans ou avec contraintes. Il peut avoir différents usages en statistique, en particulier pour la régression non linéaire. Microsoft affirme avoir amélioré le Solver dans la version 2010, notamment en ajoutant une méthode Multistart à l'algorithme de base GRG2 et en ajoutant un nouvel algorithme appelé Evolutionary. Même si la qualité des rapports est améliorée, sur base d'une batterie de problèmes de test, il n'y a pas d'amélioration sensible de l'algorithme GRG2 qui fournit zéro décimales correctes pour 12 tests sur 27. Les deux nouveaux algorithmes qui reposent sur des spécifications d'intervalles plausibles de variation pour les paramètres donnent des résultats catastrophiques en l'absence de ces spécifications.
5. Trente ans après Tufte (1983), les graphiques par défaut d'Excel 2010 sont toujours mauvais pour des données statistiques mais aussi pour d'autres données. La situation semble empirer avec un accent tridimensionnel prononcé et il faut de plus en plus d'efforts pour éliminer le superflu des graphiques. On ne trouve pas les graphiques statistiques spécifiques comme les boîtes à moustache.

En bref, Mélard (2014, p. 1126) dit que "The recent improvements reported in this paper should not hide the fact that Microsoft is still marketing a product that contains known errors."

Nous actualisons à Excel 2013 et 2016 les remarques précédentes dans le paragraphe 2. Ensuite, au paragraphe 3, nous exposons l'état de la littérature au sujet des tableurs pour l'enseignement en statistique. Au paragraphe 4, nous développons une étude de cas pour l'enseignement en analyse des séries chronologiques en mettant en évidence les problèmes de compatibilités entre tableurs.

2 Actualisation de l'étude relative à Excel 2010 et d'autres tableurs

Mélard (2014, p. 1126) ajoute "We didn't analyze Excel in Office 2013 but, according to Microsoft (2013), where the changes with respect to Office 2010 are collected, there are few changes to Excel and nothing about the statistical aspects is mentioned." Il nous reste donc à considérer Excel 2016 et les nouvelles versions d'Open Office et LibreOffice. Mélard (2014) avait aussi considéré Gnumeric qui était très bien coté.

Microsoft (2017) reprend les modifications successives appliquées à Excel pour la version 2016 (notamment pour les utilisateurs d'Office 365). Aucun changement des possibilités statistiques n'est mentionné sauf l'apparition de trois nouveaux graphiques statistiques. D'abord la procédure de génération de nombres pseudo-aléatoires selon une loi uniforme sur [0, 1) donne des résultats différents mais toujours un résultat 1 et le générateur du VBA est inchangé. Nous avons vérifié que rien n'a changé dans les procédures statistiques, par comparaison avec Mélard (2014, paragraphe 4.2). De même, le Solver d'Excel 2016 fournit des résultats identiques pour les deux colonnes "GRG2 set 1" du tableau 12. Les trois nouveaux graphiques statistiques sont un histogramme, un diagramme de Pareto et un diagramme de boîtes à moustache. Pour plus de détails sur tous ces aspects, ainsi que

pour ce qui est relatif à OpenOffice Calc et LibreOffice Calc, voir l'annexe A de la version complète de cet article parmi les ressources sur http://www.itse.be. En plus des références données par Mélard (2014), voir aussi Botchkarev (2015) pour les générateurs de nombres pseudo-aléatoires et Cooke *et al.* (2016), pour une critique portant sur les graphiques.

3 La littérature sur le sujet des tableurs en enseignement de la statistique

Le sujet n'est pas neuf mais n'a pas beaucoup été traité. Voir toutefois Cryer (2001) et Carr (2002) pour des points de vue opposés. Nash (2008) a très bien abordé le sujet assez en détail. Il discute des activités d'enseignement pour lesquelles un tableur peut être employé, du caractère approprié ou non des tableurs pour ces activités, et du choix d'Excel, en l'occurrence de la version 2007, quand un tableur peut être approprié. Plus récemment, Freeman (2014) traite de la manière d'effectuer des tests non paramétriques simples avec Excel et discute aussi de la manière particulière que possède Excel d'indiquer la colinéarité, Mélard (2014, paragraphe 4.2 et figure 3) mais ajoute que le nombre de degrés de liberté est alors erroné, ce qui implique des erreurs pour le R^2 corrigé, le test F de Fisher et les erreurs-types des coefficients. Citons aussi le travail de Dell'Omodarme et Valle (2006) qui combine Excel et R et sort donc du cadre de cet article.

Nash (2008) critique notamment le fait qu'une modification de données tantôt produise un changement de résultat (quand des formules sont employées), tantôt n'en produise pas (lors de l'emploi de la plupart des procédures statistiques des outils complémentaires). Tout en déplorant l'interface, notamment le ruban d'Office 2007 (et les versions suivantes), Nash (2008) reconnaît que la précision d'Excel peut être suffisante mais déplore le manque de pédagogie qu'il y a à enseigner avec un outil qu'on ne peut pas recommander pour un emploi dans la réalité professionnelle. Personnellement, cela nous gêne moins. Chaque outil a sa force et il est important de signaler aux étudiants les limites d'Excel qui est de plus en plus utilisé, et même maintenant dans l'enseignement secondaire en France. Par ailleurs, la statistique exige un esprit critique et donc un outil imparfait peut suffire pourvu que les risques d'erreur et les dangers soient traités durant l'enseignement, par le traitement d'exemples numériquement délicats tels que la présence de données aberrantes et d'effets de quasi-colinéarité.

De toutes manières, le marché de l'édition n'a pas attendu l'approbation des statisticiens. Nous avons un instant envisagé de citer tous les livres de statistiques mentionnant les mots « statistique » et « Excel » mais, même en français, ils sont trop nombreux pour être cités. Nous en avons donc pris un échantillon à notre disposition aussi bien en français qu'en anglais. Le but n'est pas de présenter une description exhaustive mais d'en tirer quelques enseignements. On remarque un auteur (Quirk, 2015a, 2015b, Quirk *et al.*, 2015) qui a, seul ou en collaboration, rédigé une douzaine de livres avec des titres semblables "Excel 20*nn* for *x* statistics", ou il faut remplacer *nn* par 07, 10 ou 13 et *x* par "business". "social sciences", "biological and life sciences", etc.

L'enseignement premier est le manque général de sens critique. Je n'y ai pas vu de mention des nombreux travaux de McCullough au sujet de la précision statistique d'Excel. C'est tout juste si certains mentionnent la difficulté de distinguer entre formules et procédures statistiques, comme Nash (2008) l'a indiqué et comme nous l'avons rappelé ci-dessus.

Commençons par les sujets traités. Certains auteurs (Quirk, 2011; Fraser, 2013) ne traitent de la statistique que les sujets qu'ils peuvent illustrer avec Excel en omettant les autres. D'autres auteurs (Bressoud et Kahané, 2010; Pupion, 2008; Salkind, 2007) essaient de couvrir tous les concepts de la statistique de base, se contentant d'illustrations partielles au moyen d'Excel. Enfin, Vidal (2010) couvre toute la matière avec Excel au prix de l'emploi de formules nombreuses et parfois d'approches concurrentes. Dans tous les cas, on voit une combinaison de fonctions, de formules et d'utilisation de procédures qui paraît fort hétéroclite. Georgin (2002) est une exception car il traite principalement des méthodes d'analyse de données, avec la régression, l'ACP et l'AFC, avec très peu de fonctions

mais une macro VBA pour calculer les valeurs propres d'une matrice par un algorithme de Jacobi. L'analyse des données n'est pas traitée dans les autres ouvrages.

Regardons ensuite la manière d'introduire Excel. Il y a souvent un chapitre ou un paragraphe introductif. On mentionne alors les fonctions et procédures pour chaque objet de l'analyse statistique. Tantôt (Quirk, 2015) on mélange la théorie et l'emploi d'Excel, généralement avec un accent sur le second, tantôt (Bressoud et Kahané, 2010), on restreint Excel aux exercices. La manière d'employer Excel est parfois sommaire, et d'autres fois pas à pas, soit en employant les outils du ruban (Quirk, 2015) ou même des raccourcis (Fraser, 2013). Cette dernière approche nous a semblé particulièrement déplaisante en plus d'être inutilisable avec une version d'Excel localisée, notamment en français.

4 Une étude de cas pour l'enseignement de l'analyse des séries chronologiques

Mélard (2010) avait déjà abordé le sujet de manière succincte en se référant à Mélard (2007) à titre d'exemple. Une table des matières est disponible dans la version complète de ce texte, en annexe B. Ici, nous réexaminons de manière aussi critique que possible le cours multimédia de notre livre « Méthodes de prévision à court terme », 2e édition. Un aspect particulier que nous traitons ici pour la première fois, est la nécessité, dans certains cas, de mettre à jour les classeurs d'Excel pour les besoins des versions récentes d'Excel. Nous traitons aussi des problèmes liés à l'emploi des autres tableurs.

L'analyse des séries chronologiques est une branche de la statistique qui possède les caractéristiques suivantes :

1. elle est employée non seulement par les statisticiens mais aussi dans la plupart des disciplines avec un accent particulier en économie et en finance;
2. elle est exigeante du point de vue de la théorie, puisque les observations ne constitue presque jamais un échantillon aléatoire simple;
3. elle demande des moyens de calcul supérieurs à celui de beaucoup de procédures statistiques, même avec peu de données.

En conséquence, la calculatrice n'est pas utilisable même dans les cas les plus simples. Les méthodes les plus abordables sont les suivantes et peuvent être traitées par Excel :

- la régression linéaire simple et la régression non linéaire ;
- les moyennes mobiles, y compris celles de Spencer et de Henderson, ainsi que des médianes mobiles ;
- la décomposition saisonnière par des méthodes élémentaires et par la méthode Census X-11 ;
- les lissages exponentiels simple, double, de Holt et de Winters ;
- la régression linéaire multiple ;
- l'autocorrélation ;
- une illustration d'analyse spectrale et de filtrage optimal.

Il est possible mais difficile de traiter des modèles ARIMA simples mais impossible d'aborder les méthodes de décomposition saisonnière récentes (Tramo-Seats et X-13ARIMA-SEATS). Pour plus de détails sur les méthodes elles-mêmes, voir Mélard (2007) ou l'article en accès libre de Mélard (2006).

Dans l'annexe C de la version complète de ce texte, nous illustrons ce qu'on peut effectuer avec Excel en matière de traitement de données chronologiques en prenant l'exemple de certains classeurs de Mélard (2007). A deux exceptions près, il s'agit de classeurs dont des versions mises à jour et corrigées sont mises à disposition (voir http://www.itse.be), essentiellement pour des problèmes de compatibilité (voir l'annexe D). Quiconque peut donc y accéder, pas seulement les lecteurs de Mélard (2007). Nous indiquons chaque fois les problèmes éventuels posés par Excel 2016 et Calc 5.0. Accessoirement, nous mentionnons aussi Gnumeric.

Ce cours multimédia est basé sur le matériel pédagogique développé au fil des années par l'auteur. Ainsi qu'expliqué par Cohen *et al*. (2003a) et Cohen *et al*. (2003b), le cours était basé sur structure de fichiers PDF, de classeurs Excel, et de traitement de données avec même un système d'auto-évaluation. Beaucoup des classeurs Excel avaient d'abord été développés pour Lotus 1-2-3. Ces derniers avaient été proposés comme suppléments lors de la première édition, Mélard (1990). Les classeurs des chapitres 8 et 13 ainsi que la plupart de ceux des chapitres 5 et 7 sont plus récents et avaient donc été développés directement pour Excel, en l'occurrence Excel 97.

La question qui est posée ici est la possibilité d'employer ces classeurs avec Excel 2010 et les versions suivantes et même les tableurs concurrents, comme OpenOffice et LibreOffice Calc et Gnumeric. Les différents aspects traités dans le corps de l'article doivent être considérés, ainsi évidemment que les aspects pratiques.

Dans le cours, chaque exercice, donc chaque classeur, est l'objet d'un fascicule d'instructions subdivisé éventuellement en plusieurs parties (avec aussi une distinction entre cours de base et cours avancé, que nous ne discutons pas ici). En principe, ces instructions devraient aussi être prises en compte ici mais, étant donné le volume de pages, nous ne pourrons que mentionner les éléments les plus critiques.

On trouvera les détails de l'analyse en annexe D. Les conclusions sont que, à part quelques petites erreurs, les classeurs conçus avec Excel 97 sont compatibles avec Excel 2010 et les versions suivantes et presque compatibles avec OpenOffice et LibreOffice Calc, sauf quelques classeurs cités dans l'annexe B. Presque tous les classeurs du cours fonctionnent dans Gnumeric sauf un qui repose trop sur des macros. Nous n'avons pas remarqué de différence dans les résultats dus à l'amélioration des fonctions statistiques d'Excel 2010 ou plus récent, ni de celle du générateur de nombres pseudo-aléatoires. A cause de l'absence des outils complémentaires, plusieurs parties d'exercices du chapitre sur la régression linéaire multiple ne sont pas disponibles dans OpenOffice 4.1.3 et Libre Office 5.2.7, sachant que Gnumeric propose des outils équivalents mais d'emploi légèrement différent. Compte tenu de la nature des données temporelles, il n'y a pas eu de problème avec les graphiques, essentiellement des graphes linéaires et des diagrammes de dispersion. Egalement, on a discuté des macros VBA, maintenant acceptées par OpenOffice et LibreOffice depuis la version 3.0 (mais pas par Gnumeric) qu'il a parfois fallu corriger, des hyperliens, des tables de données et d'autres aspects pratiques. Finalement, nous avons signalé des corrections d'erreurs diverses.

En définitive, les classeurs du cours multimédia de Mélard (2007) sont très bien acceptés mais l'interface ruban d'Excel a posé plus de problèmes, au point que la compatibilité s'avère globalement meilleure avec OpenOffice/LibreOffice Calc (à l'exception d'un classeur), et reste acceptable avec Gnumeric (à l'exception de quatre classeurs). Pour les différentes raisons mentionnées en annexe D, de nouvelles versions de 15 classeurs sont proposées sur le site du cours.

Pour résumer les problèmes, disons que les classeurs du cours, créés dans une version antérieure, sont ouverts par Excel 2010 et versions suivantes en mode compatibilité. A cause des macros, cela produit souvent (mais pas toujours) des messages « Avis de sécurité » qui peuvent être dissuasifs.

Bibliographie

[1] Botchkarev, A. (2015), Assessing Excel VBA suitability for Monte Carlo simulation, *Spreadsheets in Education* 8 (2), article 3.
[2] Bressoud, E. et Kahané, J. (2010), *Statistique descriptive: Applications avec Excel et calculatrices*, Pearson Education, Paris.
[3] Carr R. (2002), Teaching statistics using demonstrations implemented with Excel, 6th International Conference on Teaching Statistics (ICOTS), Haifa, Israel. 4 pp.
[4] Cryer, J. D. (2001), Problems with using Microsoft Excel for statistics, Proceedings of the joint statistical meetings. American Statistical Association, Atlanta.
[5] Cohen A., G. Mélard et Ouakasse, A. (2003a), Emploi d'un tableur dans un cours d'analyse de

séries temporelles, Actes des XXXVèmes Journées de Statistique, Lyon, 13-17 mai 2003, Société Française de Statistique, Tome 1, pp. 341-344. http://homepages.ulb.ac.be/~gmelard/Lyon03.pdf.
[6] Cohen A., G. Mélard et Ouakasse, A. (2003b), Une expérience de télé-enseignement en statistique pour une banque centrale : aspects technologiques, CoPSTIC'03, Première conférence en sciences et techniques de l'information et de la communication, Université Mohammed V-Agdal et LAB.SIR-Ecole Mohammedia d'Ingénieurs, Rabat, 11-13 décembre 2003, pp 19-22, https://dipot.ulb.ac.be/dspace/bitstream/2013/13834/1/experience_tele_enseignement.pdf
[7] Cooke, D. G., Blackwell, L. F. and Brown, S. (2016), A graphical trap for unwary users of Excel 2010, *International Journal of Open Information Technologies* 4(2), 7-10.
[8] Dell'Omodarme M. and Valle G. (2006), Teaching statistics with Excel and R, http://arxiv.org/abs/physics/0601083.
[9] Fraser, C. (2013), *Business statistics for competitive advantage with Excel 2013: Basics, model building, simulation and cases*, Springer, New York.
[10] Freeman, G. L. (2014), Microsoft Excel 2010 improved for teaching statistics but caution advised, *National Social Science Journal* 43(1), 21-32.
[11] Georgin, J. (2002), *Analyse interactive des données (ACP, AFC) avec Excel 2000: Théorie et pratique*, Presses Universitaires de Rennes, Rennes.
[12] McCullough, B. D. (2008a), Editorial: Special section on Microsoft Excel 2007, *Computational Statistics and Data Analysis* 52, 4568-4569.
[12] Mélard, G. (1990), *Méthodes de prévision à court terme*, Editions de l'Université de Bruxelles, Bruxelles et Ellipses Edition Marketing, Paris.
[13] Mélard, G. (2006), Initiation à l'analyse des séries temporelles et à la prévision, *Revue Modulad* 35, 82-129.
[14] Mélard, G. (2007), *Méthodes de prévision à court terme*, 2e édition, Editions de l'Université de Bruxelles, Bruxelles et Ellipses Edition Marketing, Paris (avec CD-Rom).
[15] Mélard, G. (2010), Utiliser un tableur dans l'enseignement de statistique. Pourquoi et comment ?, Colloque international francophone d'enseignement de la statistique, Bruxelles, 8-10 septembre. http://homepages.ulb.ac.be/~gmelard/rech/Brux2010.pdf.
[16] Mélard G. (2014), On the accuracy of statistical procedures in Microsoft Excel 2010, *Computational Statistics*, 29 (5), 1095-1125.
[17] Microsoft (2013), Changes in Office 2013. [http://technet.microsoft.com/en-us/library/cc178954.aspx, accédé 29 janvier 2014
[18] Microsoft (2017), What's new in Excel 2016 for Windows, https://support.office.com/en-us/article/What-s-new-in-Excel-2016-for-Windows-5fdb9208-ff33-45b6-9e08-1f5cdb3a6c73#Audience=Office_365_subscribers, accédé 29 juin 2017
[19] Nash J. C. (2008), Teaching statistics with Excel 2007 and other spreadsheets, *Computational Statistics and Data Analysis* 52, 4602-4606.
[20] Pupion, P. (2008), *Statistiques pour la gestion: Applications avec Excel et SPSS*, Dunod, Paris.
[21] Quirk, T. J. (2015), *Excel 2013 for business statistics: A guide to solving practical problems*, Springer International Publishing, Cham, Switzerland.
[22] Quirk, T. J. (2015), *Excel 2013 for social sciences statistics: A guide to solving practical problems*, Springer International Publishing, Cham, Switzerland.
[23] Quirk, T. J., Quirk, M., Horton, H. F. (2015), *Excel 2013 for biological and life sciences statistics: A guide to solving practical problems*, Springer International Publishing, Cham, Switzerland.
[24] Salkind, N. J. (2007), *Statistics for people who (think they) hate statistics*. SAGE Publications, Thousand Oaks.
[25] Tufte (1983), *The visual display of quantitative information*, Graphic Press, Cheshire, 1983.
[26] Vidal, A. (2010), *Statistique descriptive et inférentielle avec Excel: Approche par l'exemple*, Presses universitaires de Rennes, Rennes.

ANNEXES

Annexe A. Critiques relatives aux tableurs

Les critiques des différentes versions d'Excel portaient sur :

- la précision des fonctions statistiques ;
- le générateur de nombres aléatoires ;
- la qualité des outils complémentaires ;
- les graphiques ;
- le module Solver.

A titre d'exemple, Excel 97, 2000 et 2002 calculent la variance de 3 entiers consécutifs comme indiqué dans le tableau 1. Quand la moyenne est de l'ordre de cent millions, la variance devient fausse. Ce qui est en cause (Sawitzki, 1994) : l'algorithme « calculatrice » basé sur la formule théoriquement correcte mais imprécise :

$$s_n^2 = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - \overline{x}^2 .$$

TABLEAU 1 – La variance de 3 nombres entiers consécutifs dans Excel 2002

	A	B	C	D	E	F	G	H	I	J	K
1		1 000	10 000	100 000	1 000 000	10 000 000	77 490 000	77 500 000	100 000 000	770 000 000	771 000 000
2		+	+	+	+	+	+	+	+	+	+
3	0	1000	10000	100000	1000000	10000000	77490000	77500000	100000000	770000000	771000000
4	1	1001	10001	100001	1000001	10000001	77490001	77500001	100000001	770000001	771000001
5	2	1002	10002	100002	1000002	10000002	77490002	77500002	100000002	770000002	771000002
6											
7	Variance	0,666667	0,666667	0,666667	0,666667	0,666667	0,666667	0,000000	0,000000	0,000000	85,333333
8	Ecart-type	0,816497	0,816497	0,816497	0,816497	0,816497	0,816497	0,000000	0,000000	0,000000	9,237604

De façon générale, jusqu'à Excel 2002, la plupart des fonctions et procédures statistiques étaient numériquement déficientes en présence de données extrêmes. Pendant plusieurs années Microsoft a ignoré ces critiques ce qui a déplu à la communauté statistique. Les premières améliorations (timides) sont apparues dans Excel 2003. Il a fallu attendre cette version pour que le calcul de la variance soit correct, en employant la définition au lieu de l'algorithme de la calculatrice. Pour l'optimisation non linéaire : le module Solver fournit des messages de convergence atteinte, souvent erronés (McCullough et Heiser, 2008) et ne parvient pas à résoudre de nombreux problèmes de test (Almiron *et al.*, 2010).

Une retombée positive du manque de réaction de Microsoft a été le développement par les statisticiens (en commençant par McCullough et Wilson, 1999) d'une technologie pour juger de la précision des résultats, de manière similaire à ce qui existait pour évaluer la qualité des logiciels statistiques. D'autres développeurs (voir le paragraphe A.2) ont mis l'accent sur la qualité des fonctions et procédures statistiques. D'autres tableurs ont aussi été développés, comme OpenOffice, LibreOffice et Gnumeric. L'article d'Almiron *et al.*. (2010) examine aussi d'autres tableurs (Gnumeric, NeoOffice, Oleo) et plusieurs environnements, montrant notamment qu'Office 2008 pour Apple MacOS-X se comporte différemment d'Office 2007 pour Microsoft Windows.

Le tableau 2, tiré de Yalta (2008), illustre des différences entre plusieurs versions d'Excel. Nous ne discuterons pas de la colonne intitulée ELV (voir Knüsel, 1998).

TABLEAU 2 – Tableau 2 de Yalta (2008) probabilités binomiales cumulées en k pour un exposant n = 1030 et une proportion p = 0,5.

k	EXACT	ELV Ed.2	EXCEL 97/2K/XP	EXCEL 2003/2007
1	8.96114E−308	0	8.95245E−308	0
2	4.61499E−305	0	Exact	0
100	1.39413E−169	0	Exact	0
200	5.45781E−92	Exact	Exact	0
300	2.91621E−42	Exact	Exact	0
390	3.18196E−15	Exact	Exact	0
391	5.24099E−15	Exact	Exact	2.05902E−15
400	3.89735E−13	Exact	Exact	3.86553E−13
410	3.19438E−11	Exact	Exact	3.19406E−11
420	1.76037E−09	Exact	Exact	Exact
500	1.83106E−01	Exact	#NUM!	Exact
550	9.86550E−01	Exact	#NUM!	Exact
575	9.99920E−01	Exact	#NUM!	Exact
589	9.99998E−01	Exact	#NUM!	Exact

La question naturelle est : "y a-t-il amélioration dans Excel 2010 et les versions suivantes Excel ?". C'est là l'essentiel de l'étude de Mélard (2014) qui envisage Excel 2010 et aussi OpenOffice Calc 3.3 et Gnumeric 1.7.11. Mélard (2014) présente plus de détails sur les considérations des paragraphes A.2 à A.6. Ici nous étendons ces résultats à Excel 2013 et 2016 et à des versions plus récentes du module Calc d'OpenOffice et LibreOffice, respectivement les versions 4.1.3 et 5.2.7.

A.1 Les tableurs considérés

1. Microsoft Office Excel 2010, 2013 et 2016

Parmi les avantages du point de vue statistique pour la version 2010, l'éditeur annonce : (i) amélioration du complément Solver ; (ii) précision des fonctions améliorée et apparition de nouvelles fonctions ; (iii) graphiques améliorés (nombre de points augmenté, mise en forme plus rapide, enregistrement de macros). Nous allons examiner ce qu'il en est en réalité. Rien de neuf n'est signalé pour les versions 2013 et 2016.

2. OpenOffice Calc 4.1.3

Il s'agit d'un logiciel libre basé sur la suite StarOffice, rachetée par Sun MicroSystems, sous le nom OpenOffice.org (OOo), diffusée ensuite par Oracle sous le même nom et maintenant par Apache sous le nom Apache OpenOffice (AOO).

3. LibreOffice Calc 5.2.7

La communauté OpenOffice s'est détachée d'Oracle pour fonder ce nouveau produit libre appelé LibreOffice (LO).

4. Gnumeric

Ce tableur est également un logiciel libre avec un accent statistique prononcé. Malheureusement, la version pour Windows n'est plus disponible auprès de son éditeur. On peut le trouver par exemple à l'emplacement http://download.cnet.com/Gnumeric/3000-2077_4-10968476.html dans la version 1.10.16 que Mélard (2014) avait testée.

A.2 Précision des fonctions statistiques

Nous nous basons sur les tests réalisés par Yalta (2008). Avant de recevoir son classeur original, nous avons recréé un classeur Excel avec ses différents calculs. Le tableau 3 présente les résultats synthétiques. Pour plus de détails ainsi que pour une explication détaillée de la procédure utilisée,

voir Mélard (2014) où les résultats sont présentés en termes de nombres de chiffres corrects. La ligne "Nombre de cas" rappelle le nombre de lignes dans le tableau correspondant de Yalta (2008). Les nombres des lignes suivantes montrent le nombre de lignes pour lequel le logiciel indiqué dans la première colonne fournit le résultat correct (à la précision affichée dans la table). Les nombres en grasses indiquent un succès complet pour la fonction considérée. Par rapport à Mélard (2014) qui traite les anciennes versions d'Excel et d'OpenOffice (OOo), les résultats sont complétés par ceux d'Excel 2016, OpenOffice (AOO) 4.1.3 et LibreOffice (LO) 5.2.7.

TABLEAU 3 – Synthèse des résultats corrects pour les tableaux de Yalta (2008)

Tableau	2	3	4	5	6	7	8	9	10
Nombre de cas	14	10	14	10	14	15	14	15	12
Excel 2003	5	4	6	5	12	8	1	8	7
Excel 2007	5	4	6	5	12	8	1	8	7
Excel 2010	**14**	**10**	**14**	**10**	**14**	14	**14**	**15**	**12**
Excel 2013/2016	**14**	**10**	**14**	**10**	**14**	14	**14**	**15**	**12**
OOo Calc 3.0	**14**	**10**	10	**10**	**14**	**15**	13	5	6
OOo Calc 3.3	**14**	**10**	12	**10**	**14**	**15**	**14**	**15**	**12**
AOO Calc 3.4.1	**14**	**10**	12	**10**	**14**	**15**	**14**	**15**	**12**
LO Calc 5.2.7	**14**	**10**	12	**10**	**14**	**15**	**14**	**15**	**12**
Gnumeric	**14**	**10**	**14**	**10**	**14**	**15**	**14**	**15**	**12**

Comme on peut le voir, Excel 2010 réalise une amélioration importante par rapport à Excel 2003 et Excel 2007 pour ce qui est de la précision de la plupart des fonctions statistiques. Par exemple, celui du tableau 2 ci-dessus, Yalta (2008, tableau 2), montre que pour les probabilités cumulées de la loi binomiale, Excel 2007 a eu seulement 5 succès sur 14 alors qu'Excel 2010 trouve les valeurs correctes dans chacun des 14 cas. L'amélioration est moins importante dans le tableau 6 de Yalta (2008) pour les quantiles d'une loi normale, où les anciennes versions se comportaient déjà bien. En fait la nouvelle version atteint le score maximum pour tous les tableaux excepté le tableau 7 de Yalta (quantiles d'une loi χ^2) pour laquelle un cas est faux.

Les lignes 4 et 5 du tableau 3 sont consacrées aux scores de OpenOffice Calc 3.0 et 3.3, respectivement. Calc 3.0 s'est bien comporté mais échoue dans les tableaux 4 (probabilités de Poisson), 9 (quantiles de la loi de Student) et 10 (quantiles de la loi de Fisher-Snedecor) de Yalta (2008). Les résultats de Calc 3.3 montrent une amélioration dans ces secteurs. Calc 3.3 échoue seulement dans le tableau 4 de Yalta (2008) où Excel 2010 semble maintenant excellent.

Les nouvelles versions d'Excel 2013 et 2016, d'une part, et d'OpenOffice 4.1.3 et de LibreOffice 5.2.7, d'autre part, donnent les mêmes résultats, respectivement, qu'Excel 2010 et qu'OpenOffice 3.4.1. Quand on entre les calculs des tableaux de Yalta (2008) dans Gnumeric, on trouve chaque fois les résultats exacts.

Pour conclure, la plus grande partie des problèmes d'imprécision d'Excel 2007 relevés par Yalta (2008) ont été corrigés dans la version 2010, presque aussi bonne qu'OpenOffice.org Calc 3.3. Notons l'ajout de nouvelles fonctions avec des noms plus explicites. Evidemment cela peut poser des problèmes d'incompatibilité quand on ouvre un classeur d'Excel 2010 dans une ancienne version du logiciel. Les noms français sont différents des noms anglais mais il y a toujours une conversion automatique quand on passe d'une version linguistique à une autre. Pour des raisons de compatibilité, Microsoft a maintenu les anciens noms. Voyons maintenant si les améliorations ont aussi porté sur les autres registres.

A.3 Générateur de nombres pseudo-aléatoires

A.3.1 Générateurs d'Excel 2010-2016

La discussion porte surtout sur le générateur invoqué par la fonction ALEA(), mais il y a aussi deux autres générateurs : celui fourni par l'outil complémentaire et le générateur du Visual Basic for Application (VBA). Ce dernier n'a pas changé et est connu pour être de mauvaise qualité, L'Ecuyer et Simard (2007). Dans Excel 2007, la fonction ALEA() était sensée employer l'algorithme de Wichmann-Hill (1982) mais McCullough (2008b) a démontré une erreur d'implantation de l'algorithme.

Microsoft a déclaré avoir amélioré la fonction ALEA dans Excel 2010. Il semble que le nouvel algorithme suive la suggestion de plusieurs auteurs (McCullough et Wilson, 2005, McCullough, 2008b) d'employer l'algorithme Mersenne Twister dont la période est $2^{19937} - 1$. Mélard (2014) a réalisé une étude qui semble confirmer les qualités du nouveau générateur mais que les générateurs (B) et (C) sont inchangés par rapport à Excel 2007 et les versions précédentes mais il indique comment utiliser la fonction ALEA() dans les macros VBA. Voir aussi les critiques de Botchkarev (2015) qui regrette que la nouvelle fonction ALEA() (RAND() en anglais) ne soit pas bien documentée et aussi qu'il ne soit pas possible pour l'utilisateur de poser la semence initiale. Il recommande le module VBA "MersenneTwisterVBAModule" disponible auprès des auteurs de l'algorithme, Matsumoto et Nishimura (1998).

A.3.2 OpenOffice Calc 4.1.3 et LibreOffice 5.2.7

Mélard (2014) a examiné le générateur d'OpenOffice 3.3 qui, dans sa version sous Windows, fournit des nombres avec une précision de 15 bits, soit 32768 valeurs différentes seulement. Entretemps, cela a été amélioré avec le passage à un algorithme Mersenne Twister dans OpenOffice 4.0 ainsi que dans LibreOffice 4.0.

A.3.3 Gnumeric

Gnumeric emploie un algorithme Mersenne Twister depuis plus de dix ans.

A.4 Procédures statistiques

A.4.1 Procédures statistiques d'Excel 2010-2016

Dans Excel 2007 et suivants, après installation, les procédures statistiques sont disponibles par le menu Outils > Analyse de données, voir figure 1. L'apparence a un peu changé mais pas les procédures elles-mêmes.

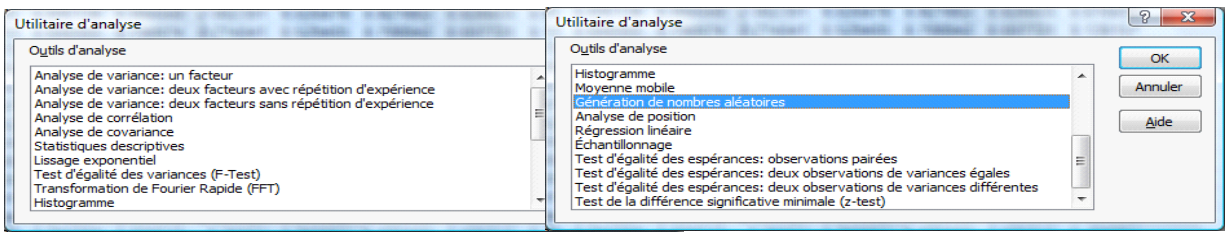


FIGURE 1 – La liste des outils complémentaires d'Excel 2007 et suivants

Les outils complémentaires d'Excel n'ont jamais beaucoup attiré l'attention pour l'enseignement. Pour les cours introductifs, il manque les tests sur une proportion, une moyenne ou une variance mais on y trouve les tests d'égalité de moyennes dans différents cas (y compris celui des variances inégales) et un test d'égalité de variances et l'analyse de variance à un et deux facteurs. La procédure de lissage exponentiel est surprenante : le complément à l'unité de la constante de lissage est employé, voir

McCullough et Heiser (2008). C'est en fait le facteur d'escompte (« discount factor ») de Brown (1962). McCullough et Heiser (2008) sont aussi surpris que les valeurs lissées soient décalées d'un point vers la droite mais c'est parce que Microsoft les a placées comme des prévisions pour le temps suivant, voir Mélard (2007, p. 144). Les prévisions ex post ne sont pas calculées, même pas pour un horizon de 1. Les moyennes mobiles sont correctes mais placées à la date de la dernière observation, ce qui est appelé les moyennes mobiles en finance par Mélard (2007, p. 96). Les prévisions ex post d'horizon supérieur à 1 ne sont pas calculées. Notons que, exceptionnellement pour les procédures statistiques d'Excel, les résultats sont donnés sous forme de formules.

Alors que les fonctions TTEST (3 cas) et FTEST fonctionnent correctement (comme dans la version 2003), le test d'égalité des moyennes, cas des observations appariées, est toujours erroné en présence de données manquantes (cellules vides) comme le montre la figure 2 basée sur Heiser (2009). Il faut insister sur le fait que cette constatation était déjà émise par Simon (2000) ! Le test d'égalité des moyennes, cas d'observations indépendantes et variances inégales, est parfois erroné parce que le nombre de degrés de liberté de la méthode de Welsch est arrondi (contrairement à la fonction TTEST où il est fractionnaire). Au risque de nous répéter, ceci n'est pas nouveau.

L'outil complémentaire de régression linéaire est relativement meilleur depuis la version 2003, sauf en cas de colinéarité. Il ne permet toujours pas plus de 16 variables explicatives comme le montre la figure 3. Notons que la fonction (mal-nommée en français) DROITEREG accepte plus de 16 variables. Il n'y a pas d'amélioration dans les graphiques optionnels qui montrent les résidus ou les valeurs ajustées en fonction de chaque variable explicative. Non seulement ils sont étriqués, comme le montre la figure 4, mais ils sont toujours aussi peu utilisables. Le diagramme de répartition de probabilités sous le titre "Probabilité normale" n'a rien de normal, McCullough et Heiser (2008). Dans la version 2007, la documentation de la procédure de régression linéaire était toujours très mauvaise, employant un langage inapproprié. Microsoft a résolu le problème en réduisant la documentation à quelques lignes pour chaque procédure.

D.A. Heiser, section XVII

	Set 1	
	A	B
1	3	2
2	4	
3	3	2
4		3
5	2	3
6	4	3
19	4	2

TTEST(A,B,2,1) 0.036878

Test d'égalité des espérances: observations pairées

	Variable 1	*Variable 2*
Moyenne	3.21052632	2.57894737
Variance	0.61988304	0.47953216
Observations	19	19
Coefficient de corrélation de Pearson	-0.17699808	
Différence hypothétique des moyenne	0	
Degré de liberté	18	
Statistique t	1.71428571	
P(T<=t) bilatéral	0.10364302	

FIGURE 2 – Exemple de Heiser comparant l'outil complémentaire d'Excel 2010 et suivants pour la comparaison des moyennes en présence de données manquantes et la fonction TTEST

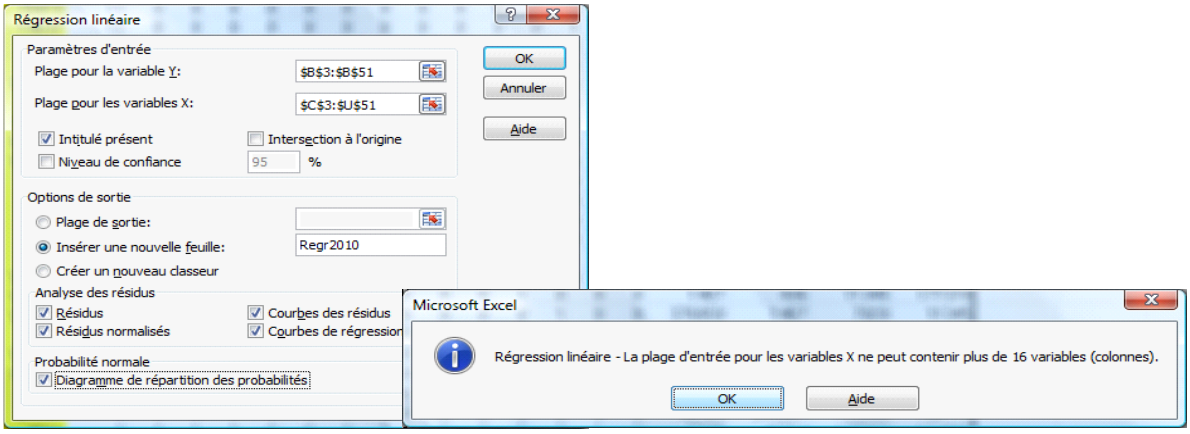


FIGURE 3 – Exemple montrant que l'outil complémentaire d'Excel 2010 et suivants pour la régression linéaire ne permet pas plus de 16 variables explicatives

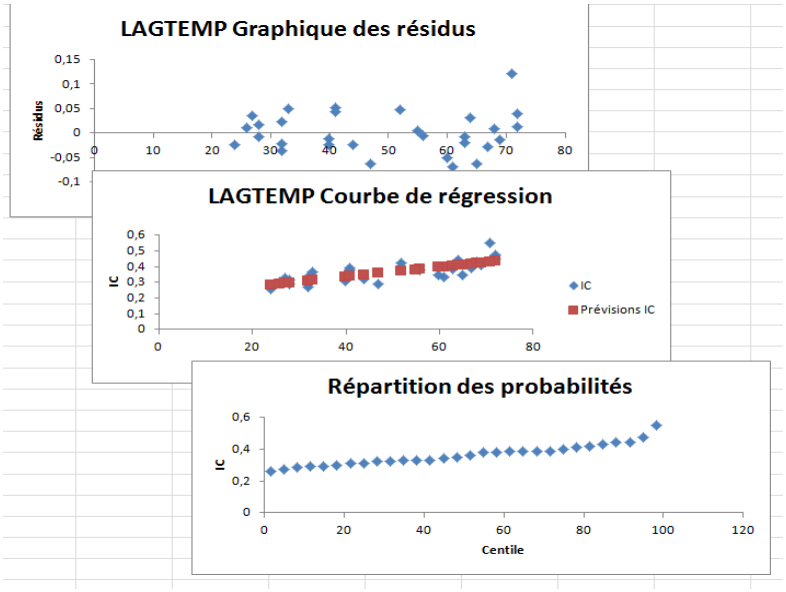


FIGURE 4 – Exemple montrant les graphiques de l'outil complémentaire d'Excel 2010 et suivants pour la régression linéaire

Mélard (2014) a aussi mentionné les batteries classiques de jeux de données de test (National Institute of Standards and Technology, 1999) ainsi que les tests de Wilkinson, voir Sawitzki (1994) mis en œuvre sur les versions précédentes par McCullough et Wilson (2005) et Almiron *et al.* (2010), voir aussi Heiser (2009). Cependant, compte tenu des remarques précédentes, il n'y a pas de modifications.

A.4.2 OpenOffice Calc 4.1.3 et LibreOffice 5.2.7

Il n'y a jamais eu d'outil complémentaire semblable dans OpenOffice Calc. Le développement d'outil a été abandonné sauf le projet R4Calc qui emploie R et sort donc du cadre de cet article. En revanche, LibreOffice a développé un ensemble d'outils statistiques similaire à ceux de Microsoft Excel. Il manque seulement l'histogramme, la transformée de Fourier rapide mais surtout la régression linéaire multiple puisque seule la régression linéaire simple est disponible (plus les régressions logarithmique et exponentielle).

Nous n'avons pas testé systématiquement les outils proposés par LibreOffice 5.2.7, nous contentant des procédures similaires à celles discutées ci-dessus, à savoir le lissage exponentiel, les moyennes mobiles et le test de comparaison de moyennes, surtout que les deux premiers sont liés avec le sujet de l'étude de cas du paragraphe 4. Comme pour Excel, c'est le facteur d'escompte qui est employé dans le lissage exponentiel. Les résultats y sont donnés sous forme de formules, comme pour les moyennes mobiles d'ailleurs. Mais le facteur d'escompte est entré numériquement dans ces formules, empêchant toute possibilité d'optimiser sa sélection. L'interface comporte des défauts flagrants pour la saisie des champs. Ni les graphiques, ni les erreurs-types ne sont fournis. L'exemple de Simon (2000) fournit ici les résultats corrects, contrairement à Excel 2016 (voir figure 2).

A.4.3 Gnumeric

Gnumeric comporte de nombreuses procédures statistiques. Elles ont été étudiées par McCullough (2004c) et se sont avérées d'excellente qualité. Leur validité a été attestée par le passage des jeux de données de test déjà mentionnées (National Institute of Standards and Technology, 1999), voir Almiron *et al.* (2010). McCullough (2004c) n'a pas discuté des procédures de lissage exponentiel et de moyenne mobile. Disons simplement que plusieurs méthodes de lissage exponentiel sont proposées, jusqu'à celle de Winters additive et multiplicatives et que plusieurs variantes de moyennes mobiles sont proposées, y compris celle de Spencer d'ordre 15. Nous recommandons la consultation de l'aide en ligne, bien réalisée, pour plus de précisions.

A.5 L'optimisation

A.5.1 Solver d'Excel 2010-2016

Le Solver d'Excel permet de réaliser de l'optimisation linéaire ou non linéaire, sans ou avec contraintes. Il peut avoir différents usages en statistique, en particulier pour la régression non linéaire. Pour les versions précédentes d'Excel, cette question a été étudiée notamment par McCullough et Wilson (2005) et Almiron *et al.* (2010), voir aussi Heiser (2009). Cependant, Microsoft affirme avoir amélioré le Solver dans la version 2010, notamment en ajoutant une méthode Multistart à l'algorithme de base GRG2 et en ajoutant un nouvel algorithme appelé Evolutionary. Aussi Mélard (2014) a effectué une mise à jour de ces études en employant une batterie de problèmes de test (National Institute of Standards and Technology, 1999). Les conclusions sont nettes pour les problèmes de régression non linéaires : il n'y a pas d'amélioration sensible de l'algorithme GRG2 (aucun chiffre correct pour 12 des 27 cas) et les deux nouveaux algorithmes qui reposent sur des spécifications d'intervalles plausibles de variation pour les paramètres ne fonctionnent pas correctement en l'absence de ces spécifications.

Notons que la version du Solver d'Excel 2016 mentionne toujours une date de copyright 2009. Les résultats inchangés mentionnés dans le paragraphe 2 ne sont donc pas étonnants.

A.5.2 OpenOffice Calc 4.1.3 et LibreOffice 5.2.7

La dernière étude pour la version 3.3 d'OpenOffice a été réalisée par Mélard (2014). Les deux algorithmes DEPS et SCO de l'extension "Solver for nonlinear programming" se sont révélées meilleures que le Solver d'Excel 2010. LibreOffice 5.2.7 inclut cette extension par défaut et elle est fonctionnelle bien que l'aide ne soit pas présente. OpenOffice Calc 4.1.3 ne contient pas cette extension par défaut. Elle est disponible sur le site en version 0.9 de 2009, bien que marquée comme de compatibilité inconnue avec les versions récentes d'OpenOffice. L'aide est aussi absente mais elle est disponible à l'emplacement https://wiki.openoffice.org/wiki/NLPSolver. Nous n'avons pas poussé plus loin les investigations.

A.5.3 Gnumeric

La dernière étude pour la version 1.10.16 a été réalisée par Mélard (2014) et, en dépit de petits problèmes, NLSolve s'est révélé meilleur que le Solver d'Excel 2010.

A.6 Graphiques statistiques

A.6.1 Les graphiques d'Excel 2010-2016

Trente ans après Tufte (1983), il est regrettable que les graphiques par défaut d'Excel soient aussi mauvais, pour des données statistiques mais aussi d'autres données. La situation semble empirer avec un accent tridimensionnel prononcé et il faut de plus en plus d'efforts pour éliminer le superflu des graphiques. Avant Excel 2016, on ne trouve pas les graphiques statistiques spécifiques comme les boîtes à moustache. Mélard (2014) met à jour les constations générales de Su (2008) et celles propres aux ajustements réalisés sur base de graphiques de Hargreaves et McWilliams (2010), tous deux sur Excel 2007 mais qui restent d'actualité. Voir aussi les critiques sévères de Cooke *et al.* (2015) pour les graphiques statistiques les plus simples, à savoir les graphiques linéaires, surtout quand ils sont transformés en diagrammes de dispersion. En plus de trois autres nouveaux graphiques, Excel 2016 a proposé trois graphiques statistiques : un histogramme, un diagramme de Pareto et un diagramme de boîtes à moustache. Le diagramme de boîte à moustache possède une option curieuse de changer la définition du premier et troisième quartiles en ignorant la littérature à ce sujet, par exemple Hyndman et Fan (1996). L'exemple de la figure 5 fourni par Microsoft (2015) est relatif aux données suivantes : 1, 2,5, 7, 10, 14, 15.

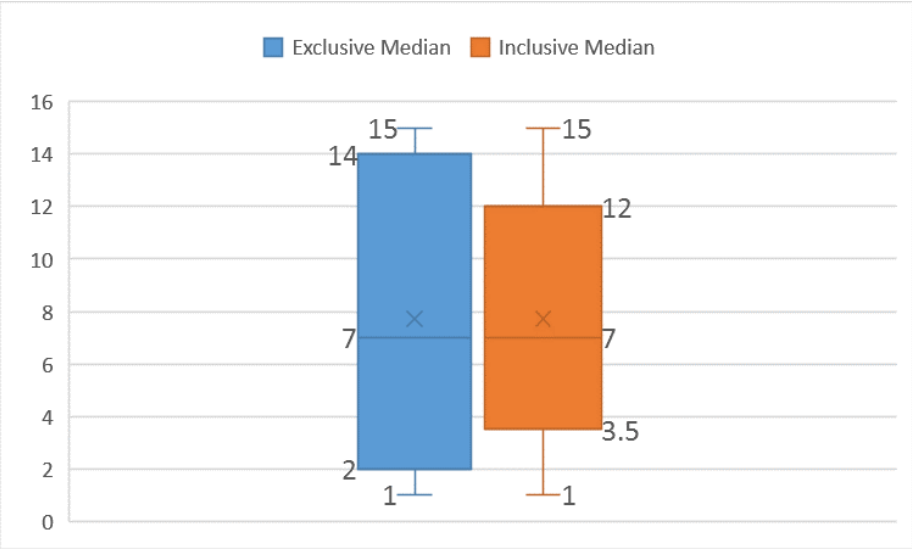


FIGURE 5 – Exemple montrant les deux variantes de boîtes à moustache d'Excel 2016 (source : Microsoft)

A.6.2 OpenOffice Calc 4.1.3 et LibreOffice 5.2.7

Les graphiques de Calc ne sont pas meilleurs du point de vue statistique, même s'ils sont toutefois moins chargés. A ce jour, les boîtes à moustaches ne sont pas disponibles.

A.6.3 Gnumeric

Gnumeric dispose d'un certain nombre de graphiques statistiques. Ils n'ont pas été étudiés par McCullough (2004c) mais s'avèrent à première vue de bonne qualité. Ils sont simples au départ mais peuvent être enrichis. Les boîtes à moustache y sont disponibles depuis longtemps. Il est possible que le nombre de ces graphiques statistiques ait augmenté depuis 2014 mais l'absence de version récente pour Windows ne permet pas de répondre à la question.

Annexe B. Un cours multimédia de méthodes de prévision

L'objectif de cette annexe est d'examiner les classeurs Excel d'un cours d'analyse de séries temporelles en tenant compte des constatations du texte. Le tableau 4 contient la liste des chapitres du CD-Rom de Mélard (2007) et le nombre de classeurs Excel principaux concernés. Nous n'avons pas considéré les nombreux classeurs des exercices supplémentaires qui ne servent qu'à contenir des données.

TABLEAU 4 – *Les chapitres de Mélard* (2007) *et les nombres de classeurs mis en œuvre*

Chapitres	Nombre de classeurs
1. Concepts et définitions	3
2. Régression simple	4
3. Courbes de croissance	5
4. Moyennes mobiles	3
5. Décomposition saisonnière	10
6. Lissage exponentiel	8
7. Régression multiple	9
8. Autocorrélation et stationnarité	3
9. Modèles ARMA	0
10/11. Méthode de Box et Jenkins	0
12/13. Méthodes X12-ARIMA/TRAMO-SEATS (*)	2
Total	47

Annexe C. Examen de quelques classeurs du cours

CH02EX04.xls — Pour commencer, on rappelle la synthèse des données par la médiane et par la moyenne tronquée (qui sera employée lors de la décomposition saisonnière au chapitre 5). On peut ensuite examiner les formules pour calculer les coefficients de la régression linéaire simple par la méthode des moindres carrés, mais aussi par la méthode des deux points (aussi connue sous le nom de Mayer) et par la méthode de Theil.

Dans les instructions, on recommande de visualiser l'effet du changement d'une donnée, et donc d'évaluer la sensibilité de ces méthodes à la présence de données extrêmes. Ces deux méthodes sont activées par des macros parce qu'elles comportent des tris (bien qu'en fait seules les médianes soient requises) et qu'il n'y a pas de fonction de tri simple dans Excel. A titre d'exemple, la figure 6 montre le calcul des pentes parmi les 10 droites joignant les 5 points de données et l'obtention de leur médiane.

$x_i - x_j$	$y_i - y_j$	Pentes	Pentes triées		Médiane
0.9	-6	-6.667	-6.667		
1.1	6	5.455	-2.105		
0.2	12	60.000	0.667		
3.0	2	0.667	1.795		
2.1	8	3.810	2.600	b_1 =	3.2048
1.9	-4	-2.105	3.810		
5.0	13	2.600	4.634		
4.1	19	4.634	5.455		
3.9	7	1.795	5.500		
2.0	11	5.500	60.000		

FIGURE 6 –Exercice CH02EX04.xls : calcul de la pente de la droite de Theil

E28 =E27+MAX(0;ENT(B28-B27+D28-D27))

	A	B	C	D	E	F	G	H
1	alpha	0.1						
2	beta	0.7						
3	gamma	0.001						
4	min	14.0845						
5	max	1000.0000		40	1000			
6	X				Données	Inverses	Logarithmes	Exp.mod.
7	1	14.0845		8.27403376	22	0.045455	3.091042	-31.6142
8	2	20.0000		5.39994466	25	0.040000	3.218876	-9.9911
9	3	28.3286		-5.7912437	25	0.040000	3.218876	13.8074
10	4	39.9840		-1.9450272	40	0.025000	3.688879	40.0000

FIGURE 7 –Exercice CH03EX05.xls : génération des données

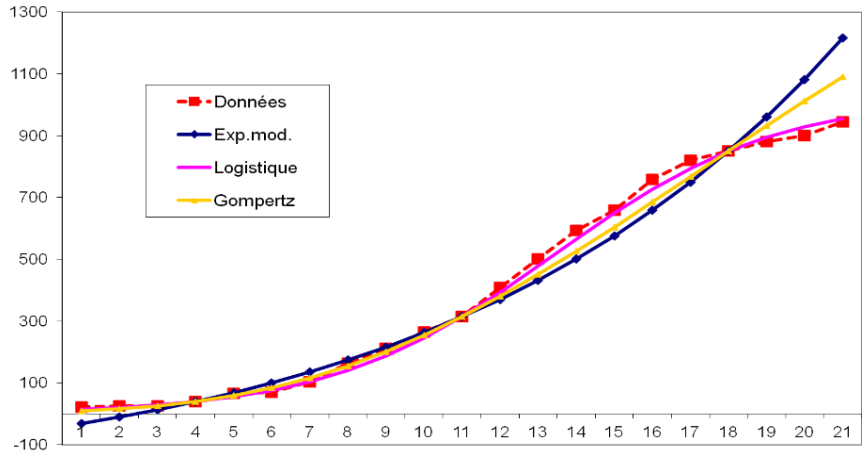


FIGURE 8 –Exercice CH03EX05.xls : ajustement des données par trois courbes de croissance

CH03EX05.xls — On traite ici des courbes de croissance non linéaires. Des données artificielles sont simulées selon une courbe logistique donnée caractérisée par des constantes α, β et γ données, par défaut 0,1, 0,7 et 0,001. Elles simulent des ventes journalières cumulées d'un article déterminé d'une société de ventes par correspondance. Le but est d'estimer le niveau plafond des ventes à partir de 21 jours de commandes de la clientèle. La construction des données artificielles mérite quelques mots d'explication. Partant d'une courbe logistique sur laquelle on trouve des ordonnées $y_t = 1/(\alpha\beta^t + \gamma)$, $t = 0, ..., 21$, et de déviations aléatoires e_t (employant la fonction ALEA), plus précisément une suite de nombres pseudo-aléatoires entre −20 et 20, on a calculé les ventes v_t, $t = 1, ..., 21$, comme suit (voir figure 7) :

- notons $w_t = \max(\lfloor y_t - y_{t-1} + e_t - e_{t-1} \rfloor, 0)$, où $\lfloor x \rfloor$, est la partie entière de x,
- accumulons les w_t : $v_t \leftarrow v_{t-1} + w_t$, $t = 2, ..., 21$.

Il faut évidemment une valeur initiale $v_1 = \max(\lfloor y_1 + e_1 \rfloor, 0)$.

On emploie la méthode des trois points pour réaliser un ajustement non linéaire sur les données simulées. Trois courbes de croissance exponentielle modifiée, courbe de Gompertz et enfin courbe logistique sont employées. On s'attend à ce que cette dernière soit plus appropriée puisqu'elle a servi à générer les données. Chaque pression de la touche de recalcul (la touche F9) permet de générer une nouvelle série et on visionne les trois ajustements, comme dans la figure 8.

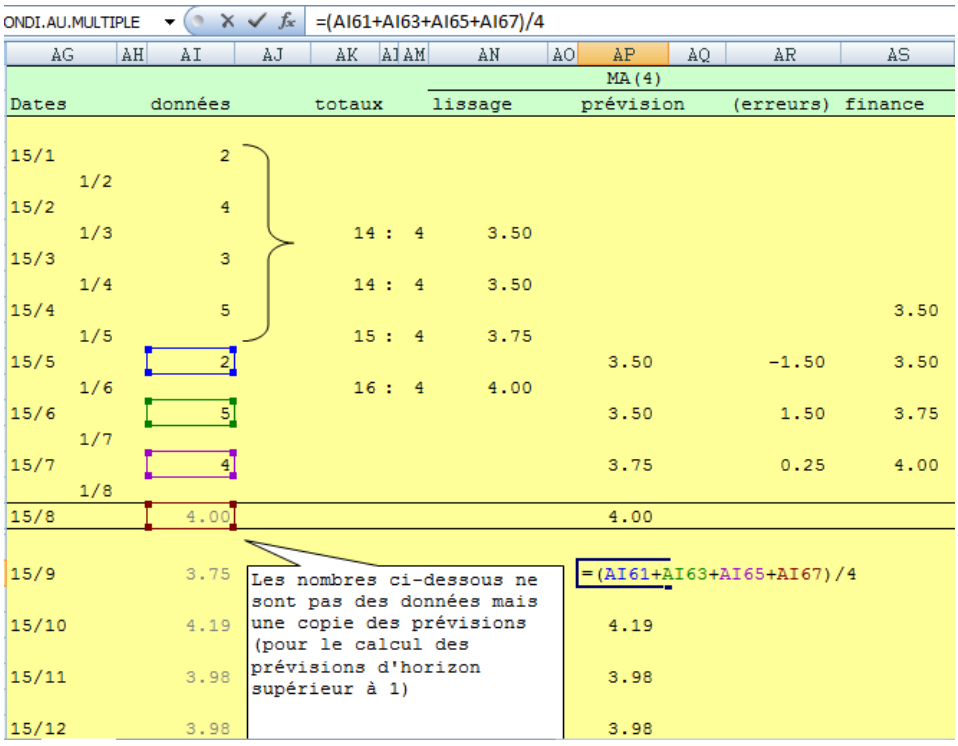


FIGURE 9 –Exercice CH04EX01.xls : calcul des moyennes mobiles de lissage, de prévision et pour l'analyse technique en finance

CH04EX01.xls — Ce classeur sert à montrer la détermination des moyennes mobiles pour le lissage, la prévision et pour l'analyse technique en finance, et ceci sur une série artificielle constituée de petits nombres entiers, de manière à faciliter la vérification mentale des calculs. De plus on peut visualiser ces trois usages des moyennes mobiles, en l'occurrence d'ordre 4, dans un tableau, voir la figure 9, ou dans des graphiques, voir la figure 10. On y voit, grâce à la pression de la touche F2 qui permet d'inspecter les cellules employées dans une formule, la prévision d'horizon 2 calculée comme moyenne de trois données et de la prévision d'horizon 1, préalablement copiée en dessous des données. On peut examiner des moyennes mobile d'ordre pair, à savoir 4, et d'ordre impair, à savoir 3, et ceci en ajoutant successivement une nouvelle observation. Ceci est réalisé à l'aide de plusieurs macro Visual Basic for Applications (VBA) pour l'initialisation, l'ajout d'une nouvelle donnée et le rétablissement des données initiales, voir la figure 11. Cela permet aussi d'illustrer les principales

propriétés des moyennes mobiles de lissage, comme le caractère centré des moyennes mobiles d'ordre impair et non centré de celles d'ordre pair, la perte de $k - 1$ valeurs à cause d'une moyenne mobile d'ordre k, le fait que l'application sur une série chronologique qui est périodique de période k d'une moyenne mobile d'ordre k donne une constante, et quelques autres propriétés moins importantes. Pour certaines de ces propriétés on demande à l'utilisateur de changer les données, par exemple remplacer l'avant-dernière donnée 5 par 4, de manière à avoir une périodicité de période 4 : 2, 4, 3, 5, 2, 4, 3. D'où l'importance de pouvoir rétablir les données d'origine. Dans ce chapitre comme dans les chapitres suivants, on signale évidemment l'avantage de copier une formule et de la coller dans une plage. Cela permet aisément de réaliser des formules de récurrence.

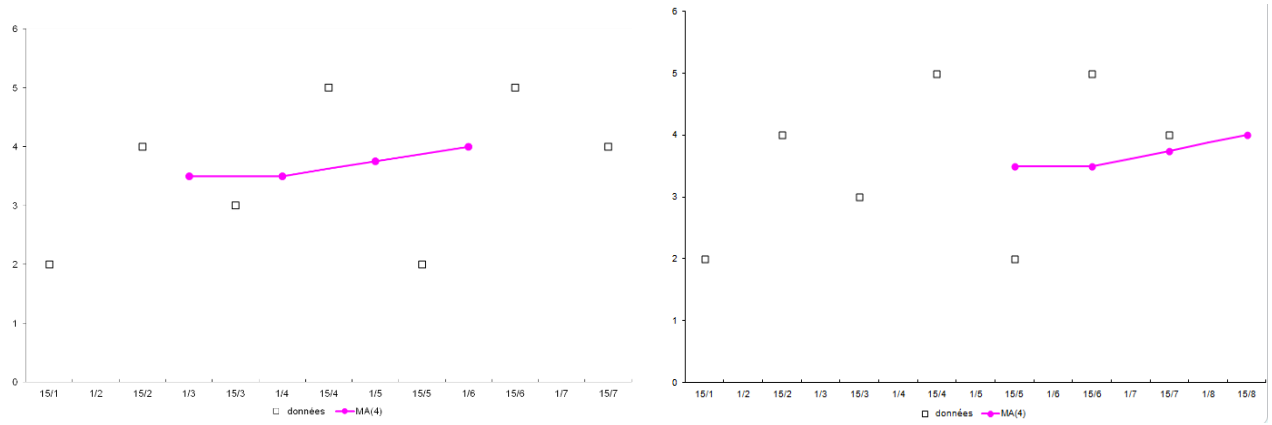


FIGURE 10 –Exercice CH04EX01.xls : graphique des moyennes mobiles d'ordre 4 de lissage et de prévision

J10 =G7+I7+K7+M7

METHODES DE PREVISION A COURT TERME, par GUY MELARD (c) 1991, 1999
Chapitre 4, tableaux 4.1 (en animation), et 4.2. Moyennes mobiles d'ordre 4, MA(4)

dates	15/1	15/2	15/3	15/4	15/5	15/6	15/7	dates
	1/2	1/3	1/4	1/5	1/6	1/7	1/8	
données	2	4	3	5	2	5		données
sommes mobiles		14.0	14.0	15.0				sommes mobiles
moyennes mobiles		3.50	3.50	3.75				moyennes mobile

ANIMATION Pressez toujours CTRL SHIFT A pour démarrer l'animation.
Tant que FIN n'apparaît pas ci-dessus, pressez CTRL SHIFT R .
On y voit les données apparaître à droite et disparaître à gauche, pendant que les totaux sur 4 mois se calculent, ainsi que les moyennes mobiles d'ordre 4, MA(4).
Ne pas quitter la ligne des sommes mobiles.
Pour réinitialiser les données, pressez CTRL SHIFT I

CTRL SHIFT A
CTRL SHIFT R
CTRL SHIFT I

FIGURE 11 –Exercice CH04EX01.xls : calcul progressif des moyennes mobiles de lissage d'ordre 4

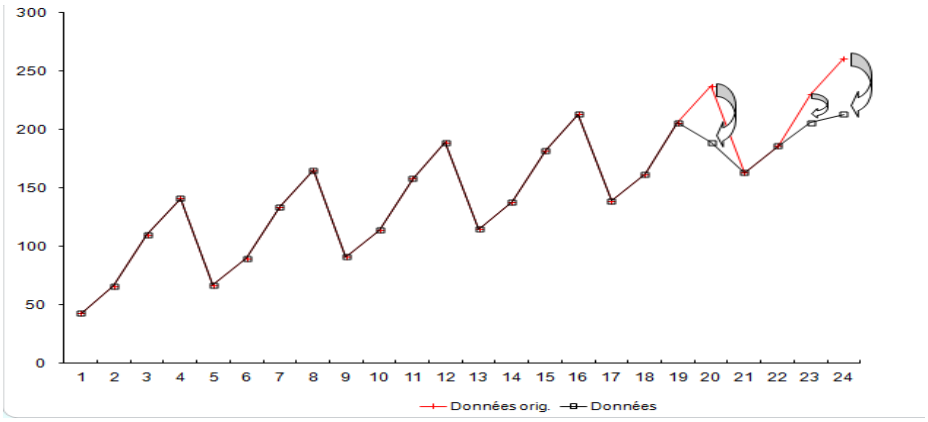


FIGURE 12 –Exercice CH05EX05.xls : données artificielles employées pour la décomposition saisonnière

Comparaisons à la tendance linéaire

	Année	I	II	III	IV	
	1980	-43.35	-25.45	13.45	39.35	
	1981	-39.75	-21.85	17.05	42.95	
	1982	-36.15	-18.25	20.65	46.55	
	1983	-32.55	-14.65	24.25	50.15	
	1984	-28.95	-11.05	27.85	5.75	
	1985	-25.35	-7.45	7.45	9.35	Somme
Coefficients saisonniers		-34.35	-16.45	18.45	32.35	0

FIGURE 13 –Exercice CH05EX05.xls : calcul des coefficients saisonniers sur les données de la figure 12 par la méthode de comparaison à la tendance linéaire

CH05EX05.xls — Ce classeur réalise l'application de plusieurs méthodes élémentaires d'ajustement saisonnier sur une série artificielle qui simule une série trimestrielle avec une saisonnalité marquée pendant les premières années et comportant quelques irrégularités vers la fin, voir la figure 12. On emploie un modèle additif $y = T + C + S + E$, où T représente la tendance, C , le cycle conjoncturel, S, le saisonnier et E, les erreurs. Les méthodes illustrées sont (1) la méthode de décomposition par comparaison à la moyenne générale, (2) la méthode de comparaison à la moyenne annuelle, (3) la méthode de comparaison à la tendance linéaire et (4) la méthode de comparaison aux moyennes mobiles centrées[1] d'ordre 4. La tendance linéaire T est déterminée (en employant des fonctions d'Excel) sur les moyennes annuelles. Pour le calcul des valeurs de tendance de chaque trimestre nécessaires pour la méthode (3), une difficulté supplémentaire est qu'il faut calculer le bon temps t. Si la tendance a été déterminée aux temps $t = 1, 2, ...$, les temps t à employer sont $-0{,}375$, $-0{,}125$, $0{,}125$, ... La figure 13 montre, dans ce qu'on appelle un tableau de Buys-Ballot, les différences $y - T$ qui, en l'absence de cycle, doivent représenter $S + E$. La synthèse des variations saisonnières est effectuée par moyenne, et aussi par médiane pour la méthode (4). Outre la détermination des coefficients saisonniers, le calcul des données corrigées des variations saisonnières et l'obtention de la série des résidus sont illustrées dans le cas des deux dernières méthodes. Enfin le calcul des prévisions est discuté pour chacune des méthodes. Compte tenu de la forme de la série, les trois premières méthodes ne conviennent pas bien, voir par exemple la figure 13 qui montre que les variations saisonnières évoluent de manière assez systématique d'une année à l'autre au lieu d'être presque constantes. Il apparaît que seule la méthode de comparaison aux moyennes mobiles donné des résultats satisfaisants mais à condition d'employer la synthèse des variations saisonnières par médiane, voir la figure 14, au lieu de la moyenne employées dans les autres cas.

	Année	I	II	III	IV
	1980			0	0
Erreurs	1981	0	0	0	0
(arrondies)	1982	0	0	0	0
	1983	0	0	0	0
	1984	0	6	12	-36
	1985	15	18		
Variance :		100.69			

FIGURE 14 –Exercice CH05EX05.xls : détermination des erreurs après calcul des coefficients saisonniers par la méthode de comparaison aux moyennes mobiles avec synthèse par médiane

[1] Une moyenne mobile centrée d'ordre 4 est une moyenne mobile d'ordre 2 d'une moyenne mobile d'ordre 4. On l'emploie parce que les moyennes mobiles d'ordre 4 (4 étant le nombre de trimestres par an) ne sont pas placées aux dates des observations.

Année	JAN	FEV	MARS	AVR	MAI	JUIN	JUIL	AOUT	SEP	OCT	NOV	DEC
1962							0.661	*0.644*	0.851	1.240	*1.634*	1.998
1963	0.704	0.684	0.835	0.891	*1.016*	0.847	0.779	0.447	0.899	1.104	1.681	2.035
1964	0.750	0.724	*0.978*	0.845	0.931	*0.927*	0.736	0.347	*0.781*	1.146	1.651	*1.986*
1965	*1.144*	0.654	0.779	*0.934*	0.928	0.916	0.741	0.334	0.950	*1.088*	1.669	2.133
1966	0.724	*0.853*	0.823	0.804	0.885	0.890	0.736	0.320	0.937	*1.279*	1.815	2.083
1967	0.741	0.732	0.833	0.789	0.910	0.834	*0.623*	0.328	0.956	1.274	*2.044*	*2.691*
1968	*0.510*	*0.558*	*0.649*	*0.723*	*0.572*	*0.791*	*0.833*	0.339	1.009	1.223	1.830	2.377
1969	0.708	0.568	0.766	0.828	0.883	0.862	0.819	*0.291*	*1.040*	1.217	1.720	2.212
1970	0.759	0.624	0.804									
Provisoire	0.731	0.664	0.806	0.831	0.907	0.870	0.745	0.352	0.933	1.201	1.728	2.139
Définitif	0.737	0.669	0.813	0.838	0.914	0.876	0.751	0.355	0.941	1.210	1.741	2.156

FIGURE 15 –Exercice CH05EX06.xls : calcul des coefficients saisonniers sur les données de ventes de champagne en France par la méthode de comparaison aux moyennes mobiles avec synthèse par moyenne élaguée

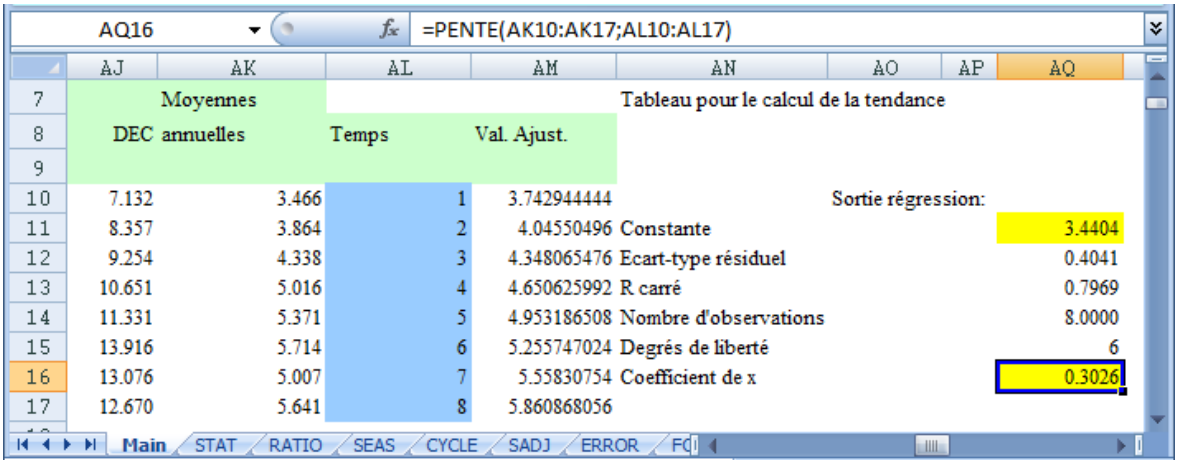


FIGURE 16 –Exercice CH05EX06.xls : calcul de droite de tendance sur les moyennes annuelles des données de ventes de champagne en France

CH05EX06.xls — Le classeur porte sur l'étude des ventes mensuelles de champagne en France entre 1962 et 1970. On commence par discuter du choix entre modèle additif et modèle multiplicatif. On choisit le modèle multiplicatif qu'on utilisera dans le reste de l'exercice. On extrait alors la composante saisonnière par la méthode de comparaison aux moyennes mobiles (centrées d'ordre 12), donc les rapports $y/(T\times C)$. Cette fois la synthèse du tableau de Buys-Ballot montré dans la figure 15 est réalisée par moyenne élaguée au lieu de la moyenne ou de la médiane employées dans les autres exercices. On procède à la détermination des autres composantes : la tendance, le cycle conjoncturel, la composante d'erreurs ainsi que la série corrigée des variations saisonnières. Comme dans l'exercice précédent, une partie délicate des calculs consiste à calculer les valeurs de tendance pour chaque mois mais à partir de la droite de tendance déterminée par la méthode des moindres carrés à partir des moyennes annuelles, avec recours à des fonctions d'Excel, voir figure 16. On termine par l'examen des erreurs dans la figure 17. Deux erreurs importantes sont interprétables : janvier 1968 a vu l'introduction de l'introduction de la taxe à la valeur ajoutée en France, avec un taux de luxe de 33% pour le champagne ce qui a produit une augmentation de prix anticipée par les consommateurs; mai 1968 avec les grèves étudiantes et la grève générale qui a suivi.

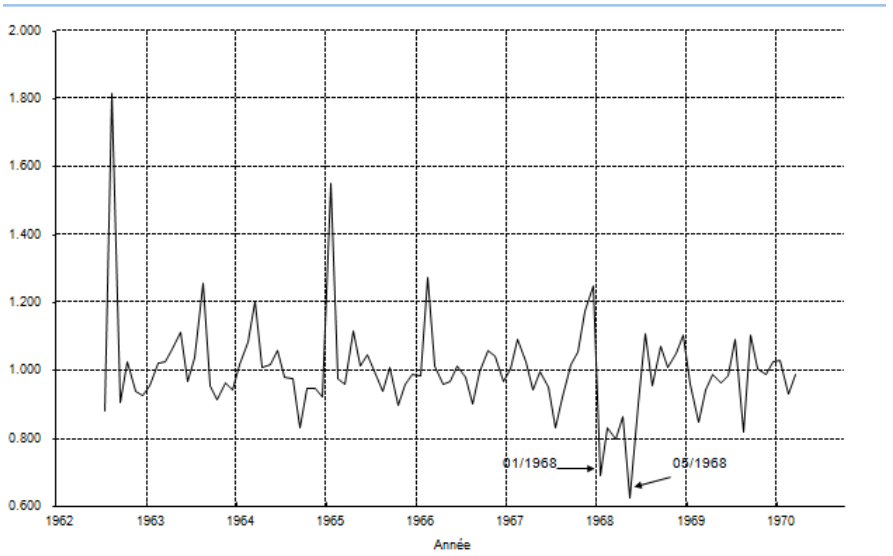


FIGURE 17 –Exercice CH05EX06.xls : graphique des résidus en fonction du temps sur les données de ventes de champagne en France

CH05EX09.xls — Il s'agit à nouveau des ventes de mensuelles de champagne en France, mais cette fois la partie de la série employée pour la détermination des facteurs saisonniers s'arrête en décembre 1969 et l'années 1970 est réservée pour la comparaison avec les prévisions. A part cela, on procède comme dans l'exercice précédent avec un modèle multiplicatif[2]. Dans la figure 18, on voit les différentes composantes : la tendance, le cycle conjoncturel, les facteurs saisonniers (dont 0,733 pour janvier), les erreurs ainsi que les données corrigées des variations saisonnières ou désaisonnalisées. On montre des prévisions pour l'année 1970, y compris la détermination d'un intervalle de prévision à 80 %, voir la figure 19.

G33 =D33/F33

	A	B	C	D	E	F	G	H	I	J	K	L	M
24	Année	Mois	Numéro	Ventes	MA(12)	CMA(12)	Rapport	Tendance		Cycle	Saisonnier	Résidu	TxCxS
25													
27	1962	Janvier	1	2.851				3.604			0.733		
28		Février	2	2.672				3.629			0.680		
29		Mars	3	2.755				3.655			0.816		
30		Avril	4	2.721				3.680			0.840		
31		Mai	5	2.946				3.705			0.917		
32		Juin	6	3.036				3.730			0.879		
33		Juillet	7	2.282	3.466	3.453	0.661	3.756		0.920	0.738	0.895	2.549
34		Août	8	2.212	3.440	3.432	0.644	3.781		0.908	0.363	1.776	1.245
35		Septembre	9	2.922	3.424	3.435	0.851	3.806		0.903	0.928	0.916	3.189
36		Octobre	10	4.301	3.447	3.470	1.240	3.831		0.906	1.210	1.024	4.199
37		Novembre	11	5.764	3.492	3.527	1.634	3.856		0.915	1.748	0.935	6.164
38		Décembre	12	7.132	3.562	3.570	1.998	3.882		0.920	2.148	0.930	7.666
39	1963	Janvier	13	2.541	3.578	3.609	0.704	3.907		0.924	0.733	0.960	2.647
40		Février	14	2.475	3.640	3.621	0.684	3.932		0.921	0.680	1.006	2.461
41		Mars	15	3.031	3.602	3.630	0.835	3.957		0.917	0.816	1.024	2.961
42		Avril	16	3.266	3.658	3.665	0.891	3.982		0.920	0.840	1.060	3.080

Main | DATA | Chap5 Forcst | SEAS | SADJ | Chap6 Forcst | SESOptim | EWSOptim

FIGURE 18 –Exercice CH05EX09.xls : détermination des différentes composantes, parmi lesquels les valeurs de tendance et les coefficients saisonniers sur les données de ventes de champagne en France

[2] Deux autres exercices montrent respectivement, l'un le choix d'un modèle additif, qui convient moins bien, et l'autre le choix d'un modèle additif sur les logarithmes des données qui fournit des résultats proches mais pas identiques.

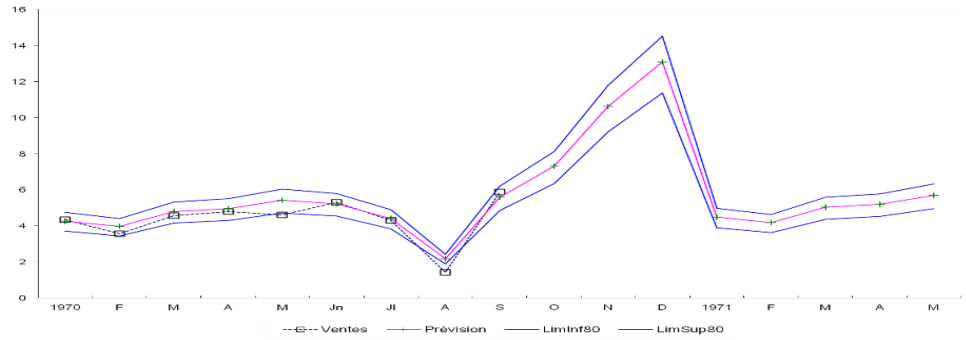


FIGURE 19 –Exercice CH05EX09.xls : prévisions et intervalles de prévision à 80% sur les données de ventes de champagne en France

CH06EX05.xls — A la suite de l'exercice précédent, trois autres méthodes sont ajoutées : la méthode de lissage exponentiel simple, la méthode de lissage exponentiel simple avec correction saisonnière et la méthode de lissage exponentiel de Winters. Pour le lissage exponentiel simple, les prévisions en *t* et d'horizon *h* sont données par la formule

$$\hat{y}_t(h) = S_t,$$

où S_t représente un niveau qui est calculé par la récurrence

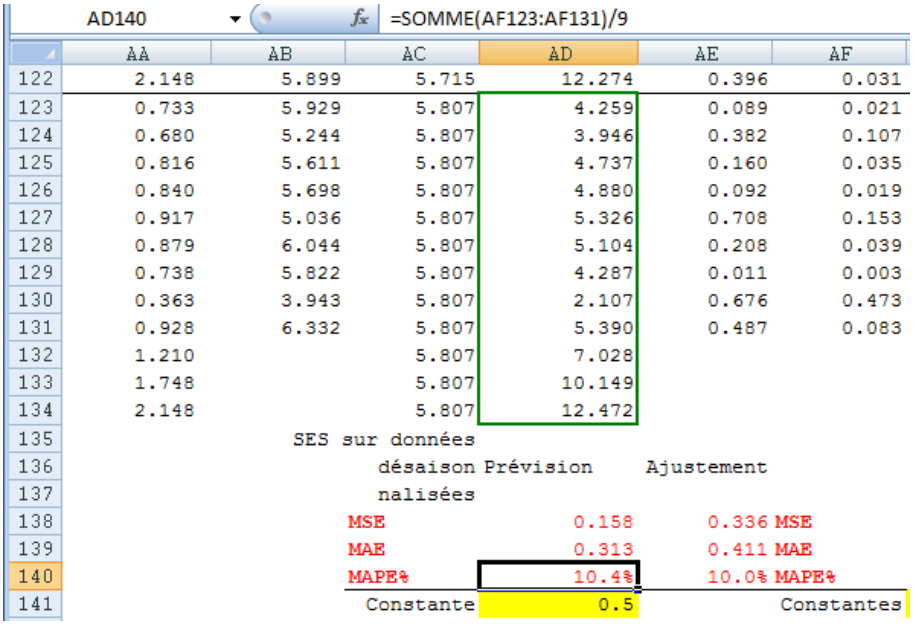
$$S_t = \alpha y_t + (1-\alpha)S_{t-1},$$

avec une valeur initiale S_0. Il consiste donc, voir figure 20, à calculer la valeur lissée en *t* comme moyenne pondérée entre la donnée en *t* et la valeur lissée en $t-1$, avec un poids 0,89 dans la cellule de nom ALPHA appelée "constante de lissage" pour la première et 1 − 0,89 = 0,11 pour l'autre. La constante de lissage a été estimée de manière à minimiser la moyenne des carrés des résidus sur l'historique, donc jusqu'en décembre 1969 atteignant 7,024. On remarque que toutes les prévisions sont basées sur la donnée de décembre et sur la valeur lissée de décembre 1969. Ceci peut-être effectué au moyen du Solver d'Excel mais un scénario d'Excel préparé à l'avance a permis de récupérer les paramètres. On aurait pu aussi déterminer cette constante de lissage optimale par balayage (réalisable avec le ruban Données > Analyse de scénarios > Tables de données) comme dans l'exercice CH06EX01.xls non montré. Les erreurs de prévision relatives en valeur absolue sont élevées, comme le montre le critère MAPE ("Mean absolute percentage error"), moyenne de colonne Z pour les 9 données de 1970, égal à 234,7 %.

ARRONDI.AU.MULTIPLE | =ALPHA*W122+(1-ALPHA)*X122

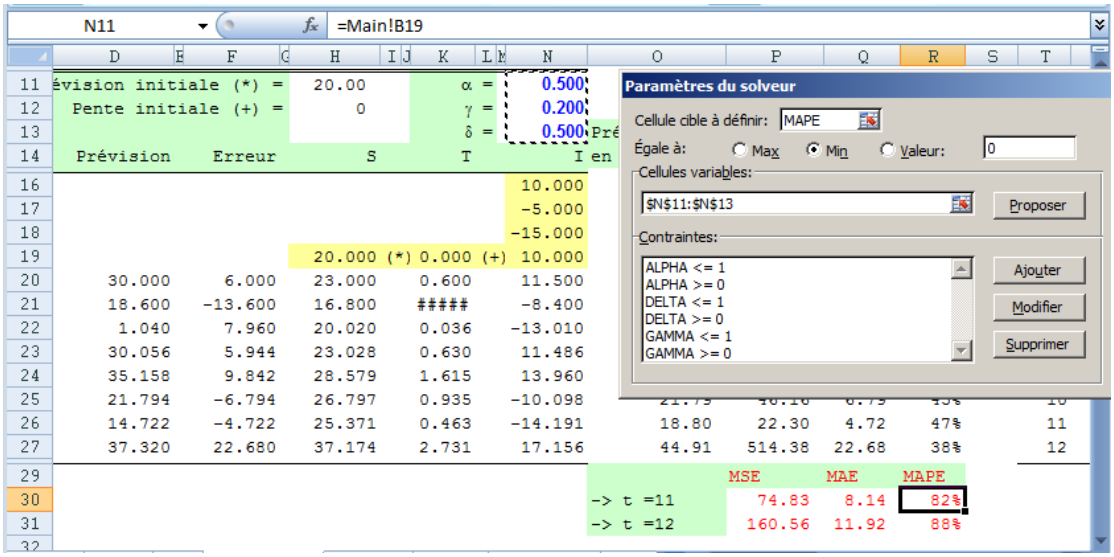
	T	U	V	W	X	Y	Z	AA
121		Novembre	95	9.851	6.820	3.031	0.308	1.748
122		Décembre	96	12.67	9.518	3.152	0.249	2.148
123	1970	Janvier	97	4.348	=ALPHA*W122+(1-ALPHA)*X122			0.733
124		Février	98	3.564	12.324	8.760	2.458	0.680
125		Mars	99	4.577	12.324	7.747	1.693	0.816
126		Avril	100	4.788	12.324	7.536	1.574	0.840
127		Mai	101	4.618	12.324	7.706	1.669	0.917
128		Juin	102	5.312	12.324	7.012	1.320	0.879
129		Juillet	103	4.298	12.324	8.026	1.867	0.738
130		Août	104	1.431	12.324	10.893	7.612	0.363
131		Septembre	105	5.877	12.324	6.447	1.097	0.928
132		Octobre	106		12.324			1.210
133		Novembre	107		12.324			1.748
134		Décembre	108		12.324			2.148
135				Lissage				
136	Tableau 5.36			exponentiel	Prévision	Ajustement		
137	----------			simple				
138	Prévision			MSE	65.593	7.024		
139	----------			MAE	8.011	1.818		
140				MAPE%	234.7%	45.0%		
141	TAB5.36			Constante	0.89019745			

FIGURE 20 –Exercice CH06EX05.xls : utilisation du lissage exponentiel simple pour prévoir les ventes de champagne en France



	AA	AB	AC	AD	AE	AF
122	2.148	5.899	5.715	12.274	0.396	0.031
123	0.733	5.929	5.807	4.259	0.089	0.021
124	0.680	5.244	5.807	3.946	0.382	0.107
125	0.816	5.611	5.807	4.737	0.160	0.035
126	0.840	5.698	5.807	4.880	0.092	0.019
127	0.917	5.036	5.807	5.326	0.708	0.153
128	0.879	6.044	5.807	5.104	0.208	0.039
129	0.738	5.822	5.807	4.287	0.011	0.003
130	0.363	3.943	5.807	2.107	0.676	0.473
131	0.928	6.332	5.807	5.390	0.487	0.083
132	1.210		5.807	7.028		
133	1.748		5.807	10.149		
134	2.148		5.807	12.472		
135		SES sur données				
136		désaison	Prévision	Ajustement		
137		nalisées				
138		MSE	0.158	0.336	MSE	
139		MAE	0.313	0.411	MAE	
140		MAPE%	10.4%	10.0%	MAPE%	
141		Constante	0.5		Constantes	

FIGURE 21 –Exercice CH06EX05.xls : utilisation du lissage exponentiel simple avec correction saisonnière pour prévoir les ventes de champagne en France, avant optimisation



	D	F	H	K	N	P	Q	R	T
11	évision initiale (*) =		20.00	α =	0.500				
12	Pente initiale (+) =		0	γ =	0.200				
13				δ =	0.500				
14	Prévision	Erreur	S	T	I en				
16					10.000				
17					-5.000				
18					-15.000				
19			20.000 (*)	0.000 (+)	10.000				
20	30.000	6.000	23.000	0.600	11.500				
21	18.600	-13.600	16.800	#####	-8.400				
22	1.040	7.960	20.020	0.036	-13.010				
23	30.056	5.944	23.028	0.630	11.486				
24	35.158	9.842	28.579	1.615	13.960				
25	21.794	-6.794	26.797	0.935	-10.098				
26	14.722	-4.722	25.371	0.463	-14.191	18.80	22.30	4.72	47%
27	37.320	22.680	37.174	2.731	17.156	44.91	514.38	22.68	38%
29						MSE	MAE	MAPE	
30						-> t =11	74.83	8.14	82%
31						-> t =12	160.56	11.92	88%

Paramètres du solveur

Cellule cible à définir: MAPE

Égale à: Max Min Valeur: 0

Cellules variables: $N$11:$N$13 — Proposer

Contraintes:
ALPHA <= 1
ALPHA >= 0
DELTA <= 1
DELTA >= 0
GAMMA <= 1
GAMMA >= 0

Ajouter / Modifier / Supprimer

FIGURE 22 –Exercice CH06EX05.xls : utilisation du lissage exponentiel de Winters multiplicatif pour prévoir les ventes de champagne en France, au moement de l'optimisation

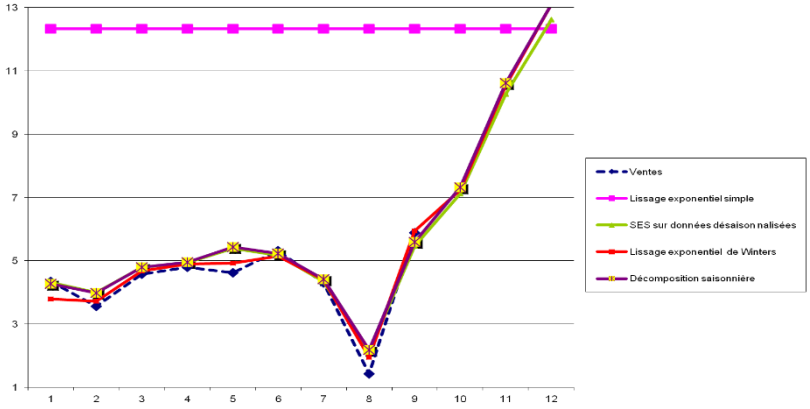


FIGURE 23 –Exercice CH06EX05.xls : comparaison des prévisions fournies par plusieurs méthodes pour prévoir les ventes de champagne en France

On peut obtenir de meilleurs résultats de lissage exponentiel simple avec correction saisonnière. Pour cela on emploie les facteurs saisonniers de la décomposition saisonnière, par exemple 0,733 pour janvier, voir la figure 21, et les données corrigées des variations saisonnières, par exemple 5,899 = 12,67/2,148 pour décembre 1969. Cette fois on n'a pas encore optimisé la constante de lissage choisie égale à 0,5. Quand les prévisions désaisonnalisées sont obtenues, 5,879 pour tous les mois, elles sont multipliées par les facteurs saisonniers, ce qui donne par exemple 4,312 pour janvier 1970. On note que le critère MSE sur l'historique vaut 0,336, donc beaucoup moins que pour le lissage exponentiel simple sans correction saisonnière. Le critère MAPE de prévision de 10,8 % est aussi meilleur.

Dans la figure 22, on montre l'application du lissage exponentiel de Winters en version multiplicative, une méthode plus complexe dont nous verrons les détails lors de l'examen du classeur suivant mais en version additive toutefois. On y voit l'emploi du Solver pour déterminer les trois constantes de lissage qui minimisent le critère MSE calculé jusqu'en décembre 1969. Disons simplement que les prévisions du lissage exponentiel simple avec correction saisonnière et du lissage exponentiel de Winters, montrées dans la figure 23, sont très bonnes.

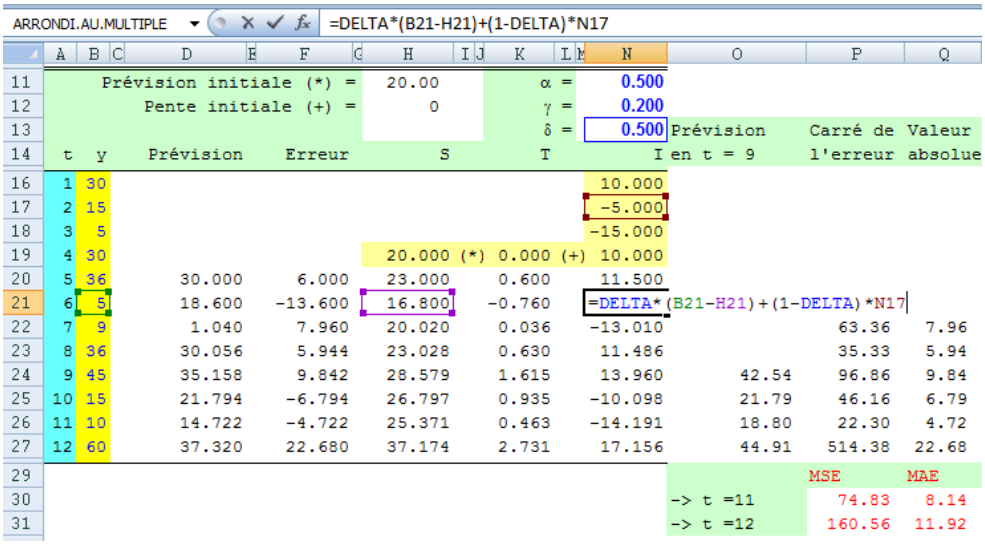


FIGURE 24 –Exercice CH06EX08.xls : calcul des valeurs lissées et des prévisions pour une méthode de Winters additive sur une série trimestrielle

CH06EX08.xls — Dans cet exercice, on étudie la méthode de Winters en version additive et en version multiplicative sur une série trimestrielle. Les prévisions en *t* et d'horizon *h* sont obtenues à partir d'une tendance (localement) linéaire avec une saisonnalité, par exemple dans le cas additif

$$\hat{y}_t(h) = S_t + hT_t + I_{t+h}, \text{ (pour } h < s)$$

où, comme pour le lissage exponentiel simple, S_t représente un niveau ou une valeur lissée, T_t est une pente de tendance, et I_t est un coefficient saisonnier utilisant la périodicité. Les trois composantes sont mises à jour chacune par une formule de lissage exponentiel spécifique, c'est-à-dire des moyennes pondérées entre deux informations. Par exemple, le niveau est évalué à partir de la donnée sous forme corrigée de la variation saisonnière, $y_t - I_{t-s}$, et du niveau précédent augmenté de la pente de la tendance précédente, $S_{t-1} + T_{t-1}$. Le poids du premier constituant de chaque équation est une constante de lissage et il y en a donc trois, α pour la valeur lissée, γ pour la pente de tendance et δ pour les coefficients saisonniers. Ces équations sont les suivantes :

$$S_t = \alpha[y_t - I_{t-s}] + (1-\alpha)[S_{t-1} + T_{t-1}]$$
$$T_t = \gamma[S_t - S_{t-1}] + (1-\gamma)T_{t-1}$$
$$I_t = \delta[y_t - S_t] + (1-\delta)I_{t-s}.$$

On montre un choix simple de valeurs initiales pour les trois composantes. Les valeurs optimales des trois constantes de lissage sont obtenues en minimisant un critère sur l'historique des données. On inspecte aussi les formules de la forme de correction d'erreurs de la méthode, un ensemble alternatif d'équations où les composantes sont mises à jour en employant la dernière erreur de prévision, ainsi que la forme ARIMA, une équation reliant les données et les erreurs de prévision. Enfin, on présente sur la même série trimestrielle la méthode de Winters multiplicative, où $y_t - I_{t-s}$ et $y_t - S_t$ sont remplacés dans les équations précédentes par y_t/I_{t-s}, et y_t/S_t. On expérimente alors les effets d'un changement des valeurs des constantes de lissage.

	C	HAB	SIT	Y	PRIX
X	1	15	0		33.0
	1	24	0		30.0
	1	19	0		37.5
	1	29	0		42.0
	1	22	1		44.5
	1	21	0		34.0
	1	20	1		40.0
	1	18	0		24.5
	1	27	1		48.0
	1	25	0		36.5

X'										X'X			X'Y
1	1	1	1	1	1	1	1	1	1	10	220	3	370
15	24	19	29	22	21	20	18	27	25	220	5006	69	8288
0	0	0	0	1	0	1	0	1	0	3	69	3	132.5

INV(X'X)			INV(X'X).X'Y
3.02031802	−0.1333922	0.04770318	18.2835689
−0.1333922	0.00618375	−0.0088339	0.725265018
0.04770318	−0.0088339	0.48881037	9.202002356

FIGURE 25 –Exercice CH07EX01.xls : calcul de la régression linéaire multiple par la méthode de la matrice inverse (voir texte)

CH07EX02.xls — Signalons auparavant que dans un exercice non traité ici, voir figure 25, on montre une manière dynamique d'effectuer dans Excel les calculs de la régression multiple[3] qui permet de visionner instantanément des changements de données, dont l'introduction de quasi-colinéarité, et les problèmes numériques. Le but de l'exercice est d'introduire — ou de rappeler pour les apprenants qui ont déjà bénéficié d'un cours de statistique de base — les tests d'hypothèse sur les coefficients du modèle de régression linéaire simple (ou multiple) ainsi que la détermination d'intervalles de confiance. Les données sont relatives à 20 échantillons de taille 10 prélevés dans une population d'habitations pour lesquelles on étudie le prix en fonction de la surface habitable. Le but est d'examiner la sensibilité du coefficient de régression estimés à l'échantillon considéré et d'avoir la possibilité d'évaluer cette sensibilité à partir d'un échantillon unique. Plus concrètement, dans la figure 26, on calcule les rapports de Student *t* employant la valeur population du coefficient de régression (en l'occurrence 0,658) donc *t* = 0,892 pour le premier échantillon. On montre empiriquement dans la figure 27 que la distribution des 20 rapports est approximativement normale de moyenne 0 et de variance 1 et on effectue même la comparaison avec la distribution de Student à 8 degrés de liberté. On utilise ici les fonctions statistiques d'Excel pour calculer une probabilité cumulée. Notons aussi un exercice qui permet d'étendre ceci au moyen de données artificielles, de manière à juger l'importance relative des différentes conditions d'application, voir figure 28.

[3] Ceci requiert l'emploi de fonctions matricielles d'Excel, assez mal documentées car il faut savoir que pour les entrer on doit entrer la combinaison de touches Maj-Ctrl-Entrée.

D43 =(B43-C$38)/C43

	A	B	C	D
38	Coefficient(s) X		0.657961935	
39	Erreur-type de coef.		0.129849259	
40				
41	BRUT:	b1	erreur-type	statistique t
43	1	0.8916	0.4832	0.4835
44	2	0.7490	0.2844	0.3201
45	3	0.4808	0.4308	-0.4111
46	4	1.4502	0.4171	1.8993
47	5	0.5385	0.4801	-0.2488
48	6	0.1890	0.4333	-1.0823
49	7	1.1321	0.4266	1.1115
50	8	1.0308	0.3409	1.0939
51	9	1.5905	0.4114	2.2669
52	10	0.4184	0.2810	-0.8525
53	11	0.4336	0.4627	-0.4850
54	12	0.5478	0.2257	-0.4880
55	13	1.3978	0.6559	1.1279
56	14	0.6667	0.3983	0.0219
57	15	0.6991	0.4138	0.0995
58	16	0.7480	0.4104	0.2194
59	17	0.6460	0.3960	-0.0301
60	18	0.0776	0.3968	-1.4627
61	19	0.3968	0.1900	-1.3747
62	20	1.4128	0.2963	2.5476

FIGURE 26 –Exercice CH07EX02.xls : calcul des statistiques de Student (par rapport au vrai coefficient de régression) pour 20 échantillons

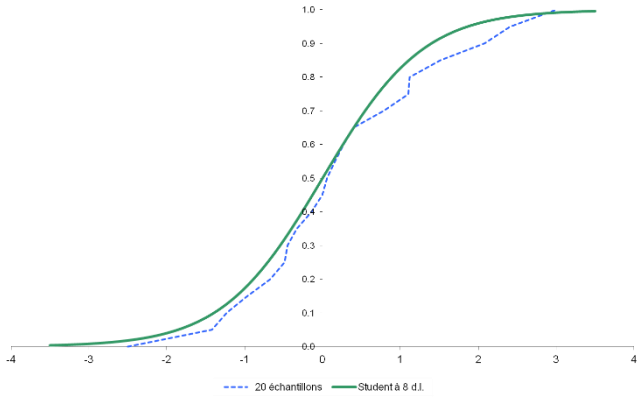


FIGURE 27 –Exercice CH07EX02.xls : fonction de distribution empirique des statistiques de Student pour les 20 échantillons, comparée à la fonction de distribution de Student à 8 degrés de liberté

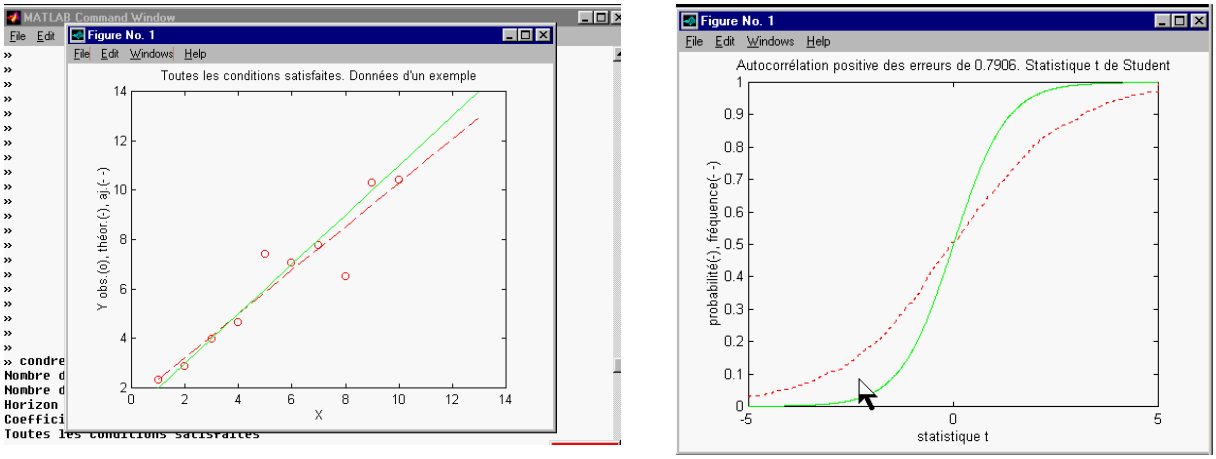


FIGURE 28 –Exercice CH07EX04 : extraits d'une vidéo montrant la distribution des statistiques de Student dans le cas où toutes les conditions d'application sont satisfaites sauf que les erreurs sont autocorrélées

CH07EX08.xls — Le jeu de données est tiré de Rao Kadiyala (1970) mais décalé de 33 années et est destiné à illustrer la régression linéaire multiple sur des données temporelles. Il s'agit d'élaborer un modèle explicatif de ventes de crème aux Etats-Unis sur le marché nord-américain. On se doute qu'elles dépendent du revenu des consommateurs et du prix de la crème glacée, de la température extérieure moyenne (en degrés Fahrenheit) de la période et de la période précédente, voir figure 29. Les données sont relatives aux 30 périodes de 4 semaines allant du 18 mars 1984 au 12 juillet 1986, et on dispose aussi des jours fériés tombant dans cet intervalle[4]. On emploie l'outil de régression d'Excel pour illustrer quelques régressions sur ces variables, d'abord en employant toutes les variables explicatives disponibles, celles citées plus le temps, ensuite en employant toutes les combinaisons parmi 4 variables explicatives. On procède ensuite en employant une méthode de sélection pas à pas, en commençant par la variable la plus corrélée aux ventes, la température de la période précédente, voir la matrice de corrélation dans la figure 30, puis en cherchant parmi les variables restantes. Les variables introduites sont successivement le temps et la température de la période. L'inspection des résidus montre que les résidus les plus élevés correspondent à la présence de jours fériés. En ajoutant des variables binaires pour chacun des quatre types de jours fériés, et la variable correspondant à la fête de l'indépendance (4 juillet) a un effet statistiquement significatif. L'analyse est détaillée dans Mélard (2006, $3.4).

IC	DATE	PRICE	INCOME	TEMP	LAGTEMP	HOLIDAY				holiday		
0.386	1	0.27	78	41		0	0	18/03/1984	14/04/1984	0		
0.374	2	0.282	79	56	41	0	0	15/04/1984	12/05/1984	0		
0.393	3	0.277	81	63	56	1	0	13/05/1984	9/06/1984	1	28/05/1984	Memorial
0.425	4	0.28	80	68	63	1	0	10/06/1984	7/07/1984	1	4/07/1984	Independance
0.406	5	0.272	76	69	68	0	0	8/07/1984	4/08/1984	0		
0.344	6	0.262	78	65	69	0	0	5/08/1984	1/09/1984	0		
0.327	7	0.275	82	61	65	1	0	2/09/1984	29/09/1984	1	3/09/1984	Labour
0.288	8	0.267	79	47	61	0	0	30/09/1984	27/10/1984	0		
0.269	9	0.265	76	32	47	0	0	28/10/1984	24/11/1984	0	22/11/1984	Thanksgiving
0.256	10	0.277	79	24	32	1	0	25/11/1984	22/12/1984	1		
0.286	11	0.282	82	28	24	0	1	23/12/1984	26/01/1985	0		
0.298	12	0.27	85	26	28	0	1	27/01/1985	23/02/1985	0		
0.329	13	0.272	86	32	26	0	1	24/02/1985	23/03/1985	0		
0.318	14	0.287	83	40	32	0	1	24/03/1985	20/04/1985	0		
0.381	15	0.277	84	55	40	0	1	21/04/1985	18/05/1985	0		
0.381	16	0.287	82	63	55	1	1	19/05/1985	15/06/1985	1	27/05/1985	Memorial
0.47	17	0.28	80	72	63	1	1	16/06/1985	13/07/1985	1	4/07/1985	Independance
0.443	18	0.277	78	72	72	0	1	14/07/1985	10/08/1985	0		
0.386	19	0.277	84	67	72	0	1	11/08/1985	7/09/1985	0	2/09/1985	Labour
0.342	20	0.277	86	60	67	1	1	8/09/1985	5/10/1985	1		
0.319	21	0.292	85	44	60	0	1	6/10/1985	2/11/1985	0		
0.307	22	0.287	87	40	44	1	1	3/11/1985	30/11/1985	1	28/11/1985	Thanksgiving
0.284	23	0.277	94	32	40	0	1	1/12/1985	28/12/1985	0		
0.326	24	0.285	92	27	32	0	2	29/12/1985	25/01/1986	0		
0.309	25	0.282	95	28	27	0	2	26/01/1986	22/02/1986	0		
0.359	26	0.265	96	33	28	0	2	23/02/1986	22/03/1986	0		
0.376	27	0.265	94	41	33	0	2	23/03/1986	19/04/1986	0		
0.416	28	0.265	96	52	41	0	2	20/04/1986	17/05/1986	0		
0.437	29	0.268	91	64	52	1	2	18/05/1986	14/06/1986	1	27/05/1985	Memorial
0.548	30	0.26	90	71	64	1	2	15/06/1986	12/07/1986	1	4/07/1986	Independance

FIGURE 29 –Exercice CH07EX08.xls : les données de vente de crème glacée, avec indication des jours fériés durant chaque intervalle de temps

	IC	*DATE*	*PRICE*	*INCOME*	*LAGTEMP*	*HOLIDAY*
IC	1					
DATE	0.259105	1				
PRICE	-0.253	-0.11197	1			
INCOME	0.064643	0.840474	-0.1351	1		
LAGTEMP	0.493423	-0.19787	-0.0927	-0.46916	1	
HOLIDAY	0.332694	-0.01734	0.115384	-0.12181	0.354016	1
Note: first row (DATE = 1) was omitted because of missing LAGTEMP						

FIGURE 30 –Exercice CH07EX08.xls : matrice de corrélation entre quelques variables pour les données de ventes de crème glacée

CH07EX09.xls — Le jeu de données est tiré de Vatter *et al.* (1978). Il s'agit d'élaborer un modèle explicatif des ventes mensuelles de céréales pour le petit déjeuner d'une certaine marque pendant quatre années en employant les dépenses de promotion qui s'adressent aux consommateurs (CP), d'une part, et aux distributeurs (DA), d'autre part. Comme ces dépenses de promotion n'ont pas

[4] Cet intervalle n'est pas explicitement spécifié dans Rao Kadiyala (1970) mais nous avons pu le retrouver à partir des jours fériés tombant dans l'intervalle en employant les fonctions de date d'Excel, notamment JOURSEM.

nécessairement d'effet le même mois, on envisage aussi les variables décalées d'un mois et de deux mois, par exemple CP_1 et CP_2, pour les dépenses de promotion consommateurs et DA_1 et DA_2 pour les autres. On considère aussi douze variables binaires pour chaque mois (JAN à DEC) et la variable de temps (TREND = 1, 2, ...) pour représenter respectivement la saisonnalité et la tendance. Il y a donc en tout 19 variables explicatives potentielles plus la constante, trop pour l'outil de régression d'Excel limité à 16 paramètres (Mélard, 2014). Il s'avère intéressant de traiter séparément les trois groupes de variables explicatives : (a) les variables binaires (sauf DEC pour éviter la colinéarité) et TREND pour représenter la saisonnalité et la tendance, (b) les dépenses de promotion consommateurs en t, $t-1$ et $t-2$, (c) les dépenses de promotion distributeurs en t, $t-1$ et $t-2$. La figure[5] 31 montre les données et les valeurs prédites obtenues pour les jeux de variables et pour une sélection combinant la plupart des variables (toutes sauf DEC, CP_2, DA_1, DA_2). Comme les dépenses de promotion peuvent être quantifiées, il est possible de déduire des recommandations de gestion à partir des coefficients de régression.

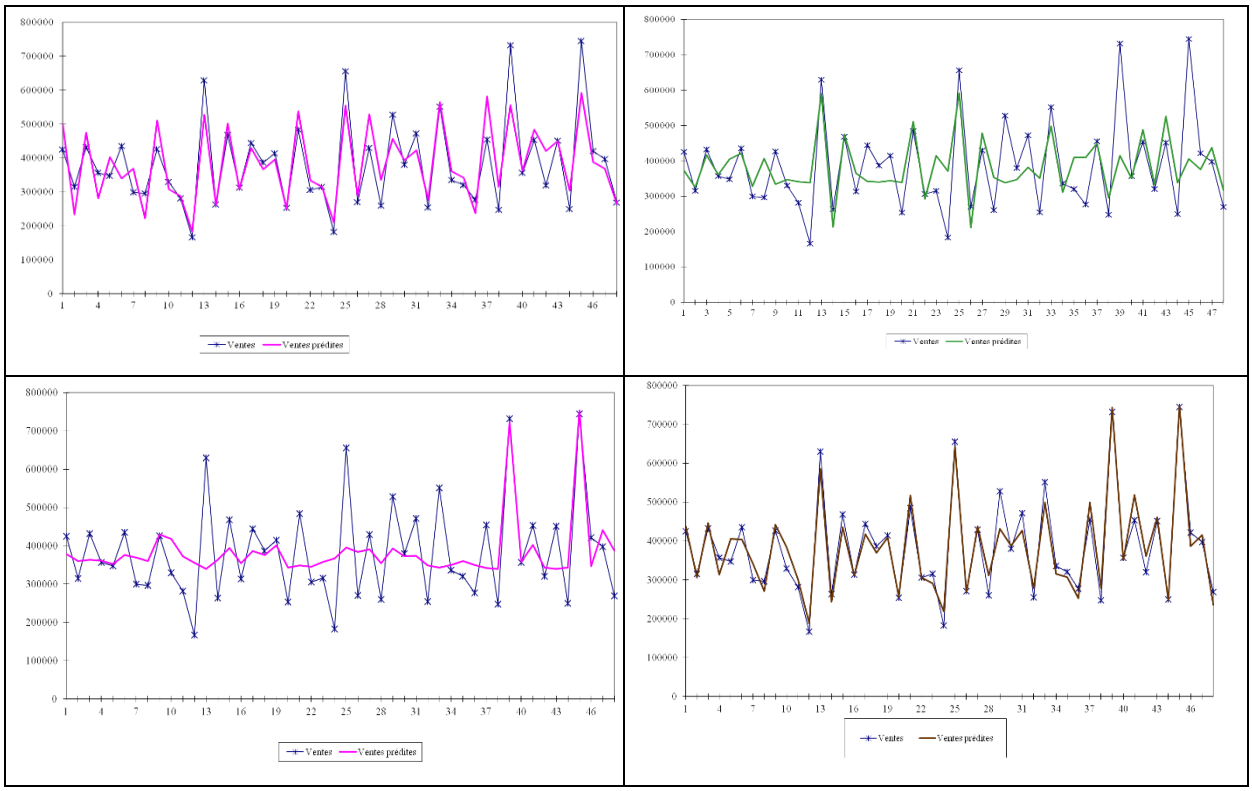


FIGURE 31 –Exercice CH07EX09.xls : graphiques montrant les données de ventes de céréales pour le petit déjeuner (en rouge) et les valeurs ajustées obtenues en employant quatre sous-ensembles de variables explicatives

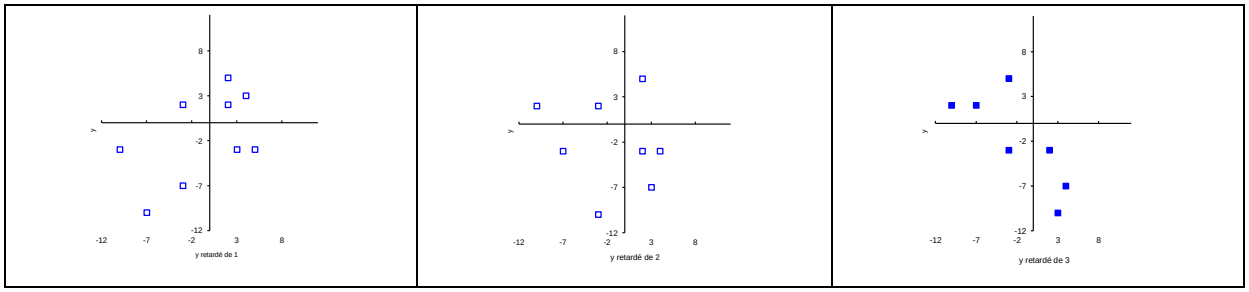


FIGURE 32 –Exercice CH08EX02.xls : diagrammes de corrélation avec retard pour une série artificielle

CH08EX02.xls — L'exercice a pour but d'introduire en termes simples le concept d'autocorrélation. Il concerne une série chronologique formée de $T = 10$ données artificielles notées $\{y_t\}$. Nous

[5] Figures proposées par un groupe d'étudiants lors d'une discussion de cas.

commençons par la visualisation de la corrélation avec retard, en l'occurrence pour les retards 1, 2 et 3, c'est-à-dire de la corrélation de la série $\{y_t\}$ avec la série $\{y_{t-1}\}$, ou entre la série $\{y_t\}$ et la série $\{y_{t-2}\}$, ou encore entre la série $\{y_t\}$ et la série $\{y_{t-3}\}$. Ces séries sont représentées dans le tableau à côté de la série des données. La visualisation s'effectue dans des diagrammes de dispersion avec la donnée du temps *t*, y_t, en ordonnées et les valeurs retardées, respectivement y_{t-1}, y_{t-2}, y_{t-3}, en abscisses, voir figure 32. Ensuite, on calcule les corrélations avec retard en discutant les calculs des moyennes et variances des 4 séries (de 10 à 7 données), puis des covariances et corrélations. Ceci justifie l'introduction des autocorrélations[6] qui se basent sur la seule moyenne $\bar{y}$ et la seule variance de la série complète et sur les sommes de produits des écarts à la moyenne $(y_t - \bar{y})(y_{t-k} - \bar{y})$, $k = 1, 2, 3$.

$$r_k = \frac{\frac{1}{T}\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2},$$

Voir figure 33 où on ne peut pas voir que les autocorrélations ne diffèrent pas énormément des corrélations avec retard. Dans la suite du cours on emploiera les autocorrélations d'une série de résidus $\{e_t\}$. Notons enfin que la quantification de la corrélation avec retard à l'aide des coefficients de corrélation ou d'autocorrélation n'enlève rien à l'intérêt des graphiques, notamment compte tenu du manque de résistance des corrélations vis à vis de données aberrantes.

F58 : =+F56/$D56

	A	B	C	D	E	F	G	H	I	J
41	AUTOCORRELATION ENTRE			y_t	ET	y_{t-k}				
42			$y^* = \_$							
43	t	y_t	$y_t - \bar{y}$	y^{*2}	y^*_{t-1}	$y^*_t y^*_{t-1}$	y^*_{t-2}	$y^*_t y^*_{t-2}$	y^*_{t-3}	$y^*_t y^*_{t-3}$
44										
45	1	4	5	25						
46	2	3	4	16	5	20				
47	3	-3	-2	4	4	-8	5	-10		
48	4	-7	-6	36	-2	12	4	-24	5	-30
49	5	-10	-9	81	-6	54	-2	18	4	-36
50	6	-3	-2	4	-9	18	-6	12	-2	4
51	7	2	3	9	-2	-6	-9	-27	-6	-18
52	8	2	3	9	3	9	-2	-6	-9	-27
53	9	5	6	36	3	18	3	18	-2	-12
54	10	-3	-2	4	6	-12	3	-6	3	-6
55										
56	Moy.=	-1	0	224		105		-25		-125
57	Var.=	22.4								
58			Autocorrelation =			0.47		-0.11		-0.56

FIGURE 33 –Exercice CH08EX02.xls : calcul des autocorrélations de retards 1 à 3 pour une série artificielle

CH08EX04.xls — Cet exercice est en partie similaire au précédent, sauf qu'ici les données artificielles sont plus nombreuses, 400 au lieu de 100, sont générées dans le classeur au moyen de nombres pseudo-aléatoires (en employant notamment la fonction ALEA) et qu'une interprétation statistique est ajoutée. Chaque pression de la touche de mise à jour (F9) permet de générer de nouvelles données. Dans le texte d'accompagnement, on en profite pour introduire le concept de processus bruit blanc, une suite de variables aléatoires indépendantes. On demande de consulter les formules pour le calcul des corrélations avec retard et des autocorrélations, proches cette fois compte tenu du grand nombre de données, voir figure 34.

Ensuite on peut examiner la distribution des autocorrélations sur 100 réalisations du processus (avec une table de données, comme mentionné dans l'exercice CH06EX05.xls). On peut alors

[6] Cette introduction est en fait plus due aux développements théoriques plus simples et aux propriétés mathématiques qu'à la simplification des calculs.

visionner la distribution en classes des 100 autocorrélations pour le retard 1 sous forme de tableau et aussi de graphique, voir figure 35. De plus la moyenne et la variance de la distribution empirique sont données. Conformément à la théorie asymptotique qui annonce une loi normale, on peut donc visionner une distribution empirique en forme de cloche symétrique, de moyenne autour de 0 (de l'ordre de −1/400) et de variance de l'ordre de 1/400 = 0,0025), donc d'écart-type 1/20 = 0,05. De plus on peut aussi visionner la distribution de la somme des carrés des trois premières autocorrélation fois 400, qui doit suivre une distribution e forme de cloche dissymétrique à droite, en théorie de type χ^2 à 3 degrés de liberté. En demandant la mise à jour, on peut obtenir de nouveaux résultats assez proches. Ces résultats empiriques vont donc justifier les intervalles de $-1{,}96/\sqrt{T}$ à $1{,}96/\sqrt{T}$ est approximativement égale à 0,05. Par conséquent, cela justifie empiriquement que si la série temporelle est produite par des variables aléatoires indépendantes, on s'attend avec une probabilité de 0,95 que l'autocorrélation de retard 1 soit comprise entre $-1{,}96/\sqrt{T}$ et $1{,}96/\sqrt{T}$. Il en est de même pour ce qui concerne les autocorrélations pour les autres retards.

D424 =+B422

	A	B	C	D	E	F	G	H	I	J
422	398	0.820	0.042	-2.686	0.971					
423	399	1.049	0.820	0.042	-2.686					
424	400	-1.752	1.049	0.820	0.042					
425										
426	Correlation =		0.017	-0.006	0.008					
427	Moy.	{y_t}	0.06	0.07	0.06					
428	Moy.	{y_{t-k}}	0.07	0.08	0.08					
429	Var.	{y_t}	0.96	0.95	0.94					
430	Var.	{y_{t-k}}	0.94	0.95	0.95					
835	399	1.049	0.985	0.971	0.756	0.745	-0.022	-0.022	-2.749	-2.709
836	400	-1.752	-1.816	3.297	0.985	-1.789	0.756	-1.373	-0.022	0.040
837										
838	Moy.=	0.06397	0	381.4465		5.038419		-2.289293		14.395429
839	Var.=	0.95362								
840			Autocorrelation =			0.013		-0.006		0.038
841										
842	On a généré cent échantillons et calculé les autocorrélations de retard 1, 2 et 3									
843	SAMPLE		Echantillon			1		2		3
844						0.013		-0.006		0.038
845				1		-0.108		-0.039		-0.027
846				2		-0.025		-0.011		0.016

FIGURE 34 –Exercice CH08EX04.xls : calcul des corrélations avec retards 1, 2 et 3 et des autocorrélations de retards 1, 2 et 3 pour une série artificielle de longueur 400 générée par un processus bruit blanc

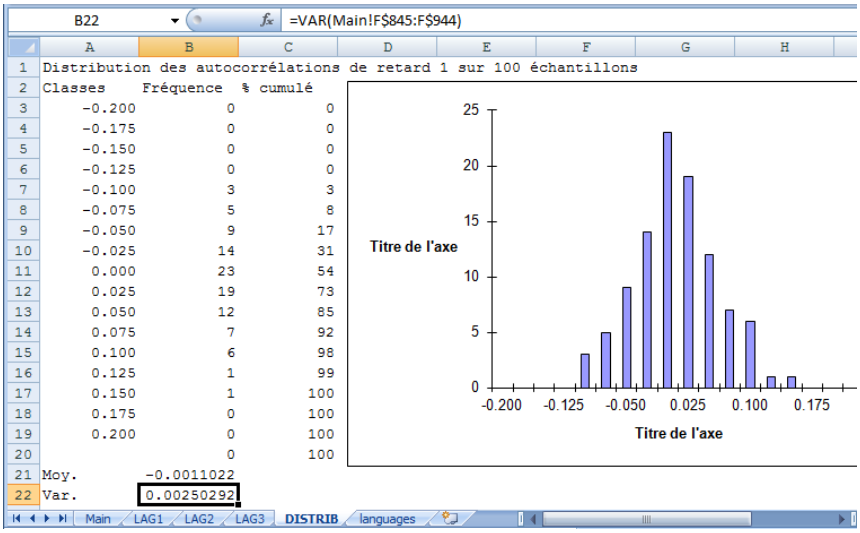


FIGURE 35 –Exercice CH08EX04.xls : distribution des autocorrélations de retard 1 pour 100 séries artificielles de longueur 400 générées par un processus bruit blanc

CH13EX06.xls — Le but de l'exercice est d'étudier la méthode de décomposition saisonnière du logiciel Tramo-Seats sur une série sans saisonnalité, une série notée TICD de taux d'intérêt des certificats de dépôt, aux Etats-Unis, entre décembre 1974 et décembre 1979. Cela permettra de

vérifier la compréhension des principes sur un exemple simple et de vérifier les calculs. L'essentiel des calculs, la modélisation ARIMA avec un modèle très simple d'équation $\nabla TICD_t - m = e_t - \theta\, e_{t-1}$, où les innovations e_t constituent un processus bruit blanc de moyenne 0 et de variance σ^2. et la décomposition proprement dite, sont effectués par les programmes extérieurs Tramo et Seats, respectivement. Il y a évidemment des valeurs spécifiques pour les estimations des paramètres.

L'exemple est exagérément simple par le fait que nous n'avons pas retenu de composante saisonnière, seulement une composante permanente notée P_t et une erreur E_t. Le modèle de décomposition est donc le suivant : $TICD_t = P_t + E_t$. On peut alors montrer que les deux composantes sont décrites comme suit en termes d'innovations indépendantes notées $e^P{}_t$ et $e^E{}_t$:

$$P_t = \frac{1+B}{\nabla} e_t^P, \qquad I_t = e_t^I$$

```
WIENER-KOLMOGOROV FILTERS (ONE SIDE)
TREND COMPONENT
 0.7474  0.1888 -0.0934  0.0462 ...
IRREGULAR COMPONENT
 0.2526 -0.1888  0.0934 -0.0462 ...
```

Calcul des autocovariances

	Tendance	Irrégulier
AR1	-0.4947	-0.4947
MA1	-1	1
$Sigma^2$	0.55856	0.06382

Retard	Tendance	Irrégulier
0	0.7474	0.2526
1	0.1888	-0.1888
2	-0.0934	0.0934

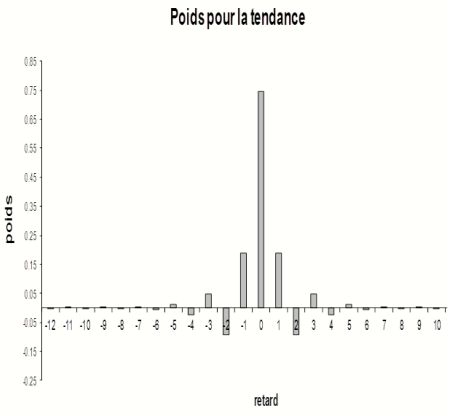


FIGURE 36 –Exercice CH13EX06.xls : comparaison entre les poids de la moyenne mobile pondérée pour la composante permanente (= tendance), ce qu'on appelle filtre de Wiener-Kolmogorov, déduite de Tramo-Seats et ceux obtenus au moyen du calcul d'autocovariances

Connaissant les deux modèles, Seats détermine les filtres appropriés pour extraire la composante de tendance et la composante irrégulière, c'est-à-dire des moyennes mobiles à appliquer aux données. Ces filtres portent le nom de Wiener-Kolmogorov. Excel intervient ici pour évaluer les poids de ces filtres de moyenne mobile pondérée en fonction d'autocovariances d'un processus ARMA(1, 1), voir figure 36. On reproduit même dans Excel la quasi-totalité des résultats intermédiaires de calculs affichés dans la sortie de Tramo-Seats, notamment le calcul des prévisions et des révisions, voir figure 37. On montre aussi empiriquement en employant l'analyse spectrale généralisée[7] que le spectre du processus générateur est égal à la somme des spectres des deux composantes. Enfin, on ajoute encore une approche par simulation, où on simule les deux composantes et on considère aussi leur somme, série chronologique dont on examine les autocorrélations.

Pour plus de détails avec un modèle légèrement différent, voir Mélard (2016).

[7] L'analyse spectrale généralisée permet la présence de racines unités, voir par exemple Gouriéroux et Montfort (1990). Elle est détaillée dans l'exercice suivant CH13EX07.

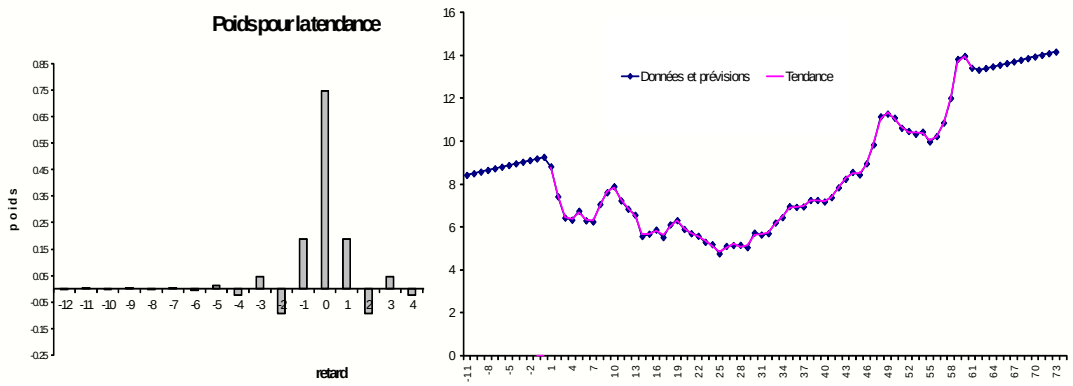


FIGURE 37 –Exercice CH13EX06.xls : poids de la moyenne mobile pondérée pour la composante permanente (= tendance) pour une série de taux d'intérêts et détermination de la série filtrée; on voit aussi les prévisions avant et arrière calculées par Tramo-Seats et vérifiées au moyen d'Excel

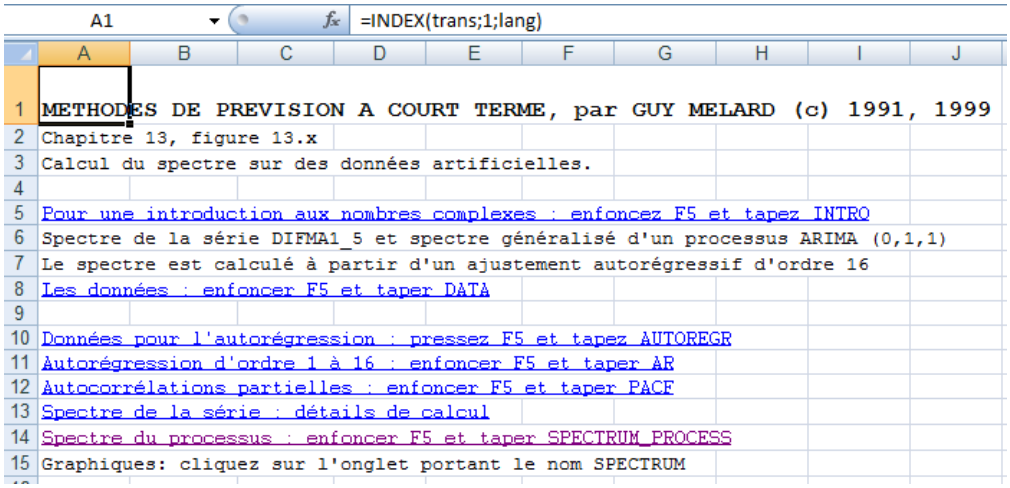


FIGURE 38 –Exercice CH13EX07.xls : les liens vers les parties du classeur

CH13EX07.xls — L'introduction de l'analyse spectrale est effectuée d'abord de manière intuitive, puis de manière plus détaillée. Voir la figure 38 pour les différentes parties de l'exercice.

Pour l'approche intuitive, on introduit la notion de fréquence et on interprète d'abord le spectre de différents processus AR, MA et ARMA que le programme permet de calculer, en entrant les valeurs des coefficients dans trois cellules pour la partie AR et dans trois autres pour la partie MA. Notons qu'on peut même spécifier les degrés associés à ces coefficients, de manière, par exemple, en spécifiant des degrés 1, 4 et 5, et des coefficients −0,6, −0,8 et 0,48 =(−0,6)×(−0,8), à traiter un produit de polynômes de degrés 1 et 4. Immédiatement on peut visualiser le spectre et son logarithme, voir figure 39. Voir Mélard (2006, $2.8) pour une introduction.

Pour une approche plus avancée, on introduit d'abord les fonctions qui permettent de définir et de manipuler des nombres complexes dans Excel, comme on en voit dans la figure 39 pour le calcul d'un produit et d'une puissance. Ceci est fait en relation avec d'autres parties du cours où ils se sont avérés utiles comme les racines de polynômes AR ou MA. On parle de puissances de nombres complexes et on effectue le lien avec les fonctions trigonométriques. Ensuite on montre comment calculer le spectre comme carré du module d'une fonction rationnelle en une exponentielle imaginaire de la fréquence et ceci pour 145 fréquences entre 0 et 1/2. On peut aussi considérer l'analyse spectrale généralisée, simplement en prenant des racines unité pour la partie AR. Après le spectre d'un processus, on explique ensuite comment on peut approcher en employant Excel le spectre d'une série, en l'occurrence une série artificielle générée par un processus ARIMA(0, 1, 1). On procède en considérant un modèle AR(16) de la série et on calcule le carré du module d'une fonction rationnelle en une exponentielle imaginaire de la fréquence, comme ci-dessus.

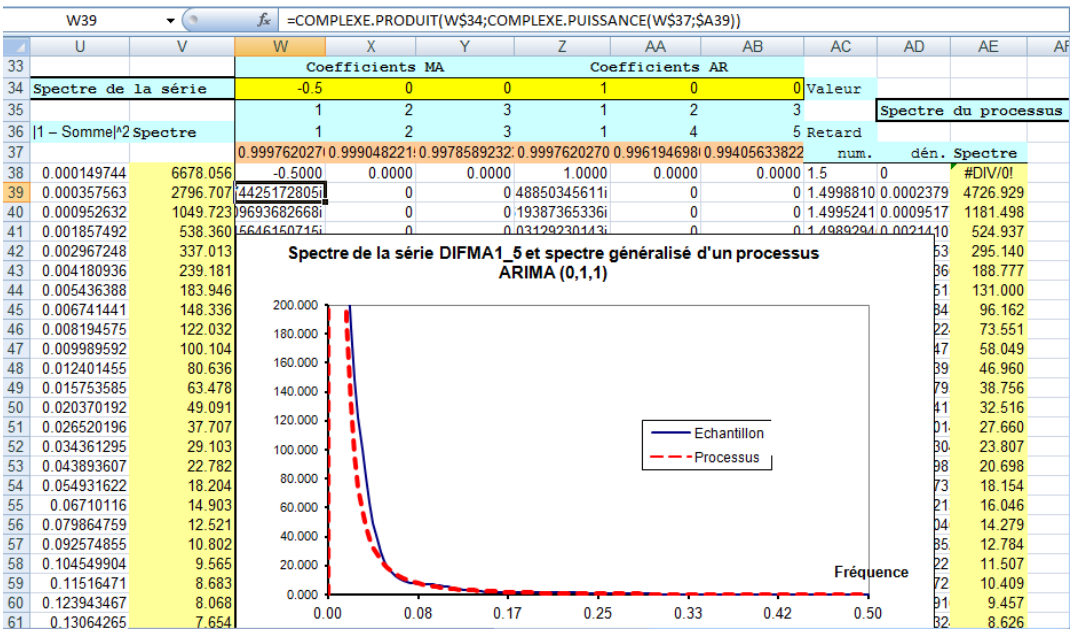


FIGURE 39 –Exercice CH13EX07.xls : analyse spectrale d'une série artificielle et de son processus générateur

Annexe D. Mise à jour des classeurs pour les tableurs récents

Précisons d'abord les parties du cours qui sont réalisées avec Excel et pour quelle activité. Pour des illustrations, voir l'annexe C. Nous indiquons chaque fois les problèmes éventuels posés par Excel 2016 et OpenOffice/LibreOffice Calc. Accessoirement, nous mentionnons aussi Gnumeric.

Tout d'abord, disons que les classeurs du cours, créés dans une version antérieure à savoir Excel 97 et Excel 2003, sont ouverts par Excel 2010 et suivants en mode compatibilité. A cause des macros, cela produit souvent (mais pas toujours) des messages « Avis de sécurité » qui peuvent être dissuasifs.

Dans les exercices du cours multimédia, il s'agit d'abord d'un examen des formules par rapport à la théorie. On discute aussi de l'effet immédiat de changement d'une cellule, par exemple l'introduction d'une donnée extrême. On indique aussi que beaucoup de formules utilisant des récurrences peuvent être obtenues par copier et coller, par exemple pour les moyennes mobiles, le lissage exponentiel et la décomposition saisonnière. On illustre aussi des ajustements statistiques (linéaire, exponentiel, …) et des graphiques standards. Il n'y a pas de problème pour tout ceci avec Excel 2010 et suivants et OpenOffice Calc à partir de la version 3.2, ni avec Gnumeric.

Pratiquement aucune des opérations mentionnées dans les instructions des exercices ne doivent être adaptées pour tenir compte de l'absence des menus d'Excel 2007 à 2016. Les exceptions principales sont l'accès au ruban Révision pour ôter la protection d'une feuille ou au ruban Données pour l'utilisation des outils et cela ne concerne que le document principal appelé CH00_DOC.pdf.

Plusieurs fonctions statistiques et autres sont employées : moyenne, variance, médiane, probabilités et quantiles de quelques lois classiques, estimation par moindres carrés, génération de nombres pseudo-aléatoires, nombres complexes pour l'analyse spectrale (exercices 6 et 7 du chapitre 13). Notons que ces derniers n'étaient pas disponibles dans les anciennes versions françaises d'Excel. Par exemple, dans la cellule M17 de la feuille Second du classeur CH13EX06, la version française d'Excel 2000 trouvait une erreur, là où la version française d'Excel 2003 a reconnu la fonction IMSUM (pour calculer la somme de nombres complexes) que la version française d'Excel 2007 appelle COMPLEXE.SOMME. Il n'y a pas de problème ici avec Excel 2010 ou suivants ou avec Gnumeric. En revanche, Calc 3.2 n'acceptait pas la syntaxe COMPLEXE.SOMME(G17:I17) à corriger en COMPLEXE.SOMME(G17;H17;I17). La correction a été plus longue dans le classeur CH13EX07 où des sommes de 16 termes apparaissent. Ces problèmes sont résolus avec les versions courantes d'OpenOffice et de LibreOffice.

Nous n'avons pas remarqué de différence dans les résultats dus à l'amélioration des fonctions statistiques d'Excel 2010, seulement gênés par le manque de maîtrise du ruban. Evidemment, la feuille Second du classeur CH02EX01, voir tableau 1, qui montrait les problèmes dans le calcul de la variance d'Excel 97 (ainsi que dans les versions 2000 et 2002) n'a plus beaucoup d'objet, sinon historique.

Quelques fonctions de type tableau (« array ») sont utilisées, comme du calcul matriciel au chapitre 7 (à entrer par la combinaison de touches Maj-Ctrl-Entrée !) pour montrer l'effet de la colinéarité en régression (classeur CH07EX01).

Parmi les outils complémentaires d'analyse des données, on insiste sur le besoin de leur installation (rendu plus difficile dans Office 2010 et suivants) mais on n'emploie que la régression linéaire aux chapitres 2 (exercice CH02EX01) et 7 (CH07EX01) et les corrélations (CH07EX01). A cause de l'absence des outils complémentaires, plusieurs parties d'exercices du chapitre 7 ne sont pas disponibles dans OpenOffice. Bien que LibreOffice 5.2.7 possède des procédures statistiques, la régression linéaire multiple n'en fait pas partie, donc cela ne permet pas de résoudre les exercices du chapitre 7. Gnumeric propose des outils équivalents mais d'emploi légèrement différent (par exemple choisir les variables explicatives avant la variable dépendante). Heureusement Time Series Expert (Mélard et Pasteels, 1997) est proposé chaque fois comme alternative, surtout que l'outil complémentaire de régression oblige à placer les données des variables explicatives dans des colonnes

adjacentes et ne permet pas facilement de réaliser, ni les diagrammes des résidus (résidus en fonction des valeurs ajustées), ni le calcul de la statistique de Durbin-Watson. Dans un classeur (CH07EX09) on a ajouté une fonction DROITEREG pour obtenir les résultats dans une régression avec 19 variables explicatives. Attention : cette fonction n'est pas recalculée dans Excel 2003.

Plusieurs expériences de simulation sont réalisées en pressant la touche fonction F9, également dans les graphiques. Les déficiences du générateur de nombres aléatoires ne sont jamais apparues. Les table de données d'Excel, appelés autrefois tableaux d'hypothèse, ou opérations multiples dans Calc, permettent d'examiner l'effet de changements de valeur de paramètres, par exemple la constante de lissage pour le lissage exponentiel (par exemple CH06EX01). Il n'y a pas eu de problème ici. Cependant, dans un exercice (CH08EX04) pour le calcul d'autocorrélations sur des données artificielles basées sur des nombres pseudo-aléatoires (et accessoirement la réalisation d'un histogramme basé sur ces résultats) les tables de données fonctionnent toujours parfaitement dans Excel 2010 et suivants mais pas du tout dans OpenOffice 4.1.3, ni d'ailleurs dans Gnumeric. L'artifice qui a été trouvé (ajouter le numéro de la simulation à la donnée, ce qui n'a pas d'effet sur l'autocorrélation) fonctionne dans Gnumeric mais pas dans les opérations multiples de OpenOffice Calc. Pour assurer la compatibilité avec Calc, il a fallu réaliser une macro Visual Basic for Applications (VBA) qui force le calcul à chaque ligne du tableau et copie les résultats dans le tableau. Le problème ne se pose pas dans LibreOffice 5.2.7.

Les scénarios sont employés souvent en connexion avec le module Solver qui permet d'estimer (souvent avec un succès relatif il faut l'avouer) les paramètres de modèles non linéaires comme ceux des courbes de croissance ou de la méthode de lissage exponentiel simple ou de celle de Winters. A une exception près (CH03EX04), les résultats sont pratiquement identiques. Nous n'avons donc pas pu mesurer l'efficacité du Solver amélioré d'Excel 2010 et suivants.

Les classeurs comportent de nombreux graphiques, certains basés sur des nombres pseudo-aléatoires qui se mettent à jour par pression de F9. Compte tenu de la nature des données temporelles, il n'y a pas eu de problème ici parce que les graphiques utilisés sont essentiellement des graphes linéaires et des diagrammes de dispersion. Notons un classeur (CH02EX02) qui contient des feuilles graphiques figées, lesquelles n'apparaissent pas du tout dans Gnumeric.

Les hyperliens avaient été placés dans les classeurs, principalement dans le coin gauche de la feuille Main pour accéder rapidement à des parties de la feuille de calcul. Ces hyperliens ne fonctionnent pas dans OpenOffice.org Calc 4.1.3 mais bien dans Gnumeric. Ce n'est pas trop grave parce que le texte réfère toujours à la touche fonction F5 qui permet de se déplacer en un endroit défini de la feuille. Les liens apparaissent correctement dans LibreOffice 5.2.7. Dans Excel 2010 et suivants, il n'y a pas la barre d'outils Web de la version 2003 et l'outil Précédent n'est disponible dans aucun ruban. Pour revenir au début d'un classeur il faut donc employer la combinaison de touches Ctrl Début ou parvenir à ajouter l'outil Précédent à la barre d'accès rapide.

Les classeurs sont multilingues (français, néerlandais, anglais, avec possibilité d'ajouter d'autre langues) grâce à un artifice expliqué dans Cohen *et al*. (2003b) avec une feuille de traduction et une cellule de choix de la langue. Ceci fonctionne aussi dans les autres tableurs qu'Excel.

On recourt modérément aux macro-instructions VBA, essentiellement pour restaurer les données initiales après modification par l'apprenant mais plus fondamentalement pour commander l'une ou l'autre animation. L'utilisation des macros oblige de configurer le logiciel pour qu'il accepte les macros, et dans le cas d'OpenOffice ou LibreOffice qu'il accepte les macros VBA. Gnumeric n'accepte pas de macros VBA ou d'OpenOffice. Ce tableur est plutôt conçu pour fonctionner avec R. Les combinaisons de touches prévues dans les exercices (par exemple Ctrl Maj I) sont inopérantes dans Calc et il faut donc passer par les menus. Seul un classeur (CH04EX01) emploie des boutons, également inopérants dans Calc. On recourt aussi à du code VBA pour la copie du jeu de données à traiter (CH02EX03 et CH07EX03) et la mise en œuvre des tris pour la méthode des trois points et la méthode de Theil (CH02EX03, CH02EX04 et CH07EX03) et le contrôle de l'animation pour l'obtention des moyennes mobiles successives en ajoutant une donnée l'une après l'autre

(CH04EX01). Tout ceci ne pose pas de problème ni dans Excel 2010 et suivants ni dans OpenOffice/LibreOffice Calc 3.2 et suivants alors que ce n'était pas possible dans la version 2.0 de Calc. Comme expliqué ci-dessus, nous avons dû ajouter une macro à un classeur (CH08EX04). Il a toutefois fallu corriger les macros d'un classeur (CH04EX01) pour enlever la protection de la feuille et la remettre après. On peut considérer que Calc est difficile à employer pour un classeur (CH04EX01). Gnumeric est partiellement inutilisable pour ce même classeur ainsi que pour les trois autres qui recourent aux macros VBA pour des traitements.

En résumé, à part quelques petites erreurs, les classeurs conçus avec Excel 97 sont compatibles avec Excel 2010 et suivants et presque compatibles avec OpenOffice/LibreOffice Calc, sauf les quelques classeurs cités ci-dessus. Presque tous les classeurs du cours fonctionnent dans Gnumeric sauf, évidemment, CH04EX01 qui repose trop sur des macros.

Cette vérification a permis accessoirement de détecter et de corriger des erreurs dans quelques classeurs : CH02EX04 et CH06EX08 (erreurs dans des macros), CH03EX05 (erreur de valeurs initiales), CH05EX05 (suppression de macros inutiles), CH05EX09 et CH06EX05 (nom de macro changé), CH07EX02 (recours à des fonctions statistiques plutôt qu'à des approximations), CH07EX08 et CH07EX09 (erreurs dans des données et nettoyage). Ces deux derniers classeurs ont été profondément modifiés. La vérification a aussi permis de corriger un ensemble de données pour Time Series Expert (CH07EX09) incompatible avec le classeur correspondant.

Annexe E. Références additionnelles

[27] Almiron, M. G., Lopes, B., Oliveira, A. L. C., Medeiros, A. C. and Frery, A. C. (2010), On the numerical accuracy of spreadsheets, *Journal of Statistical Software* 34(4), 1-29.

[28] Brown, R. G. (1962), *Smoothing, forecasting and prediction of discrete time series*, Prentice-Hall, Englewood Cliffs.

[29] Hargreaves B. R. et T. P. McWilliams (2010), Polynomial trendline function flaws in Microsoft Excel, *Computational Statistics and Data Analysis* 54, 1190-1196.

[30] Heiser, D. A. (2009), Microsoft Excel 2000, 2003 and 2007 faults, problems, workarounds and fixes, http://www.daheiser.info/excel/frontpage.html. (consulté 26 février 2013).

[31] Hyndman, R. J. and Fan, Y. (1996), Sample quantiles in statistical packages, The American Statistician 50 (4), 361-365

[32] Knüsel, L. (1998), On the accuracy of statistical distributions in Microsoft Excel 97, *Computational Statistics and Data Analysis* 26, 375-377.

[33] L'Ecuyer, P. and R. Simard (2007), TESTU01: A C library for empirical testing of random number generators, *ACM Transactions on Mathematical Software* 33, 22/1-40.

[34] Levy, S. M. (2014), Fooled by pseudo-randomness, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2507276.

[35] Matsumoto, M. and Nishimura, T. (1998), MersenneTwister: a 623-dimensionally equidistributed uniform pseudorandom number generator, *ACM Transaction Modelling, Computation and Simulation* 8, 3-30.

[36] McCullough B. D. (2004c) Fixing statistical errors in spreadsheet software: the cases of Gnumeric and Excel. CSDA Statistical Software Newsletter. www.csdassn.org/software_reports.cfm.

[37] McCullough, B. D. (2008b), Microsoft Excel 2007's 'Not The Wichmann-Hill' random number generator, *Computational Statistics and Data Analysis* 52, 4587-4593.

[38] McCullough B. D. and D. A. Heiser (2008), On the accuracy of statistical procedures in Microsoft Excel 2007, *Computational Statistics and Data Analysis* 52, 4570-4578.

[39] McCullough B. D. and B. Wilson (1999), On the accuracy of statistical procedures in Microsoft Excel 97, *Computational Statistics and Data Analysis* 31, 27-37.

[40] McCullough B. D. and B. Wilson (2005), On the accuracy of statistical procedures in Microsoft Excel 2003, *Computational Statistics and Data Analysis* 49, 1244-1252.

[41] Mélard, G. (2016), On some remarks about SEATS signal extraction, SERIEs 7, 53–98

[42] Mélard G., Pasteels J.-M. (1997), Manuel d'utilisateur de Time Series Expert (TSE version 2.3), 3e édition. Bruxelles: Institut de Statistique et de Recherche Opérationnelle, Université Libre de Bruxelles.

[43] National Institute of Standards and Technology (1999), Statistical Reference Datasets: Archives, http://www.itl.nist.gov/div898/strd/general/dataarchive.html.

[44] Koteswara Rao, K. (1970), Testing for the independence of regression disturbances, *Econometrica*, 38, 1970, 97-117

[45] Sawitzki G. (1994), Report on the Numerical Reliability of Data Analysis Systems, *Computational Statistics and Data Analysis* 18(2), 289-301.

[46] Simon, Gary (2000), https://www.jiscmail.ac.uk/cgi-bin/wa.exe?A2=ind0012&L=ASSUME&F=&S=&X=31BDFD6E4E1E0E4F11&P=4325. Consulté 5 septembre 2010 (Mélard, 2014 Online resource 1).

[47] Su, Y.-S. (2008), It's easy to produce chartjunk using Microsoft Excel 2007 but hard to make good graphs, *Computational Statistics and Data Analysis* 52, 4594-4601.

[48] Vatter, P. A., Bradley, S. P., Frey, S. C. and Jackson, B. B. (1978), *Quantitative methods in management*, Irwin, Homewood Ill.

[49] Wichmann, B. A. and Hill, I. D. (1982), Algorithm AS 183: An efficient and portable pseudo-random number generator, *Journal of the Royal Statistical Society Series C Applied Statistics* 31 188-190, réimprimé avec une correction in Hill I. D. and P. Griffith (Eds) *Applied Statistics Algorithms*, Ellis Horwood, Chichester, 1985.

[50] Yalta, A. T. (2008), The accuracy of statistical distributions in Microsoft Excel 2007, *Computational Statistics and Data Analysis* 52, 4579-4586.