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.