# A graphical trap for unwary users of Excel 2010

Delwyn G. Cooke, Leonard F. Blackwell, and Simon Brown

***Abstract*—Excel's "line chart" has several peculiarities that can yield misleading representations of data. However, in Excel 2010 it is possible to combine "line charts" with "scatter charts" which exacerbates the capacity to be misleading.**

***Keywords*— Excel, graph, spreadsheet.**

## I. Introduction

Microsoft Excel is among the most widely used software in business and education [1] and, in various versions, has been for about 30 years. It has been suggested that Excel might be "…the most dangerous software on the planet" [2] because of the coding or logic errors that even expert users introduce into their spreadsheets and because of difficulties in data auditing [3]. The impact of these sorts of problem has been significant, as is apparent from the list of extraordinary examples maintained by EuSpRiG [4]. Irrespective of how Excel is used, in-built errors in calculation and statistics have been reported since at least the mid-1990s [5, 6] and this literature has grown considerably [7-15]. The intrinsic difficulties with Excel also extend to its graphics [6, 16]. The number of supported "chart types" has steadily increased and it is possible to generate many others [17-19], but the potential for the generation of poor and misleading graphs has also increased. For example, pseudo-three-dimensional charts are difficult to interpret, features of the automatic layout of the scatter plot lead to data being obscured, there is often too much "chartjunk" and a variety of other issues that mean that "charts" produced using Excel are frequently unsatisfactory.

Even the most basic of the problems identified have not always been rectified [20, 21] and some have even been carried forward into the Excel Web App [22]. Given this, and the ubiquity of Excel, the only pragmatic response open to the practioner is to become familiar with its limitations. However, beginners often find quite unlikely capabilities, so teachers must also be familiar with these. Here, we outline some of the well-known peculiarities of Excel's "line chart" and then we describe some of the consequences arising from the developments in Excel 2010.

## II. Established properties of the "line chart"

The "line chart" is described in Excel 2010 as being "… used to display trends over time". It is widely used in some disciplines, such as economics [23], and in one survey more than 85% of those charts that were not bar or pie charts (which together accounted for 60% of the total) were "line charts" of some sort [24]. However, to many students (and some who should know better) several properties of this "chart type" are not apparent from the description.

1. No matter what the spacing between the independent values the points are always evenly spaced. For example, if $x = \{1, 1.1, 10, 1000\}$ the points are evenly spaced on a "line chart", despite the enormous differences in the intervals between the coordinates.
2. The points are always located halfway between tick marks, even when the $x$ value would correspond to a tick mark. For example, if $x = \{1, 1.1, 10, 1000\}$ then $x = 1.1$ and $x = 1$ would be located midway between the adjacent pairs of tick marks.
3. The data are always distributed across the entire range of the plot.
4. The $x$ values used in these plots are actually categorical, so it would not matter if $x$ = {John, Paul, George, Ringo} rather than $x = \{1, 1.1, 10, 1000\}$, the plot would be the same except for the labels on the abscissa.
5. A corollary of this is that order is important. First, the order in which the data are listed in the spreadsheet determines the internal representation. Second, where more than one set of data is plotted on a "line chart" the order in which the data are selected determines the axis labels.
6. Despite the categorical nature of the $x$ values, Excel will fit a "trendline" to the data [12], which can be very misleading. For example, if $y = x^2$, for $x = \{-10, -9, \ldots, 10\}$, is plotted as a "line chart", the least squares second-order polynomial obtained from the data (the "trendline") is
$$y = x^2 - 22x + 121 = (x - 11)^2 \qquad (1)$$
(Figure 1). Given this, it is clear that the internal representation of the coordinates is actually $(i, y_i)$, for $i = 1, 2, \ldots, n$, irrespective of the $x$ values provided, consistent with the insensitivity of the plot to changes in labels (point 4, above). It is also clear that simply omitting or adding data can change the "trendline" dramatically. For example, omitting just one coordinate (say (-5, 25)) causes the expression to become
$$y = 1.1239x^2 - 23.478x + 122.49 \qquad (2)$$
and each omission or addition yields a different

Manuscript received December 27, 2015.

D. G. Cooke is with the Creative Diagnostics Laboratory, Massey University, Palmerston North, New Zealand (e-mail: D.G.Cooke@massey.ac.nz).

L. F. Blackwell is with the Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand (e-mail: L.F.Blackwell@massey.ac.nz).

S. Brown is with the Deviot Institute, Deviot, Tasmania 7275 Australia (e-mail: Simon.Brown@ deviotinstitute.org).

"trendline" (Figure 2) even if all points lie on a particular curve (such as $y = x^2$) because of the internal representation of the data.

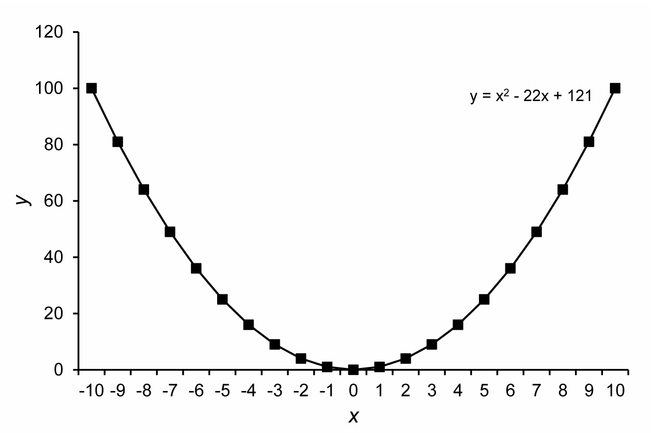

Figure 1. "Line chart" of $y = x^2$ for $x = \{-10, -9, \ldots, 10\}$ and the "trendline" reported by Excel 2010.

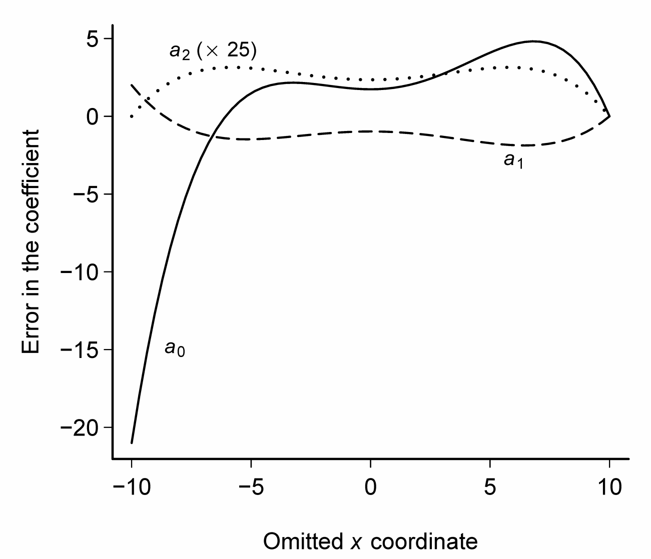

Figure 2. The error in the coefficients of the quadratic "trendline" ($y = a_2x^2 + a_1x + a_0$) fitted to the curve shown in Figure 1 when a single coordinate is omitted. In each case, the error is the difference between the coefficient reported by Excel and the "trendline" given in Figure 1 (that is $a_0 - 121$, $a_1 + 22$ and $a_2 - 1$) rather than those for the ideal fit for $y = x^2$. For clarity, the error associated with $a_2$ has been multiplied by 25.

In case it might be thought that phenomena such as these are not observed in practice, we describe two very short examples from the literature [25, 26] and omit any unnecessary details. First, Holzinger *et al.* [26] simply plot data without regard to the uneven distribution of the $x$ values and then they report the results in terms of linear least squares regression. In the second example, Patel *et al.* [25] report a least squares fit of a quadratic ($f(x)$) to data in which $x$ ranges from about 40 to 70 and $y$ ranges from about 8 to 12. The problem is obvious if a value of $x$ in the middle of the range (say $x = 55$) is chosen: $f(55) = -352.8$. Not only is this value well outside the range of the $y$ values, but, because $y$ represents a chemical concentration, it is also physically impossible. On the other hand, if $x = 3$ is chosen because of the clustering of the data: $f(3) = 10.68$. This value lies within the range of uncertainty indicated [25].

Excel provides no indication to the unwary that the "line chart" can be misleading. An experienced scientist will know that it is relatively rare to obtain data at perfectly regular intervals, but it is, perhaps, not unreasonable that many students simply think that the "line chart" is the correct way to plot a line between experimentally obtained data points. Moreover, for at least some, "line chart" is synonymous with "line graph", which is often interpreted as a plot in which a function is represented by a series of straight lines between coordinates [27], as can be achieved in Excel using a "scatter chart". The properties described above mean that the "line chart" does not plot the data reliably which deters most experienced users of Excel for scientific purposes from employing this "chart type" without a very good reason.

## III. What Excel 2010 has to teach even experienced users

In Excel 2010 the "line chart" has three properties we have not encountered before.

1. Excel facilitates the combination of a "scatter chart" and a "line chart". For example, having made a "scatter chart" of two sets of data, one can be converted to a "line chart" by right clicking on a data point, selecting "Change Series Chart Type …" from the pop up menu and then "line chart" from the options displayed. To illustrate the consequences of this, in the example shown in Figure 3 the same data were plotted twice on a "scatter chart" and then one series was converted to a "line chart". Two unexpected features are apparent from Figure 3.
   a. The data in these two representations have different ranges: the original data ranged from $x = -11$ to $x = 13$, but in the "line chart" the range is from $x \approx -14.5$ to $x \approx 14.5$.
   b. Only one of the 25 points (that at $x = 6$) coincides in the two representations. In fact the "line chart" is stretched relative to the "scatter chart" (for $x > 6$ or $x < 6$, each point in the "line chart" is plotted at too high or low, respectively, an x value and the discrepancies increase as $x$ diverges from $x = 6$).
2. The discrepancy between the actual and correct x coordinates can be modified by changing the range of the abscissa (Figure 4).
3. The precise behaviour observed need not appear to be consistent.

These three properties are consequences of the basic properties of the "line chart" listed above and the scaling of the "scatter chart". In Figure 3, the default range of the "scatter chart" abscissa is $x = -15$ to $x = 15$, not all of which is required by the data. In contrast, in the "line chart" representation the data are always uniformly distributed over the entire plot width, so the two representations do not coincide in Figure 3 and the discrepancy can be made worse (Figure 4) or reduced by changing the range of the "scatter chart". The apparently inconsistent behaviour of the combination of a "scatter chart" and a "line chart" results from the significance of order for the latter. While it might seem unlikely that a combination of a "line graph" and a "scatter plot" would be considered useful, we have encountered examples of this recently and its potential to be misleading is profound.

The greatest risk associated with this particular phenomenon is that it is almost impossible to detect without either (i) direct access to the raw data or (ii) plots of the data that make the discrepancy obvious, such as those in Figures 3 and 4. It is rare that either of these options would be available, leading one to conclude, for example, that the data in Figure 3 are not particularly consistent and that the discrepancy is even worse in Figure 4. In fact, the data are identical and should match perfectly both within and between the two graphs (Figures 3 and 4).

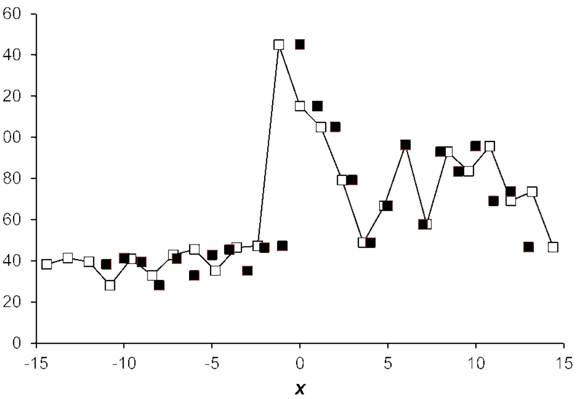

Figure 3. “Scatter chart” (■) and superimposed “line chart” (—□—) of the same data generated using the default scaling in Excel 2010. We omit details of the work from which the data were obtained, but details of very similar data can be obtained elsewhere [28].

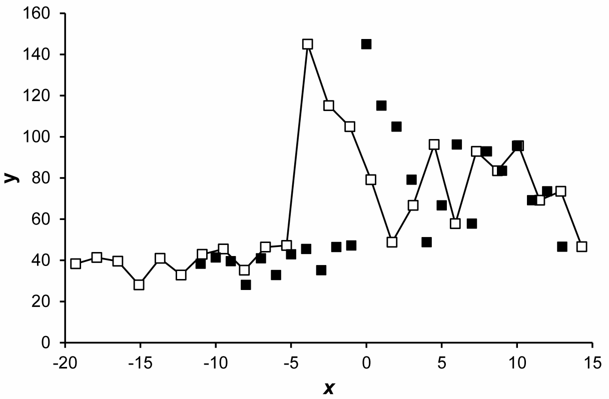

Figure 4. “Scatter chart” (■) and superimposed “line chart” (—□—) of the same data plotted in Figure 3 generated by arbitrarily changing the lower limit of the abscissa from -15 to -20.

## IV. Conclusion

The use of the “line chart” in Excel should be considered very carefully. Everything that can be done with a “line chart” can be achieved with a little thought using a “scatter chart”. However, the real danger in Excel 2010 is the capacity to combine a “line chart” and a “scatter chart” which can be virtually undetectable and have unanticipated and serious consequences.

## References

[1] D. Pianjud, O. Natakuatoong, and J. Vicheanpanya, “The development of a knowledge management system to promote the sufficiency economy philosophy for the basic education teacher,” *International Journal of e-Education, e-Business, e-Management and e-Learning*, vol. 3, pp. 219-223, 2013. http://dx.doi.org/10.7763/IJEEEE.2013.V3.227

[2] T. Worstall, “Microsoft’s Excel might be the most dangerous software on the planet,” in Forbes, 2013. http://onforb.es/12LAYaI

[3] M. P. Campbell, “Spreadsheet issues: pitfalls, best practices, and practical tips,” *Actuarial Practice Forum*, vol. 2010, pp. 1-39, 2010. https://www.soa.org/library/journals/actuarial-practice-forum/2010/february/apf-2010-02-toc.aspx#

[4] EuSpRiG, “Spreadsheet mistakes – news stories collated by the European Spreadsheet Risks Interest Group,” vol. 2015: European Spreadsheet Risks Interest Group, 2013. http://www.eusprig.org/horror-stories.htm

[5] G. Sawitzki, “Report on the numerical reliability of data analysis systems,” *Computational Statistics and Data Analysis*, vol. 18, pp. 289-301, 1994. http://dx.doi.org/10.1016/0167-9473(94)90177-5

[6] J. C. Nash and T. K. Quon, “Issues in teaching statistical thinking with spreadsheets,” *Journal of Statistics Education*, vol. 4, pp. np, 1996. http://www.amstat.org/publications/jse/v4n1/nash.html

[7] J. C. Nash, “Spreadsheets in statistical practice – another look,” *American Statistician*, vol. 60, pp. 287-289, 2006. http://dx.doi.org/10.1198/000313006X126585

[8] R. L. Berger, “Nonstandard operator precedence in Excel,” *Computational Statistics and Data Analysis*, vol. 51, pp. 2788-2791, 2007. http://dx.doi.org/10.1016/j.csda.2006.09.040

[9] B. D. McCullough and B. Wilson, “On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP,” *Computational Statistics and Data Analysis*, vol. 40, pp. 713-721, 2002. http://dx.doi.org/10.1016/S0167-9473(02)00095-6

[10] B. D. McCullough and D. A. Heiser, “On the accuracy of statistical procedures in Microsoft Excel 2007,” *Computational Statistics and Data Analysis*, vol. 52, pp. 4570-4578, 2008. http://dx.doi.org/10.1016/j.csda.2008.03.004

[11] A. T. Yalta, “The accuracy of statistical distributions in Microsoft® Excel 2007,” *Computational Statistics and Data Analysis*, vol. 52, pp. 4579-4586, 2008. http://dx.doi.org/10.1016/j.csda.2008.03.005

[12] B. D. Hargreaves and T. P. McWilliams, “Polynomial trendline function flaws in Microsoft Excel,” *Computational Statistics and Data Analysis*, vol. 54, pp. 1190-1196, 2010. http://dx.doi.org/10.1016/j.csda.2009.10.020

[13] L. Knüsel, “On the accuracy of statistical distributions in Microsoft Excel 2003,” *Computational Statistics and Data Analysis*, vol. 48, pp. 445-449, 2005. http://dx.doi.org/10.1016/j.csda.2004.02.008

[14] M. G. Almiron, B. Lopes, A. L. C. Oliviera, A. C. Medeiros, and A. C. Frery, “On the numerical accuracy of spreadsheets,” *Journal of Statistical Software*, vol. 34, pp. 1-29, 2010. http://dx.doi.org/10.18637/jss.v034.i04

[15] G. Mélard, “On the accuracy of statistical procedures in Microsoft Excel 2010,” *Computational Statistics*, vol. 29, pp. 1095-1128, 2014. http://dx.doi.org/10.1007/s00180-014-0482-5

[16] Y.-S. Su, “It’s easy to produce chartjunk using Microsoft® Excel 2007 but hard to make good graphs,” *Computational Statistics and Data Analysis*, vol. 52, pp. 4594-4601, 2008. http://dx.doi.org/10.1016/j.csda.2008.03.007

[17] J. Baker, “The charts that Excel cannot do,” *Spreadsheets in Education*, vol. 1, pp. 6, 2004. http://epublications.bond.edu.au/ejsie/vol1/iss3/6/

[18] J. Benacka, “Introduction to 3D graphics through Excel,” *Informatics in Education*, vol. 12, pp. 221-230, 2013. http://www.mii.lt/informatics_in_education/htm/INFE226.htm

[19] C. Duller, “Teaching statistics with Excel. A big challenge for students and lecturers,” *Austrian Journal of Statistics*, vol. 37, pp. 195-206, 2008. http://www.stat.tugraz.at/AJS/ausg082/082Duller.pdf

[20] B. D. McCullough, “Does Microsoft fix errors in Excel?,” presented at 2001 Joint Statistical Meetings [CD-ROM], Alexandria, 2002. http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00177.pdf

[21] B. D. McCullough, “The unreliability of Excel’s statistical procedures,” *Foresight*, vol. 2006, pp. 44-45, 2006. http://www.forecastingprinciples.com/files/McCullough.pdf

[22] B. D. McCullough and A. T. Yalta, “Spreadsheets in the cloud – not ready yet,” *Journal of Statistical Software*, vol. 52, pp. 1-14, 2013. http:// www.jstatsoft.org/article/view/v052i07/v52i07.pdf

[23] J. A. Schwabish, “An economist’s guide to visualizing data,” *Journal of Economic Perspectives*, vol. 28, pp. 209-234, 2014. http://dx.doi.org/10.1257/jep.28.1.209

[24] C. Chambers and C. Scaffidi, “Struggling to Excel: a field study of challenges faced by spreadsheet users,” in *Proceedings. 2010 IEEE symposium on visual languages and human-centric computing (VL/HCC 2010)*, C. Hundhausen, E. Pietriga, P. Diaz, and M. B. Rosson, Eds. Los Alamitos: Institute of Electrical and Electronics Engineers, Inc., 2010, pp. 187-194. http://dx.doi.orh/10.1109/VLHCC.2010.33

[25] S. S. Patel, M. Z. Molnar, J. A. Tayek, J. H. Ix, N. Noori, D. Benner, S. Heymsfield, J. D. Kopple, C. S. Kovesdy, and K. Kalantar-Zadeh, “Serum creatinine as a marker of muscle mass in chronic kidney disease: results of a cross-sectional study and review of literature,” *Journal of Cachexia, Sarcopenia and Muscle*, vol. 4, pp. 19-29, 2013. http://dx.doi.org/10.1007/s13539-012-0079-1

[26] U. Holzinger, R. Brunner, H. Losert, V. Fuhrmann, C. Modl, F. Sterz, and B. Schneeweiβ, “Resting energy expenditure and substrate oxidation rates correlate to temperature and outcome after cardiac arrest – a prospective observational cohort study,” *Critical Care*, vol. 19, pp. 128, 2015. http://dx.doi.org/10.1186/s13054-015-0856-2

[27] Ministry of Education, *Mathematics standards for years 1-8*. Wellington: Learning Media Limited, 2009. http://nzcurriculum.tki.org.nz/content/download/3166/47235/file/Maths_Standards_amended_vs3.pdf

[28] L. F. Blackwell, P. Vigil, M. E. Alliende, S. Brown, M. Fortin, and D. G. Cooke, “Monitoring of ovarian activity by measurement of urinary excretion rates using the Ovarian Monitor, Part IV: the relationship of the pregnanediol glucuronide threshold, basal body temperature and cervical mucus as markers for the beginning of the post-ovulatory infertile period” *Human Reproduction*, vol. 31, pp. 445-453, 2016. http://dx.doi.org/10.1093/humrep/dev303