The use of Charts, Pivot Tables, and Array Formulas in two Popular Spreadsheet Corpora

Jansen, Bas; Hermans, Felienne

Publication date
2018

Document Version
Accepted author manuscript

Published in
Proceedings of the 5th International Workshop on Software Engineering Methods in Spreadsheets

The use of Charts, Pivot Tables, and Array Formulas in two Popular Spreadsheet Corpora

Bas Jansen
Delft University of Technology
Email: b.jansen@tudelft.nl

Felienne Hermans
Delft University of Technology
Email: f.f.j.hermans@tudelft.nl

***Abstract*—The use of spreadsheets in industry is widespread. Companies base decisions on information coming from spreadsheets. Unfortunately, spreadsheets are error-prone and this increases the risk that companies base their decisions on inaccurate information, which can lead to incorrect decisions and loss of money. In general, spreadsheet research is aimed to reduce the error-proneness of spreadsheets. Most research is concentrated on the use of formulas. However, there are other constructions in spreadsheets, like charts, pivot tables, and array formulas, that are also used to present decision support information to the user. There is almost no research about how these constructions are used. To improve spreadsheet quality it is important to understand how spreadsheets are used and to obtain a complete understanding, the use of charts, pivot tables, and array formulas should be included in research. In this paper, we analyze two popular spreadsheet corpora: Enron and EUSES on the use of the aforementioned constructions.**

I. Introduction

The use of spreadsheets in industry is widespread. It is estimated that 90% of all analysts use spreadsheets for their calculations [1] and 95% of US firms use spreadsheets in some form of financial reporting [2]. This reporting and the calculations of annalists in spreadsheets are used by companies to make decisions. Unfortunately, spreadsheets are known to be error-prone [3]. This increases the risk that companies make decisions that are based on inaccurate information which can lead to incorrect decisions and loss of money[1].

A substantial part of current spreadsheet research is focused on improving the quality of spreadsheets by applying software engineering methods to them like testing [4] [5], reverse engineering [6] [7], code smells [8] [9], and refactoring [10] [11]. This research has in common that it concentrates on spreadsheet formulas. However, there are other constructions in spreadsheets like charts, pivot tables, and array formulas, that are also used to present decision support information to the user of the spreadsheet. We do not know how widespread the use of these constructions are and if they are somehow related to the error-proneness of spreadsheets. There is almost no research on this topic. The only study we found is the paper of Fisher and Rothermel in which they introduce the EUSES corpus. They counted the number of spreadsheets that contained charts.

To improve the overall quality of spreadsheets, we need to understand how spreadsheets are used in industry, and to obtain a complete understanding we should include constructions like charts, pivot table and array formulas in our research. Therefore, in this short paper, we analyze two popular spreadsheet corpora: Enron [12] and EUSES [13] on the use of charts, pivot tables, and array formulas.

The contributions of this paper are: 1) a tool that is capable of analyzing properties of charts, pivot tables, and array formulas in Excel files and 2) an analysis of the use of charts, pivot tables, and array formulas in the EUSES and Enron corpora.

We organize the remainder of this paper in the following way. In the next section, we describe the approach that was used to analyze the spreadsheets. In Section III we present the results of the analysis. Related work is discussed in Section IV and we end the paper with the concluding remarks in Section V.

II. Approach and Tool

In this study we use two popular spreadsheet corpora: EUSES [13] and Enron [12]. In previous work [14], [15], [16] we used the Spreadsheet Scantool, developed at Delft University of Technology. It utilizes the Gembox library to read spreadsheets. Unfortunately, Gembox is not able to analyze charts, pivot tables or array formulas. As a result, we developed a new Analyzer in Visual Basic for Applications (VBA) that is able to access the full Excel object model. We made the VBA code together with instructions about how to use it, available on Github[2]. With the Analyzer, we collect several metrics about the use of charts, pivot tables, and array formulas in spreadsheets in the aforementioned corpora.

For charts, we counted the number of spreadsheets that contained at least one chart and the total number of charts in the corpus. For each chart, we determined the chart type. In Excel there are 73 different chart types[3]. However, most chart types are a variation of their base type. For example, the 'radar', 'filled radar' and 'radar with data markers' all belong to the category of radar charts. In our analysis, we mapped the actual chart type to its category.

Also for pivot tables, we counted the number of spreadsheets that contained at least one pivot table. Next, for each pivot table we analyzed:

[1]http://www.eusprig.org/horror-stories.htm

[2]https://github.com/HeerBommel/SpreadsheetExplorer.git

[3]https://msdn.microsoft.com/en-us/vba/excel-vba/articles/xlcharttype-enumeration-excel

- the size: by counting the number of rows and columns of the underlying dataset
- the number of calculated fields: calculated fields are formulas that can refer to other fields in the pivot table
- the number of calculated items: calculated items are formulas that can refer to other items within a specific pivot field
- the aggregation functions used
- the number of pivot tables per worksheet

With respect to array formulas, we counted the number of spreadsheets that contained at least one array formula and we analyzed all array formulas to obtain a better understanding of why they are used.

III. Results

In this section, we will present the results of our analysis. The spreadsheets from both the Enron and the EUSES corpus were scanned for the usage of charts, pivot tables, and array formulas. For each construct, a more detailed analysis was executed to gain more insight into how they were used.

A. Charts

Table I shows the total number of spreadsheets we have analyzed. Some of the files were password protected, corrupted or otherwise unreadable and we have excluded them from the study (25 files in EUSES and 130 in the Enron dataset). Charts are not rare but the majority of spreadsheets do not contain any charts. They occur more in the Enron spreadsheets. This could be an indication that charts are used more frequently in an industrial setting.

TABLE I
Number of spreadsheets with and without charts

	EUSES	%	Enron	%
Charts	340	8%	1,721	11%
No Charts	4,133	92%	14,078	89%
Total	4,473	100%	15,799	100%

Within Excel there are two ways to create a chart: 1) the chart is created as a special type of worksheet, this is called a Chart Sheet or 2) the chart is embedded on an existing worksheet. Table II presents the occurrence of the two types in both corpora. When Fisher and Rothermel introduced the EUSES corpus, they stated that 105 of the 4,498 spreadsheets contained charts [13]. We found a different number. According to our analysis, there are 340 spreadsheets that contain charts. The difference is caused by the two different ways a chart can be created. Fisher and Rothermel only counted the chart sheets, while our analysis also included the embedded charts.

TABLE II
The use of Chart Sheets vs Embedded Charts

	EUSES		Enron	
Type	# Charts	%	# Charts	%
Sheet	355	25%	1,149	13%
Embedded	1,090	75%	7,686	87%
Total	1,445	100%	8,835	100%

Tables III and IV show the different chart types that were used in both corpora. The most frequently used type is the column chart. Other popular types in both corpora are the line chart and the pie chart. Notable differences between the two corpora are the occurrence of the mixed and scatter chart types. The scatter chart is used frequently in the EUSES corpus (22%), but less frequently in the Enron set (2%). The opposite is true for the mixed chart type: it is frequently used within Enron but hardly in the EUSES set. A mixed chart type means that either two y-axes are used or there are two or more data series that use a different chart type (for example the combination of a line and a column chart), see Figure 1 for an example.

TABLE III
Overview of used chart types in EUSES

Rank	Chart Type	# Charts	%
1	Column	515	35.6%
2	Scatter	322	22.3%
3	Line	258	17.9%
4	Pie	139	9.6%
5	Bar	125	8.7%
6	Mixed	55	3.8%
7	Surface	15	1.0%
8	Area	13	0.9%
9	Stock	2	0.1%
10	Radar	1	0.1%

TABLE IV
Overview of used chart types in Enron

Rank	Chart Type	# Charts	%
1	Column	4,253	48.1%
2	Line	2,815	31.9%
3	Mixed	866	9.8%
4	Pie	649	7.3%
5	Scatter	168	1.9%
6	Area	67	0.8%
7	Bar	11	0.1%
8	Surface	6	0.1%

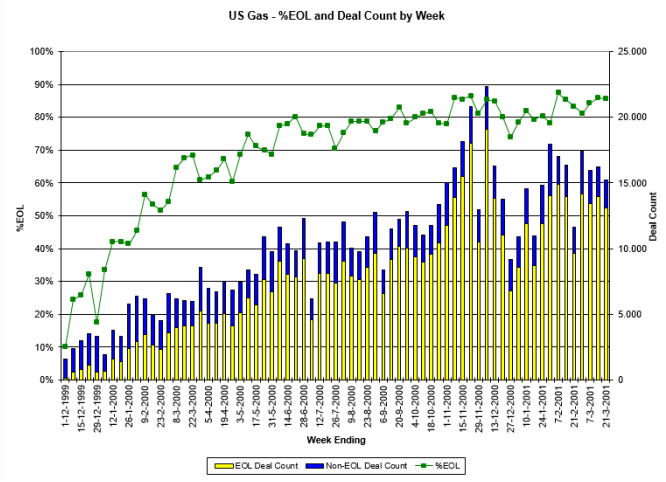

Fig. 1. Mixed chart type (Source: Enron Corpus scott_neal__38010__Helplist.xls, sheet US Gas)

TABLE V
NUMBER OF SPREADSHEETS WITH PIVOT TABLES

	EUSES	%	Enron	%
Pivot	19	0.4%	244	1.5%
No Pivot	4,454	99.6%	15,555	98.5%
Total	4,473	100.0%	15,799	100.0%

TABLE VI
OVERVIEW OF PIVOT TABLE METRICS

Corpus	avg. # records	avg. # fields	# calc. items	# calc. fields
EUSES	3,738	13	3	0
Enron	4,073	28	0	0

B. Pivot Tables

Table V shows the number of spreadsheets in both corpora that contain pivot tables. It shows that pivot tables are used very rarely but that they are used more frequently in the industrial Enron set than in the EUSES spreadsheets.

Pivot tables can be used to analyze large datasets. Therefore, we analyzed the average size of the data tables behind the pivot tables (Table VI). The maximum number of records we found is equal to the maximum number of rows up to Excel 2003 (65,535 rows). The average number of records in both datasets are comparable, but in the Enron sheets, the average number of columns (fields) in these data sets is larger. Within a pivot table, it is possible to create calculated fields or items based on other data in the table. From Table VI we can observe that this functionality is hardly used.

The main functionality that a pivot table provides is the possibility to summarize large data sets with aggregate functions. We analyzed the use of these functions and found that, in both corpora, in more than 85% of the cases, either the function Sum or Count was used to summarize the data.

When a pivot table is defined in Excel, by default it is created on a new worksheet. However, it is possible to create more than one pivot table on a worksheet and based on our analysis we can conclude that users tend to do so. The average number of pivot tables per worksheet ranges from 1.3 (EUSES) to 1.6 (Enron) and we found a maximum of nine pivot tables on one worksheet.

C. Array Formulas

Formulas in Excel always return a single value. However, with an array formula[4], it is possible to perform multiple calculations on one or more items of an array and return either a single value or multiple values. The formulas have to be entered with a special key combination (ctrl + shift + enter) and most users are unaware of their existence. Nevertheless, we checked in both corpora for array formulas and to our surprise, we found that their occurrence is similar to that of pivot tables. About 1.2% of the spreadsheets (in both corpora) contains array formulas.

[4]https://support.office.com/en-us/article/Guidelines-and-examples-of-array-formulas-7d94a64e-3ff3-4686-9372-ecfd5caa57c7?ui=en-US&rs=en-US&ad=US

In the 250 spreadsheets that contained one or more array formulas, we found a total of 3,965 unique array formulas. We have analyzed these formulas to obtain a better understanding of why array formulas are used. Table VII summarizes the results.

TABLE VII
USE CASES FOR ARRAY FORMULAS

Use case	# Formulas	% of Formulas
Multiple criteria aggregation	2,398	60.5%
Array functions	690	17.4%
Sumproduct	455	11.5%
User defined functions	264	6.7%
Unnecessary or erroneous	66	1.7%
Repeat block	47	1.2%
TABLE function	45	1.1%
Total	3,965	100.0%

In over 60% of the cases, array formulas are used to calculate aggregated values with functions like SUM, MIN, and AVERAGE, based on multiple criteria. An example of such a formula is shown in Figure 2.

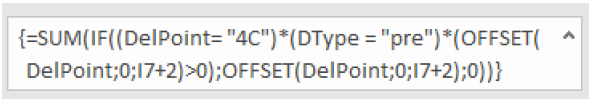

Fig. 2. Array formula for a multiple criteria SUM (Source: Enron Corpus eric_linder__9655__4_02act.xlsx, sheet Preschedule, cell I30)

The curly brackets indicate that the formula is an array formula. This specific use case is related to the age of both the Enron and the EUSES corpus. At the time the spreadsheets in the Enron and EUSES corpus were created, there were no functions available for conditional aggregation of values and it could only be accomplished by using an array formula or the SUMPRODUCT function. However, as from Excel 2010, Microsoft has added dedicated functions like SUMIFS, AVERAGEIFS, MINIFS, etc. for these type of calculations.

Another important use case for array formulas are the special array functions in Excel, like TRANSPOSE, MMULT, LINEST, FREQUENCY, and MINVERSE. Instead of returning a single value, these functions return an array as result and can only be used in an array formula.

Sometimes an array formula is entered as a multi-cell array formula. In such a case a group of cells gets exactly the same array formula and this formula can only be edited when the whole group is selected. The formula will calculate the result for each item in the array of cells. We grouped this use case under the category repeat block.

Other use cases we found were user-defined functions that are used as array formulas and what-if analyses with a data table (TABLE function [5]). Finally, we encountered a set of array formulas that were either incorrect or unnecessary. The latter meaning that the same formula would also have worked without the special array syntax.

[5]https://support.office.com/en-us/article/calculate-multiple-results-by-using-a-data-table-e95e2487-6ca6-4413-ad12-77542a5ea50b

IV. Related Work

Most related to our work are the papers introducing the corpora that we have analyzed, Euses [13] and Enron [12]. While the paper introducing EUSES describes statistics on charts too the paper on Enron does not. Another spreadsheet corpus is FUSE [17]. This corpus consists of almost 250,000 spreadsheets that were extracted from a public web archive with over 26 billion pages. We preferred the Enron corpus over the FUSE corpus because the Enron spreadsheets were used in an industrial setting.

Besides those two corpora, there are other smaller corpora: Two prominent corpora are the Galumpke and Wall corpora, containing 82 and 150 spreadsheets respectively, both derived from classroom experiments [18]. Powell and colleagues survey other corpora used in field audits, each audit examining between 1 and 30 spreadsheets [19]. To our knowledge, none of these corpora are publicly available.

Furthermore, there are papers on spreadsheet metrics, which also measure properties of spreadsheets. In 2004, Bregar published a paper presenting a list of spreadsheet metrics based on software metrics [20]. He, however, does not provide any justification of the metrics, nor did he present an evaluation.

V. Concluding Remarks

In this paper, we analyzed two popular spreadsheet corpora and focussed on the use of charts, pivot tables, and array formulas. Although the spreadsheets in both corpora are more than ten years old, we believe they still offer a valuable insight into how they are used. Especially because Microsoft did not make a lot of changes with respect to charts, pivot tables, and array formulas.

We found that charts are used in about 10% of the spreadsheets. The most commonly used chart types are Column, Line, and Pie. Pivot tables are much rarer and only found in about 1% of the spreadsheets. In more than 85% of the cases, the data in the pivot table is summarized with the aggregate functions *Sum* and *Count*. Finally, to our surprise, the complex array formulas are used as frequently as pivot tables.

Overall the use of charts, pivot tables, and array formulas in spreadsheets is limited. Still, they can have an impact on the error-proneness of spreadsheets. Especially pivot tables and array formulas can introduce new types of errors. A well-known problem with pivot tables is that they are not automatically refreshed when the underlying data changes, increasing the risk that the user is analyzing outdated data. Array formulas are difficult to understand and not very well known by the majority of the spreadsheet users. Because of their complexity, it is easy to make errors in these formulas. The information presented in the charts is coming from the underlying data. Errors in this data will lead to errors in charts. However, charts also can introduce their own errors. For example, Excel will in some cases automatically let the Y axis not start at zero, which could misrepresent the underlying data.

References

[1] W. Winston, "Executive education opportunities millions of analysts need training in spreadsheet modeling, optimization, monte carlo simulation and data analysis," *OR MS TODAY*, vol. 28, no. 4, pp. 36–39, 2001.

[2] R. R. Panko and N. Ordway, "Sarbanes-oxley: What about all the spreadsheets?" *arXiv preprint arXiv:0804.0797*, 2008.

[3] R. R. Panko, "What we know about spreadsheet errors," *Journal of Organizational and End User Computing (JOEUC)*, vol. 10, no. 2, pp. 15–21, 1998.

[4] K. J. Rothermel, C. R. Cook, M. Burnett, J. Schonfeld, T. R. Green, and G. Rothermel, "Wysiwyt testing in the spreadsheet paradigm: An empirical evaluation," in *Software Engineering, 2000. Proceedings of the 2000 International Conference on*. IEEE, 2000, pp. 230–239.

[5] S. Roy, F. Hermans, and A. van Deursen, "Spreadsheet testing in practice," in *Software Analysis, Evolution and Reengineering (SANER), 2017 IEEE 24th International Conference on*. IEEE, 2017, pp. 338–348.

[6] F. Hermans, M. Pinzger, and A. Deursen, *ECOOP 2010 – Object-Oriented Programming: 24th European Conference, Maribor, Slovenia, June 21-25, 2010. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ch. Automatically Extracting Class Diagrams from Spreadsheets, pp. 52–75. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-14107-2_4

[7] J. Cunha, M. Erwig, and J. Saraiva, "Automatically inferring classsheet models from spreadsheets," in *Visual Languages and Human-Centric Computing (VL/HCC), 2010 IEEE Symposium on*. IEEE, 2010, pp. 93–100.

[8] F. Hermans, M. Pinzger, and A. van Deursen, "Detecting and refactoring code smells in spreadsheet formulas," *Empirical Software Engineering*, pp. 1–27, 2014.

[9] J. Cunha, J. P. Fernandes, H. Ribeiro, and J. Saraiva, "Towards a catalog of spreadsheet smells," in *Computational Science and Its Applications–ICCSA 2012*. Springer, 2012, pp. 202–216.

[10] F. Hermans and D. Dig, "Bumblebee: a refactoring environment for spreadsheet formulas," in *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM, 2014, pp. 747–750.

[11] S. Badame and D. Dig, "Refactoring meets spreadsheet formulas," in *Software Maintenance (ICSM), 2012 28th IEEE International Conference on*. IEEE, 2012, pp. 399–409.

[12] F. Hermans and E. Murphy-Hill, "Enron's spreadsheets and related emails: A dataset and analysis," in *Proceedings of the 37th International Conference on Software Engineering-Volume 2*. IEEE Press, 2015, pp. 7–16.

[13] M. Fisher and G. Rothermel, "The EUSES spreadsheet corpus: a shared resource for supporting experimentation with spreadsheet dependability mechanisms," *ACM SIGSOFT Software Engineering Notes*, vol. 30, no. 4, pp. 1–5, 2005.

[14] B. Jansen, "Enron versus euses: A comparison of two spreadsheet corpora," in *Proceedings of the 2nd Workshop on Software Engineering Methods in Spreadsheets, Florence, Italy*, 2015.

[15] F. Hermans, M. Pinzger, and A. Deursen, "Detecting code smells in spreadsheet formulas," *Proceedings of the International Conference on Software Maintenance (ICSM)*, 2012.

[16] B. Jansen and F. Hermans, "Code smells in spreadsheet formulas revisited on an industrial dataset," in *Proceedings of the 2015 International Conference on Software Maintenance and Evolution*. IEEE Press, 2015, pp. 372–380.

[17] T. Barik, K. Lubick, J. Smith, J. Slankas, and E. Murphy-Hill, "F use: a reproducible, extendable, internet-scale corpus of spreadsheets," in *Proceedings of the 12th Working Conference on Mining Software Repositories*. IEEE Press, 2015, pp. 486–489.

[18] R. R. Panko, "Two corpuses of spreadsheet errors," in *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*. IEEE, 2000, pp. 8–pp.

[19] S. G. Powell, K. R. Baker, and B. Lawson, "A critical review of the literature on spreadsheet errors," *Decision Support Systems*, vol. 46, no. 1, pp. 128–138, 2008.

[20] A. Bregar, "Complexity metrics for spreadsheet models," in *Proc. of EuSpRIG '04*, 2004, p. 9.