Evolution of Statistical Software and Quantitative Methods

Brandon LeBeau[1] & Ariel M. Aloe[1]

[1] University of Iowa

Author Note

Correspondence concerning this article should be addressed to Brandon LeBeau, Psychological and Quantitative Foundations, University of Iowa, Iowa City, IA 52245. E-mail: brandon-lebeau@uiowa.edu

Abstract

Statistical software is the enabling tool of quantitative research studies and the availability and use of the software can greatly shape which methods are used by researchers. Software that is more accessible is likely to have more users and the methods implemented within the software limits the methods accessible to researchers. Open source software, (e.g. R), has reduced these barriers by making cutting edge statistical methods available to researchers through add-on packages. This paper aims to explore the evolution of statistical software within social science research using a research synthesis to establish the state of affairs.

Keywords: Research Synthesis, Statistical Software, Quantitative Methods

Evolution of Statistical Software and Quantitative Methods

How statistical software is used in primary research studies has important implications for the reproducibility and replicability of the research. Without knowledge of which software and version is used in the analysis, differences in results could be due to different software implementations, maybe not the analysis. To our knowledge, in the social sciences, no research has actively explored which software are commonly used in published research or how often research papers cite the software they use.

Software usage is only a part of the puzzle to be able to reproduce or replicate study results. The statistical method(s) used are also important to have a clear distinction about how they are used. Without clear, transparent, and open statistical methods, data processing, and other data management tasks, the ability to reproduce or replicate the study drops significantly (Peng, 2009, 2011; Stodden, 2012).

The purpose of this paper is to explore the evolution (or lack thereof) of statistical software usage, statistical methods usage, and the interaction between the two over time in the social sciences. Research synthesis methods (Cooper, 2016) will be used to explore these trends over time in published articles found in twelve social science research journals. Social science journals spanning economics, education, political science, psychology, public policy, and sociology that had strong impact factors were targeted for inclusion.

Statistical Software and Reproducibility

Statistical software is an enabling tool to performing applied data analysis. Statistical methods that are implemented within software will increase their usage (particularly if the software is also user-friendly) by applied analysts and are likely taught more frequently in methodology courses at universities. Moreover, casual users of statistics software may not distinguish between the limitations of the models and the limitations of the software. However, the user-friendly nature of software (i.e., point and click graphical interfaces,

ability to manipulate data by hand) also can severely limit the ability for research to be reproducible; a recent topic of intense discussion in biostatistics, medicine, and pyschology (Asendorpf et al., 2013; Ioannidis, 2014; Iqbal, Wallach, Khoury, Schully, & Ioannidis, 2016; Peng, 2009, 2011; Stodden, 2012). The replicability and reproducibility crisis has pointed the finger at statistical software more directly with a strong emphasis in some disciplines for analyses to be script (i.e., source code) based and posted with the published journal article, often described as the gold standard.

This idea of reproducibility has not seemed to fully enter the social science research domain. In these areas, replication has seen more focus in the literature (add citations about replication in SS research). SPSS is likely the most common statistical software program used in many social science research domains, particularly education. Although SPSS has many common and advanced statistical techniques and it is possible to have a reproducible analysis, the default behavior within SPSS is not script based and can create bad habits. For example, editing raw data directly in the graphical user interface, running analyses without saving a script, creation of variables without syntax, or marking values as missing or not possible. Statistical software that is primarily command line, programs such as R, Python, SAS, or Stata, offer easier reproducible frameworks as all data manipulations or analyses are saved in scripts that can be re-ran in the future. A data script can be thought of as a cockpit flight recorder in which every single step that was done to the original data going from data collection to final tables and figures was script based. Under a reproducible framework, the raw data are never altered directly, they should always be altered programmatically through a script. This keeps a log of the data manipulations that happened in the data analysis cycle. For example, R has packages that aid in the ability to create living data documents that contain text and analysis code within a single document (see Allaire et al., 2017; Xie, 2015, 2017).

The reproducible analysis framework has many advantages, including a transparent

analysis process that could be validated by others or even simply the ability to investigate the data analysis completed months or years previously. Unfortunately, the current academic research framework has many barriers that limit the reproducibility. First, applied researchers may not be users of primarily command line or script based statistical software. This limits the ability to create a reproducible framework from the start. Secondly, researchers are not incentivized to conduct an analysis in a reproducible framework. Namely, the publish or perish aspect of academic research limits the sharing of statistical code partly due to the increased chance of criticism upon evaluation of the code used for the analysis. Finally, many journals and even the American Psychological Association (APA) publication manual (American Psychological Association, 2010) states that common software or programming languages need not be cited. This could even be interpreted by some as not needing to mention. Unfortunately, if the software used for a data analysis is not reported, the ability to recreate the analysis drops even more due to differences in estimation, handling of missing data, or other software specific settings.

This paper aims to explore the state of affairs in statistical software usage in education. Particular attention will be made to which software is currently being used in published social science research as well as how this has changed over the last twenty years. Secondly, this paper also aims to explore how frequently open-source software tools are used and to explore evidence of reproducible analysis framework being implemented. These aims will be explored using research synthesis methods.

Statistical Software Usage

Research on the usage of statistical software has rarely been undertaken. Muenchen (2017) has explored the popularity of data science software through job advertisements, scholarly articles, and other metrics. This exploration has shown that job advertisements for R and Python have increased between 2012 and 2017, but job advertisements for SAS have stayed relatively steady over this time frame based on data using Indeed's job trends tool.

The number of scholarly articles, tracked through a Google Scholar search, mentioning research software was also explored between 1995 and 2016. The data show that SPSS has dominated between the years 2000 and 2010, but since 2010 has steadily decreased although still remains the most popular mentioned software tool. SAS has been second for most of this time frame and has followed a similar trajectory as SPSS. Recently, R has passed SAS in how often it is used in 2016. STATA has also shown strong increases as well, but not quite as steep an increase as R has shown.

The current study differs and expands from those conducted by Muenchen (2017). The current analysis focuses on social science research exclusively, which is narrowed compared to the analysis conducted by Muenchen (2017). Secondly, the current study also aims to explore statistical methods used and the interaction between statistical methods and software usage. Finally, through the methodology by Muenchen (2017), it is not directly possible to explore how many studies are not mentioning statistical software usage, but still mentioning the usage of statistical methods. Statistical software implementation has implications for the reproducibility and replicability of the analysis.

Statistical Methods Usage

Research by Tatman (2015) on the usage of statistical methods usage in linguistics. A total of 348 articles from the most recent issue of a variety of linguistics journals were explored. Of these 348 articles, about 65% of the articles listed at least one of the statistical methods coded. Most of the articles that listed a statistical method only listed a single method, but a few articles had up to 10 methods listed. The most popular inferential methods found were analysis of variance, t-tests, correlations, and Chi-Square.

To our knowledge, no other study has undertaken the synthesis of software usage, statistical methods usage, or the interaction between the two. This study aims to explore the following research questions.

Research Questions

1. To what extent has the statistical software usage shifted over time in published analyses?
    - If there is evidence of a shift, is there evidence this shift differs based on quantitative method or journal?
2. To what extent are published analyses citing statistical software?
    - Has this changed over time and across journals?
3. To what extent are open-source software tools used?
    - Is there evidence of reproducible analyses being employed?

Methods

Research synthesis methods (Cooper, 2016) will be used to explore the evolution of statistical software and quantitative methods in social science research. More specifically, the statistical software used for the analysis will be coded in additional to the specific quantitative methods (i.e. linear regression, hierarchical linear model, etc.). Additional meta data will also be coded including, journal, article title, author information, article keywords, year published, and any mention of supplementary materials. This information will be used to explore descriptive trends in the data over time, by journals, and methods.

The research synthesis will gather data from a handful of education journals that primarily publish empirical data analysis. The search will not include journals that the primary focus is methodological, the use of software in these journals would likely be a different population than those that are data analytic in nature. Therefore the following journals were selected to be searched from 1995 onward:

- American Economic Journal (AEJ)
- American Educational Research Journal (AERJ)
- American Journal of Political Science (AJPS)

- Economic Journal (EJ)
- Educational Evaluation and Policy Analysis (EEPA)
- Educational Researcher (ER)
- Higher Education (HE)
- Journal of Experimental Education (JEE)
- Journal of Public Policy (JPP)
- Political Science Quarterly (PSQ)
- Public Policy Administration (PPA)
- Sociology of Education (SE)

Data and Software

All journal articles published between 1995 through the middle of 2018 were organized into EndNote. The citation information was obtained using the Web of Knowledge online tool. Once the citations were added to EndNote, the find pdf feature was used to gather the published documents from each journal. This pdf database will then be searched using the *pdfsearch* R package (LeBeau, 2018; R Core Team, 2017). This package allows for keyword searching directly within pdf documents. This will be the primary data collection method. The software keywords searched for can be seen in Table 1. Table 1 also shows keywords to be used to search for statistical models and estimation methods. A handful of articles will be randomly selected to be coded manually by reading the document to evaluate the accuracy of coding using the *pdfsearch* package.

Additional metadata obtained from journal articles will be obtained and combined with the keyword searching data. This metadata will contain information such as, year of publication, publication keywords, author information, and other article metadata obtained from EndNote. The citation information was converted to a bibtex file and the *bib2df* package was used to parse the bibtex fields to collect the metadata (Ottolinger, 2018). These data will be used to further enhance the keyword search data obtained from the *pdfsearch*

package and the subsequent analyses discussed in the next section.

Analysis

Descriptive analyses will be performed on the research synthesis data obtained from keyword searching performed with the *pdfsearch* package. Initial exploration will focus on software and statistical models separately. Subsequent descriptive analyses will be performed to explore the if there are any interactions between software and statistical models used in published research. All of these analyses will be performed over time to explore trends in software and statistical model usage and citation rates. Figures will be the primary analysis methods and will be created in R using the ggplot2 package (Wickham, 2016).

Results

The number and percentage of PDFs able to be obtained from EndNote is shown in Table 2. Many of the journals had a high success rate of obtaining the PDF for the published studies, however some were problematic. For example, JPP and EEPA were both under 50% of PDFs obtained and AJPS, EJ, and SE were all just over 50%. The PDFs that were not obtained were attempted to be obtained over multiple occasions over a six month period, with no additional PDFs obtained between the last two attempts. Figure 1 shows the percentage of PDFs obtained by year and journal and highlights some noticeable trends. There are periods of time for specific journals that all the PDFs are obtained compared to not obtained. For example, AJPS, EEPA, EJ, and SE have periods from 1995 to just after 2000 where most of the PDFs were not obtained. These ranges are portions of time in which our University does not have digital access to these journals.

Using the obtained PDFs, keyword searching was performed for the software and model keywords. Figure 2 shows the percentage of articles from each journal with at least one match for software and model keywords. The figure shows that software keywords are much less likely to be found within the obtained PDFs compared to the statistical model keywords.

The largest percentage was in JEE with about 50% of obtained articles reporting at least one of the searched software keywords. Most of the other journals had 25% or less of the articles mentioning one of the software keywords, with none of the obtained articles in JPP mentioning a software keyword. Model keywords on the other hand were much more prevalent and the journals fall into two broad groups. One group, JEE, EEPA, AEJ, AJPS, SE, and EJ have more than 50% of the obtained articles mentioning at least one of the model keywords and the remaining journals being less than 50%. The articles in the latter group, with the exception of AERJ, were closer to 25% or less.

Expanding on the software keywords found within the PDFs, Figure 3 explores which software keywords were found in each journal. In general, mirroring results from Figure 2, the percentage of articles reporting software keywords was relatively small, most often less than 5%. R, SAS, SPSS, and STATA were the most commonly found software keywords, with R being the most common in most journals. The one exception to this was in JEE, where SAS was more common (about 20% of articles), with R and SPSS having a similar percentage (about 15%) of articles reporting their usage. One interpretation note, articles may mention more than one software keyword and those duplicate results will show up in each category. On average, the average number of software keywords identified in each article was highest for JEE at 1.71 (range: 1 to 5) and a lowest of 1 for PPA (range: 1 to 1).

A similar figure for model keywords can be seen in Figure 4. The x-axis scale here is wider compared to the software keywords showing that the methods are more commonly reported. However, there are still a sizable number of articles appearing in these journals that do not list one of the model keywords searched. The most commonly used methods are linear model, analysis of variance, meta-analysis, or linear mixed (i.e. HLM or multilevel) models. The journals EEPA, JEE, and SE have the widest array of models identified through the keyword search. On the opposite side, AEJ, AJPS, and EJ are dominated by linear models. Finally, journals such as AERJ, ER, HE, PPA, and PSQ all have a low

prevalence of articles that are using the model keywords.

Impact of publication year on keyword rates

The impact of publication year on prevalence of software and model keywords was explored in Figures 5 and 6 for common software and model keywords respectively. In general, publication year does not have a strong impact on the software keyword rates with most trajectories being flat across the publication years. The one exception to this is with SAS which on average has declined across the publication years. R has consistently been the most cited software across these publication years for these journals, but overall software was infrequently cited (see Figure 3).

Commonly used specialty software, such as AMOS, HLM, LISREL, or Mplus, are shown in the bottom-most plot of Figure 5. Not surprisingly, the specialty software on average shows up less frequently compared to the general purpose software shown in the top-most plot of Figure 5. Out of these four specialty software, AMOS tends to be used the most frequently, LISREL has decreased in usage since 1995, and Mplus has gained additional usage. However, these programs still only account for around 10% or less of statistical software used.

The usage of model keywords occurs more frequently and shows evidence of trends across publication years as shown in Figure 6. For example, the top figure shows a large increase in the prevalence of terms related to linear models (i.e. regression models) or linear mixed models and shows a decline in analysis of variance methods. Logistic regression and the variety of t-tests are rarely mentioned in the articles included. The bottom figure also shows general increases in the four models depicted, particularly the mention of IRT and SEM methods. These gains are more modest compared to the increase in the mention of linear models from the top figure, but the increase has now put IRT and SEM methods about the same percentage as ANOVA from the top figure.

Interaction between software and statistical methods

Figure 7 shows a tile plot of the interaction between software (x-axis) and model (y-axis) keywords. The darker shaded regions of the figure show more combinations of the software and model keywords. In the figure, publication year is ignored to identify which methods are most closely paired with specific software. There are duplicates in the data for this figure, for example, an article may cite both R and SAS within the document and mentioned using a linear model. In this case, this article will show up in the R/linear model cell as well as the SAS/linear model cell. Any duplicate keyword combinations would only occur once for each article. This analysis is also limited to articles that have both a software and model keyword returned. On average, only about 20% of all the PDFs obtained from the study had both a software and model keyword in them. This provides further evidence of the reporting bias, particularly with regard to software.

The figure shows that the two most common combinations are R and ANOVA and SAS and linear model. ANOVA is also common with more specialized software such as AMOS, LISREL, or Mplus that may indicate these are being used for nested model comparisons. Linear mixed models were most commonly used in SAS, with R and HLM software being the next highest percentages. Meta-analysis is most commonly associated with SPSS. Interestingly, IRT models are associated with R more than specialized IRT software, however the number of articles using IRT may be quite small and may not adequately cite the software used.

The impact of publication year is explored next for the four general purpose statistical software, R, SAS, SPSS, and STATA. In addition, the models explored were restricted to reflect more general purpose statistical procedures. Each panel in Figure 8 represents a different publication year. In the panel label the percentage of articles that are missing. More specifically, these are studies in which a PDF document was obtained but did not include software and model keyword. These values range from a low of 70% of the articles

not appearing to a high of about 88%. The percentage of missing articles does decrease slightly in recent years, however there is still evidence of reporting bias.

The tile plot shown in Figure 8 shows an increase in the percentage of keyword combinations that are found as the publications become more recent. This is particularly true for R, SAS, and SPSS software regardless of the model used. On the other hand, STATA appears infrequently within each year, but there is evidence of STATA appearing more frequently with more recent publications. There does not appear to be any significant trends over time with regard to which models are used in particular software, but there were articles using R, SAS, and SPSS with most of the statistical models shown in Figure 8 which supports the general usage of the software across a variety of model situations.

Discussion

This study explored the usage of statistical software and statistical models in published social science research from 1995 to 2018 across 12 academic journals. The journals that were selected were identified as journals within sub-fields of social science that were focused on publishing applied research spanning economics, education, political science, psychology, public policy, and sociology. After obtaining PDF documents these journals, keyword searching was performed to extract keywords associated with statistical software and statistical models (see Table 1 for a list of search words that were used).

The results show that statistical software is not cited as frequently as the statistical models that are used to analyze the data. This has important implications for the reproducibility and replicability of the research that is being done. There has been an increase in the prevalence of software citations in published research, however overall the percentage of studies that cite software is about 20%, low rates that have been found by others (Tatman, 2015) Evidence support differential software citations rates by academic journal and there was also evidence of specific statistical methods being used within academic

journals. For example, JEE had evidence of stronger citation rates of statistical software and a wider variety of statistical models. However, linear models dominated journals such as EJ and AEJ, both economic journals, these journals also had very low software citation rates.

The statistical software, R, had evidence of being highly cited compared to other statistical software over time. This may be due to R users being more motivated to cite the software or due to R providing a citation function that makes it easy to generate R and R package citations (i.e. the `citation()` function can be used for this). This is anecdotal evidence that would warrant addition exploration.

Recommendations for Practice

This study provides further evidence that statistical software is commonly not cited in published research. Without appropriate statistical software citations, it can be much more difficult for others to replicate or reproduce the work of other studies. In addition, when differences are found, it would be unclear if the differences are due to the unreplicability of the findings or due to differences in statistical software specifications. For example, default settings across general purpose or specialized statistical software could explain differences in study results.

We want to promote the citation of statistical software along with the statistical methods being used. Statistical software takes time to implement, test, evaluate, etc., therefore users of this software should at least cite the software used to acknowledge the time developers have spent. If more statistical software is cited, the transparency and reproducibility of studies will be increased as well another benefit of the citing software used. However, the citation of statistical software is not enough, the specific version used and operating system used is also needed as implementations evolve, software bugs are fixed, and default settings can change over time. Software citations are a necessary but not a sufficient condition to ensure proper transparency and reproducibility of study results.

References

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., ... Arslan, R. (2017). *Rmarkdown: Dynamic documents for r*. Retrieved from https://CRAN.R-project.org/package=rmarkdown

American Psychological Association. (2010). *Publication manual of the american psychological association*. Washington, D.C.: American Psychological Association.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... others. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119.

Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach* (Vol. 2). Sage publications.

Ioannidis, J. P. (2014). How to make more published research true. *PLoS Medicine*, *11*(10), e1001747.

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., & Ioannidis, J. P. (2016). Reproducible research practices and transparency across the biomedical literature. *PLoS Biology*, *14*(1), e1002333.

LeBeau, B. (2018). pdfsearch: Search tools for pdf files. *Journal of Open Source Software*, *3*(27), 668. doi:10.21105/joss.00668

Muenchen, R. A. (2017). The popularity of data science software. Retrieved from http://r4stats.com/articles/popularity/

Ottolinger, P. (2018). *Bib2df: Parse a bibtex file to a data.frame*. Retrieved from https://CRAN.R-project.org/package=bib2df

Peng, R. D. (2009). Reproducible research and biostatistics. *Biostatistics*, *10*(3), 405–408. doi:10.1093/biostatistics/kxp014

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science & Engineering*, *14*(4), 13–17.

Tatman, R. (2015). The state of the stats: Current use of statistical methods across linguistics subfields. *Linguistics Summer Institute*.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from http://yihui.name/knitr/

Xie, Y. (2017). *Knitr: A general-purpose package for dynamic report generation in r*. Retrieved from http://yihui.name/knitr/

Table 1

Search keywords used in search of published journal documents.

Search	Group	Keywords
Software	SPSS	SPSS Statistics, SPSS Modeler, SPSS
	R	R-project, R Project, CRAN, R core team, R software, RStudio
	SAS	SAS Institute, SAS, JMP
	STATA	STATA
	Python	Python
	Other	MATLAB, Statistica , Statsoft, Java, Hadoop, Minitab, Systat, Tableau, Scala, Julia, Azure Machine Learning
	HLM	HLM[0-9], HLM [0-9]
	IRT	BILOG, BILOG-MG, Multilog, PARSCALE, IRT Pro
	Latent Variable	Mplus, LISREL, AMOS
Statistical Models	ANOVA	Analysis of Variance, ANOVA, ANCOVA, Analysis of Covariance, multivariate analysis of variance, MANOVA, repeated measures analysis of variance, RMANOVA, RM-ANOVA
	HLM	HLM, Hierarchical Linear Model, Linear Mixed Model, LMM, Multilevel Model, Multi-level Model
	Latent Variable	item response theory, IRT, confirmatory factor analysis, CFA, exploratory factor analysis, EFA, latent variable modeling, structural equation modeling, SEM
	t-test	one sample t-test, one-sample t-test, two sample t-test, two-sample t-test, dependent samples t-test, dependent-sample t-test

Table 2

EndNote success rate of obtaining article PDf by journal.

Journal	Number of PDFs	Total Possible Articles	Percent PDFs Obtained
AEJ	363	364	99.7
AERJ	444	444	100.0
AJPS	922	1436	64.2
EEPA	188	405	46.4
EJ	1829	3376	54.2
ER	742	794	93.5
HE	1914	1914	100.0
JEE	517	525	98.5
PSQ	2722	3589	75.8
PPA	83	83	100.0
JPP	27	180	15.0
SE	261	453	57.6

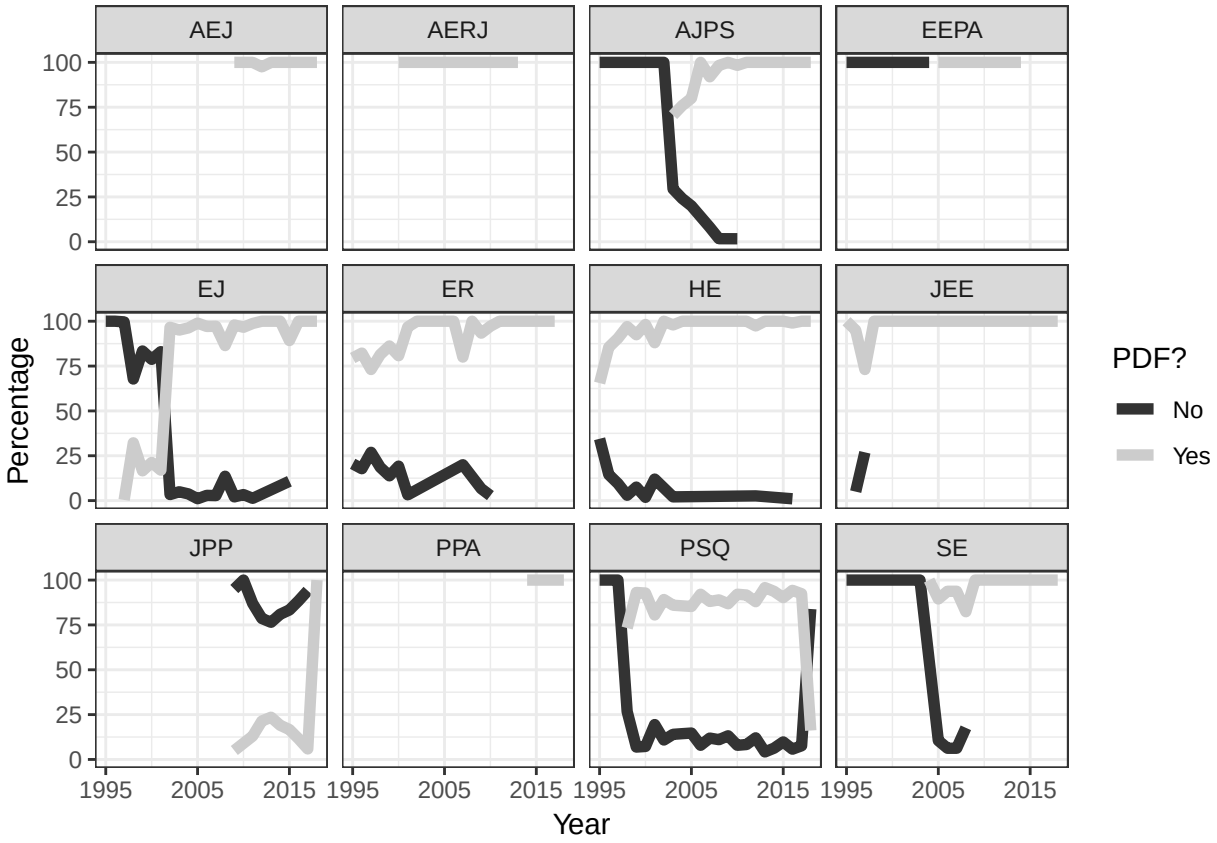

Figure 1. Number of PDFs obtained by journal and year.

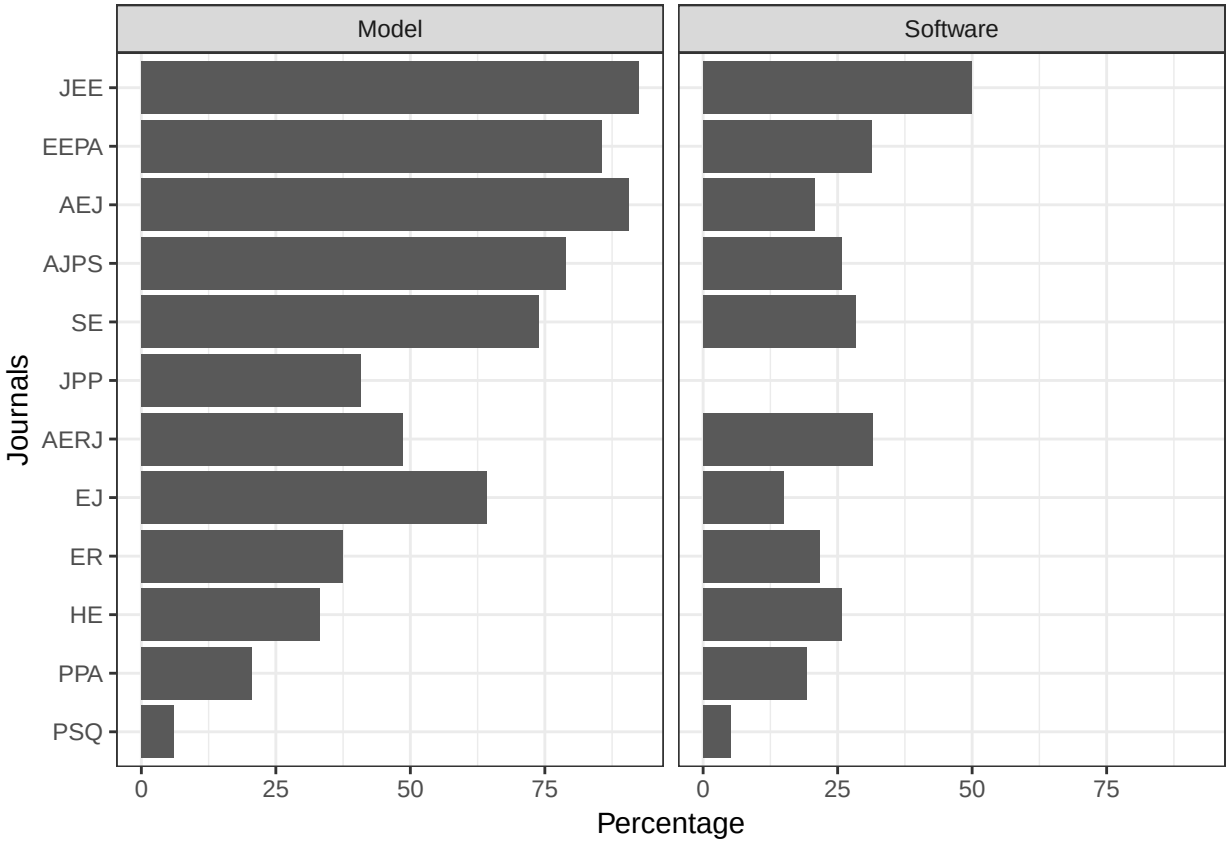

Figure 2. Number of articles with at least one model or software keyword match by journals.

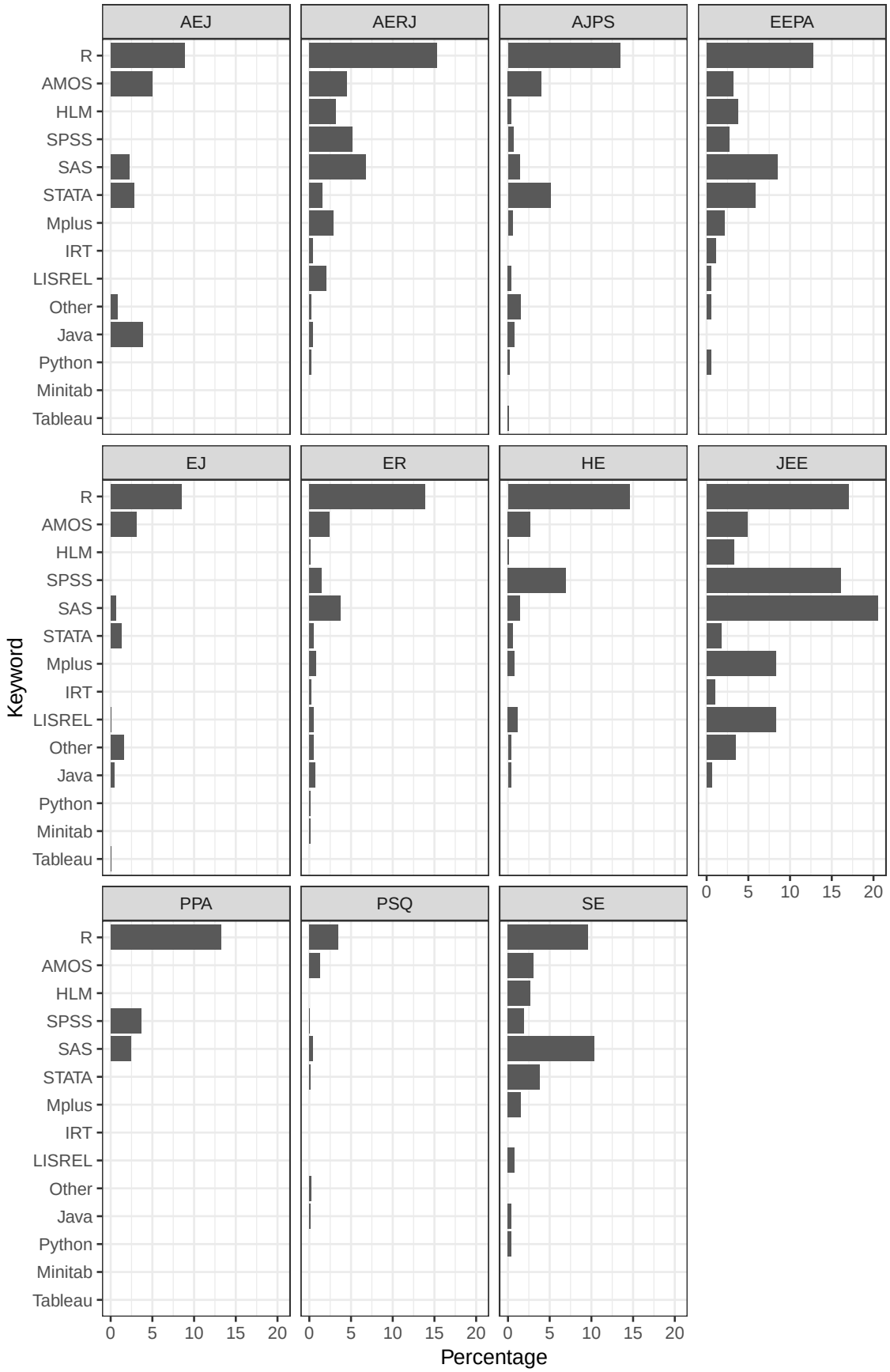

Figure 3. Software counts by journal

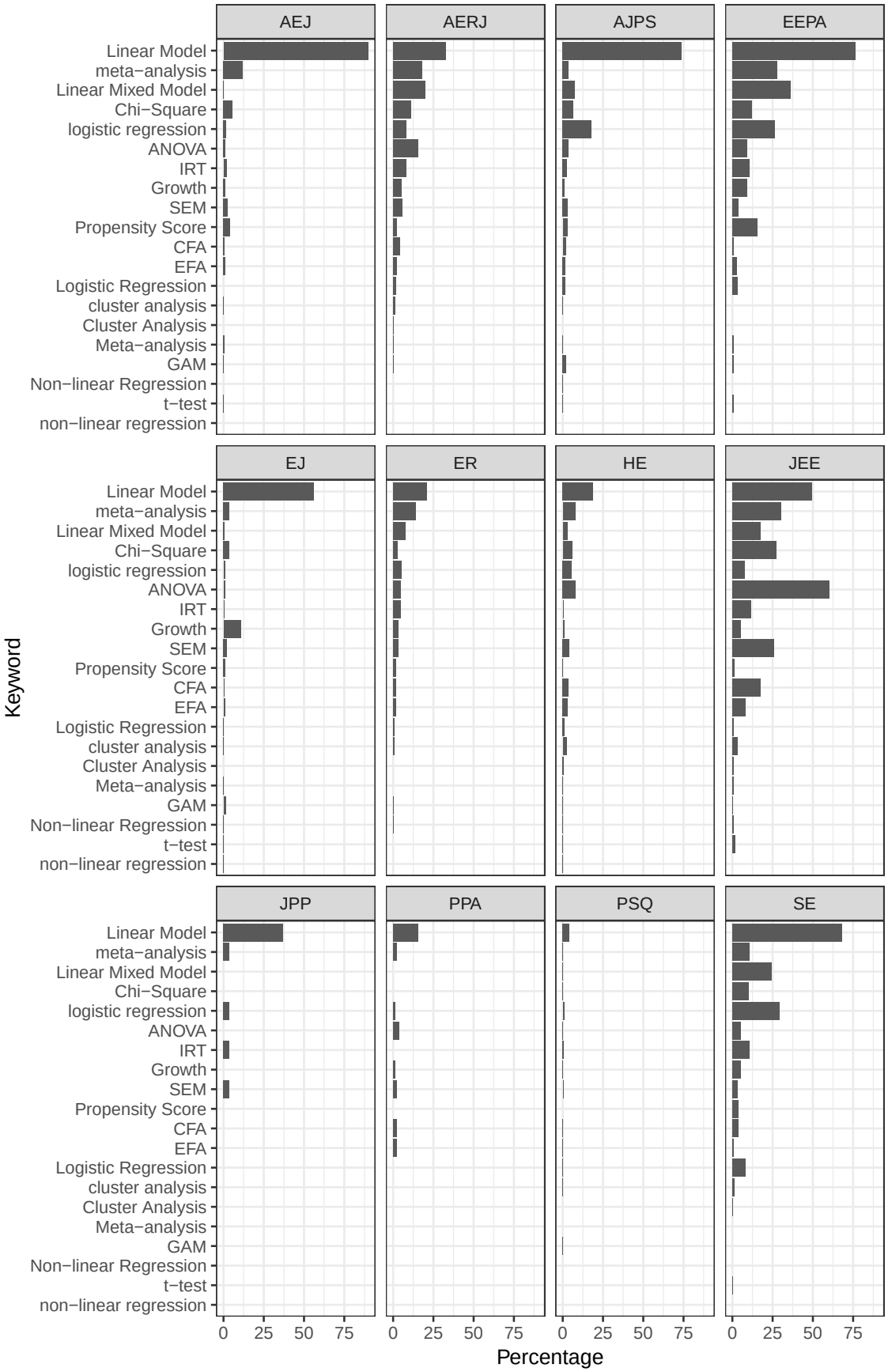

Figure 4. Statistical Model counts by journal

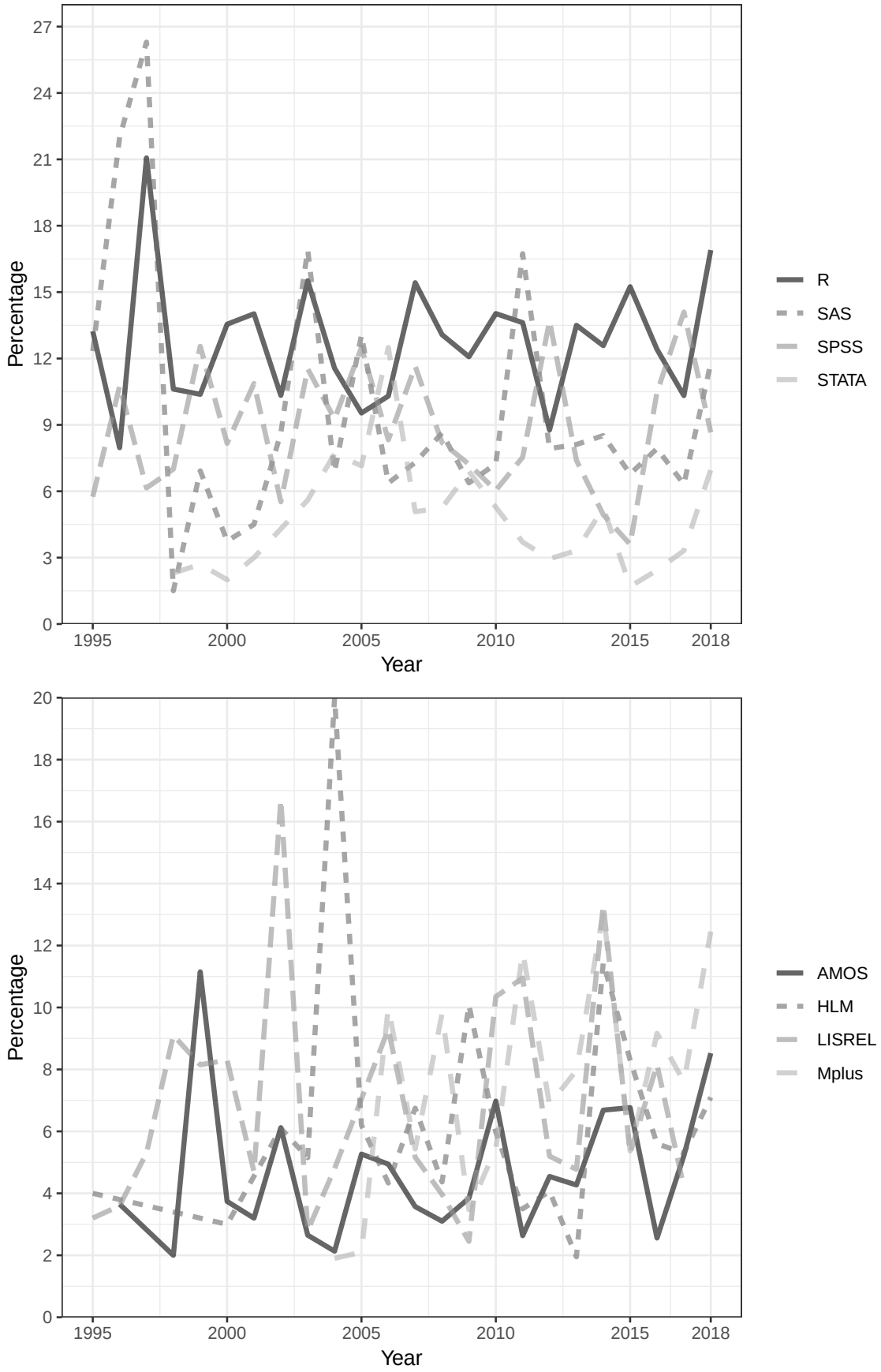

Figure 5. Software percentages by year

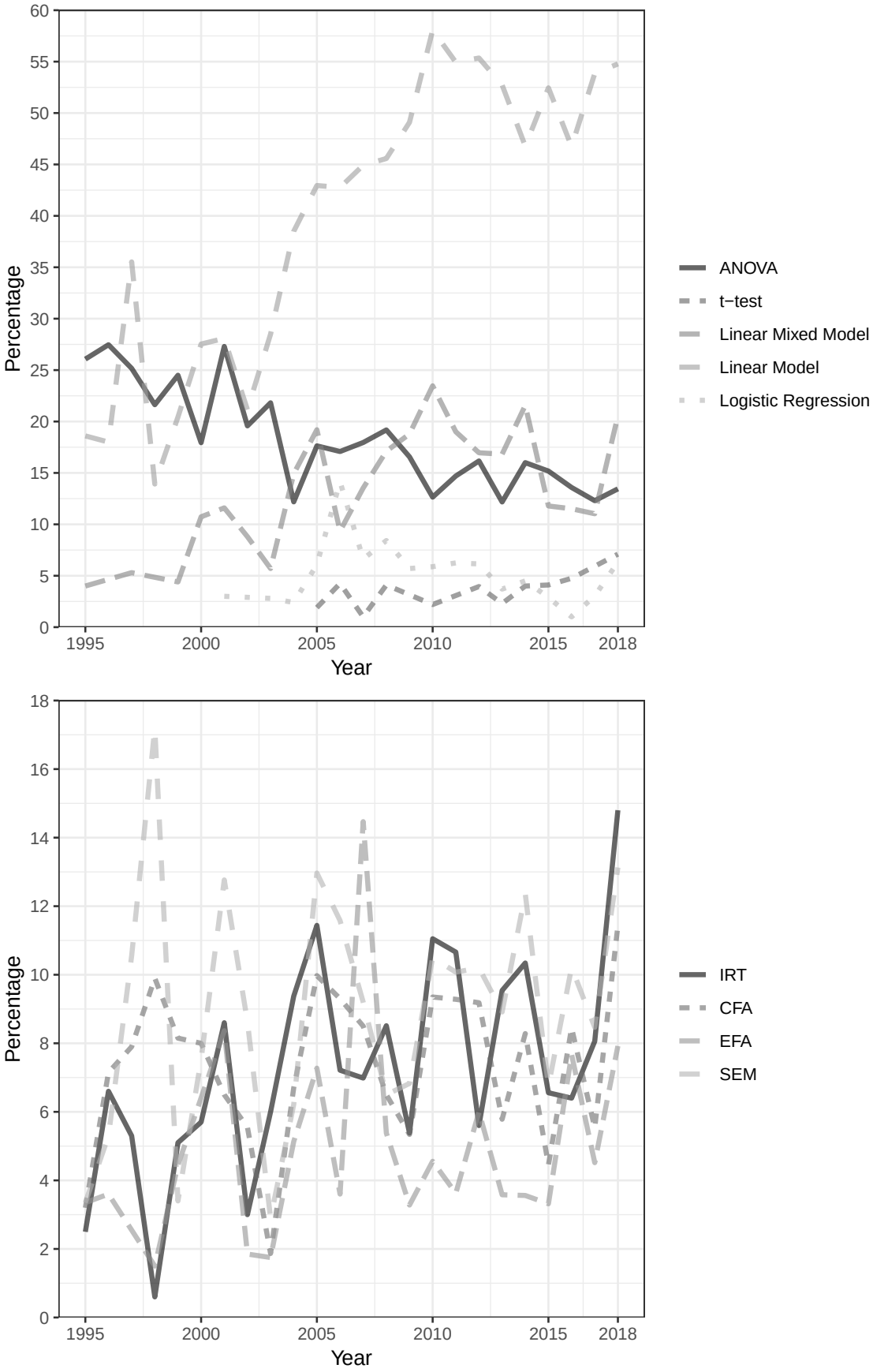

Figure 6. Model percentages by year

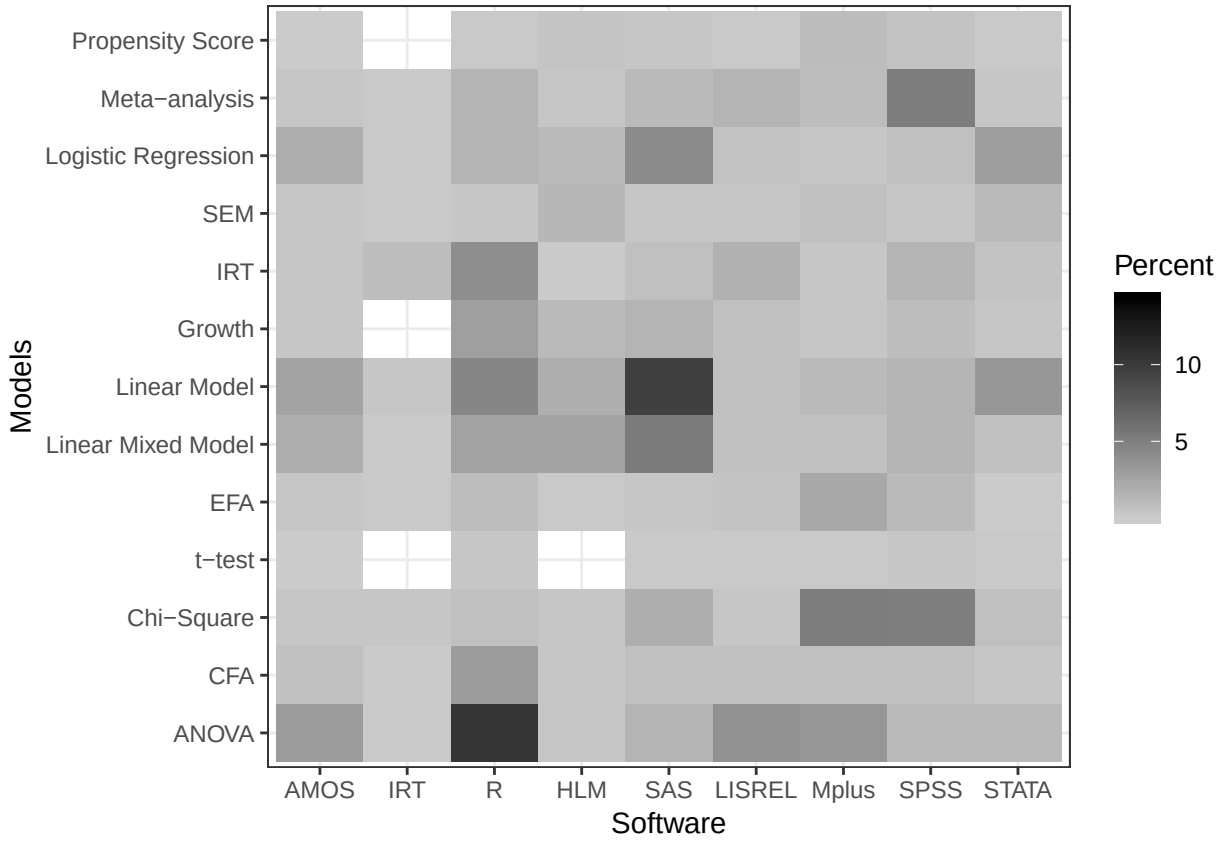

Figure 7. Tile plot showing interaction between software and statistical methods.

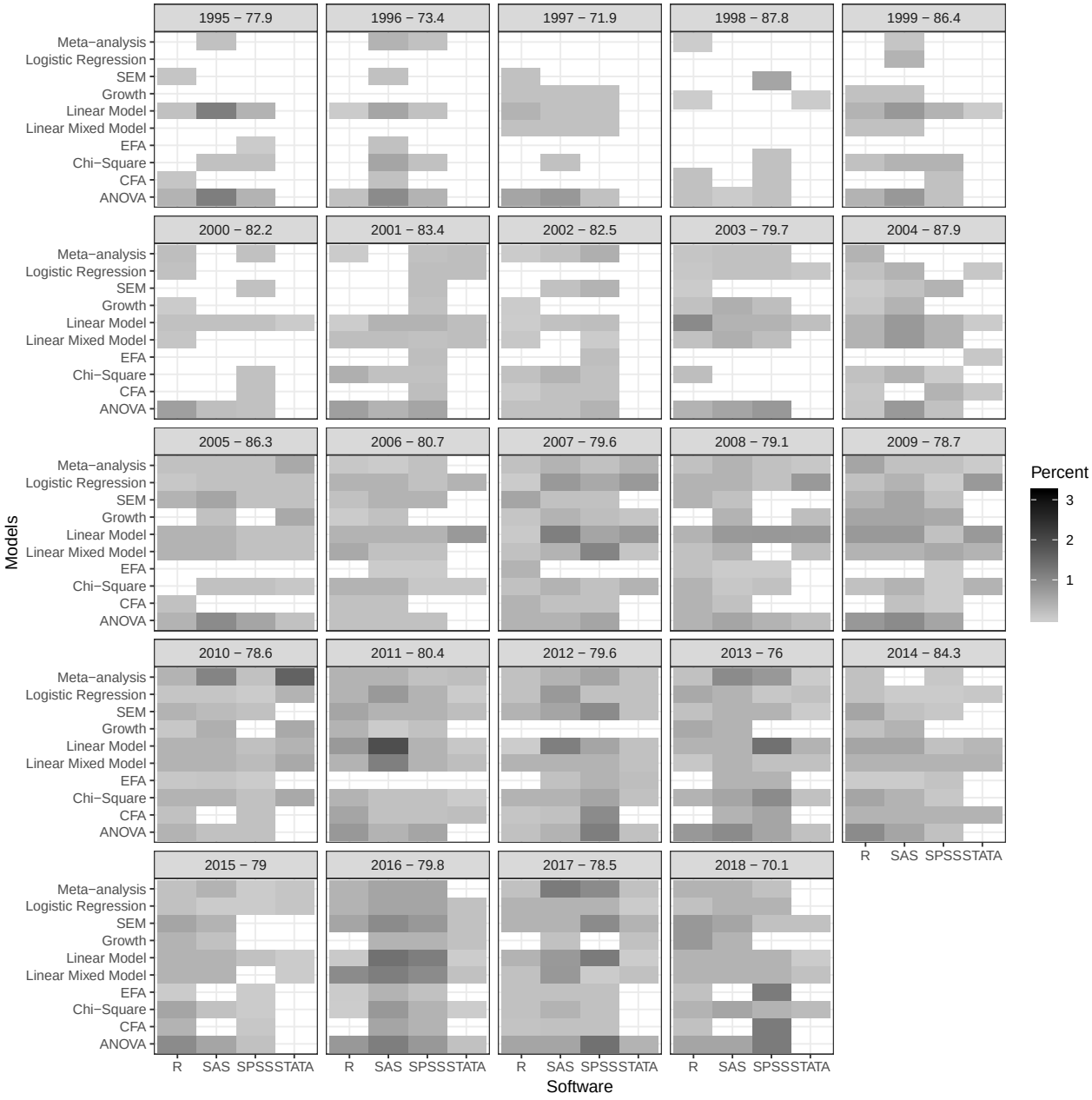

Figure 8. Tile plot showing interaction between software and statistical methods by publication year for primary software.