3 February 2013

The Popularity of Data Analysis Software

by Robert A. Muenchen

Abstract: This page presents various ways of measuring the popularity or market share of BMDP, JMP, Minitab, R, R-PLUS, Revolution R, S-PLUS, SAS, SPSS, Stata, Statistica, and Systat, as well as two implementations of the SAS Lanugage, Carolina and WPS. I update this paper several times a year at http://r4stats.com to provide an ongoing view of the software. Recent updates include: Added 2012 data to listserv plot, Fig. 1a and added a new plot of forum discussion as Fig. 1b (2/12/2013), TIOBE and Transparent Indices (2/3/2013), KDnuggets poll result, number of blogs (5/30/2012), current number of R add-on packages (4/13/2012), plots on Google Scholar data in Fig. 7a & 7b (4/12/2012), numbers of blogs for each package (3/13/2012), Listserv subscriber data (3/9/2012), StackOverflow and Crossvalidated data (3/8/2012).

Introduction

When choosing an analytical tool to use, there are many factors to consider. Does it run natively on your computer? Does the software provide all the methods you use? If not, how extensible is it? Does that extensibility use its own language, or an external one (e.g. Python, R, SQL) that is commonly accessible from many packages? Does it fully support the style (programming vs. point-and-click) that you like? Are its visualization options (e.g. static vs. interactive) adequate for your problems? Does it provide output the form you prefer (e.g. cut & paste vs. LaTeX integration)? Does it handle large enough data sets? Do your colleagues use it so you can easily share data and programs? Can you afford it?

It can also be helpful to know the size of the software's market share and whether it is growing or shrinking. Software that is popular and whose usage is growing probably meets the needs of many people well, however that certainly doesn't mean it will meet yours. That said, let's examine various ways to estimate popularity and/or market share.

Sales & Downloads

Sales figures reported by some commercial vendors include products that have little to do with analysis. Not all vendors release sales figures. For open source software such as R (Ihaka and Gentleman 1996) you could count downloads, but one confused person can download many copies, inflating the total. Conversely, many people can use a single download on a server, deflating it.

Download counts for the R-based Bioconductor project are located at http://www.bioconductor.org/packages/stats/. Similar figures for downloads of Stata add-ons (not Stata itself) are available at http://fmwww.bc.edu/fmrc/reports/report.ssc.html. A list of Stata repositories is available at http://stata.com/links/resources2.html. The many sources of downloads both in repositories and individuals' web sites makes counting downloads a very difficult task.

Language Popularity Measures

The TIOBE Community Programming Index ranks the popularity of programming languages, but from a programming language perspective rather than as analytical software (http://www.tiobe.com). It extracts measurements from blogs, entries in Wikipedia, books on Amazon, and search engine results, and combines them into a single index. In January 2012, they ranked R in 24th place and SAS at 31st. However, by February 2013, the two had reversed positions with SAS in 23rd place and R in 26th.

The only other language that focuses on data analysis that is ranked in the top 100 is S. In previous years SPSS ranked in the 50-100 group but by February of 2013 it had dropped out.

The Transparent Language Popularity Index is very similar to the TIOBE Index with except that its ranking software, algorithm and data are published for all to see. In February, 2013, it ranks R in 12th place and SAS in 25th. Those positions have been stable for at least the prior 6 months.

Langpop.com also ranks programming languages (http://langpop.com/) in a variety of interesting ways, but unfortunately their focus excludes statistical software.

Internet Discussion

There are some stable and objective measures regarding analytic software. Schwartz (2009) suggested estimating relative popularity by plotting the amount of email discussion devoted to each. The most widely used packages all have discussion lists, or "listservs" devoted to them. The less popular ones either do not have such discussions or, like the lists for Minitab or S-PLUS, may have only a dozen or so emails per year. Some software packages have multiple discussion lists. For example, there are 21 devoted to using R for various focused areas such as graphics, mapping, ecology, epidemiology, etc. (http://www.r-project.org/mail.html). A broader list, including a version of R-Help in Spanish, lists 49 discussions (https://stat.ethz.ch/mailman/listinfo).

Figure 1a shows the level of activity on only each main discussion listserv in a typical month (i.e. forums, news groups and Google groups are excluded). Each point represents the sum of the 12 monthly counts that occurred in that year. This plot contains data through the end of 2012. If you read this article in previous years, this plot used to display the mean number of emails per

month rather than the sum. Therefore the scale of the *y*-axis is different but the relative locations of the points are virtually identical. I made this change to enable better a better comparison to discussion forums (e.g. Fig. 1b).

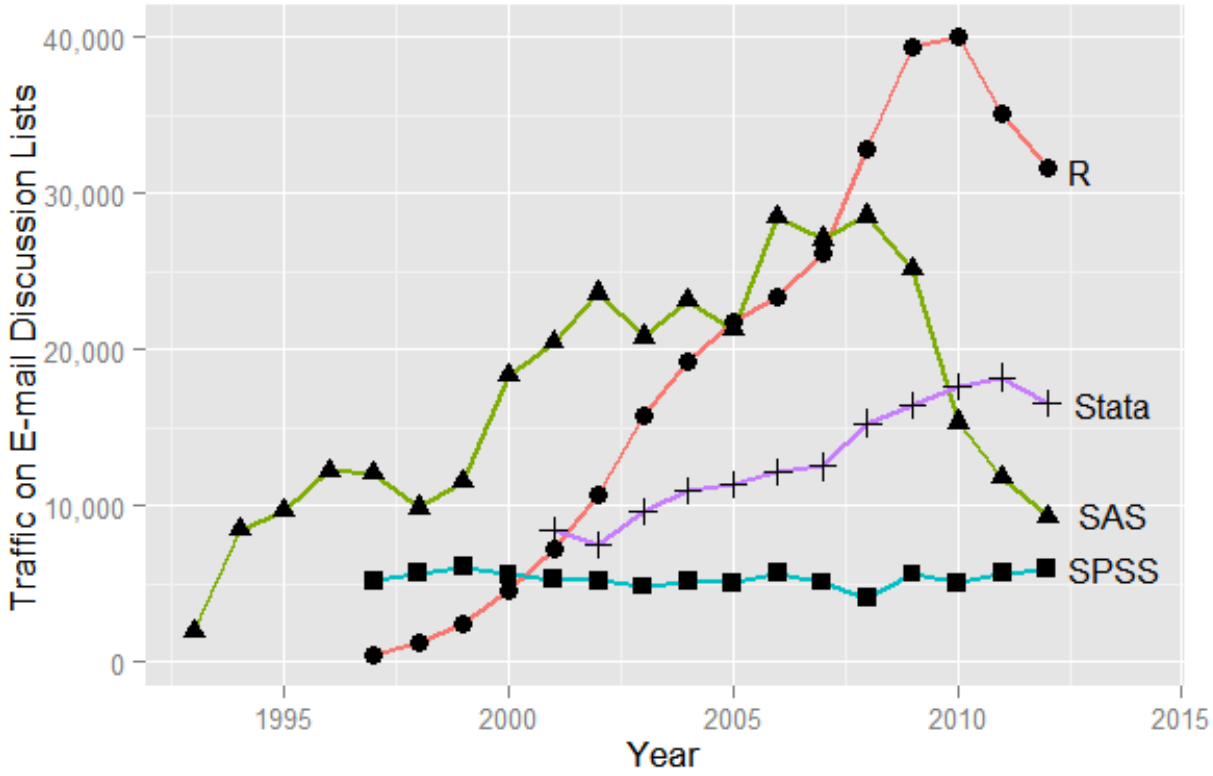


Figure 1a. Sum of monthly email traffic on each software's main listserv discussion list.

We can see that discussion of R has grown the most rapidly and, for the past few years, R is the most discussed software by an almost two-to-one margin. In recent years, it is followed by Stata, SAS and SPSS, respectively.

Stata showed steady discussion growth until it passed SAS in 2010.

SAS saw rapid growth in its discussion until 2006 when it leveled off and then declined. That decline coincided with the strong growth of both R and Stata, offering competition to SAS.

SPSS held steady at a low rate across the time frame, which may be attributable to its great ease of use relative to the other packages. With both the interface and the documentation aimed at people who prefer GUIs over programming, there's less need to ask how to do variations on an analysis. In fact, there's less *ability* to do such variations. As a result, I doubt SPSS' low showing in this graph is indicative of its popularity or market share.

It would be interesting to see what topics were most discussed on each list. The only such analysis of which I am aware was done by Arthur Tabachnek (2010) for the SAS list. The most

popular topic in 2009 turned out to be…R! You can read his full analysis here under *slides from the 2010 session.*

In the last year or two, R and Stata joined SAS in the decline in listserv discussion. Given the sharp increase in the popularity of business analytics, Big Data, and so on, it is unlikely that people are using or talking about these tools less. Instead, alternative forums of discussion have appeared. The site Stack Overflow (http://stackoverflow.com) covers a wide range of programming and statistical topics, while its sister site, Cross Validated (http://stats.stackexchange.com/), focuses only on statistical analysis. A third site, Talk Stats (http://www.talkstats.com), also focuses on statistical analysis. At all three sites, users tag their topics making it particularly easy to focus searches. Figure 1b shows the software people are discussing there.

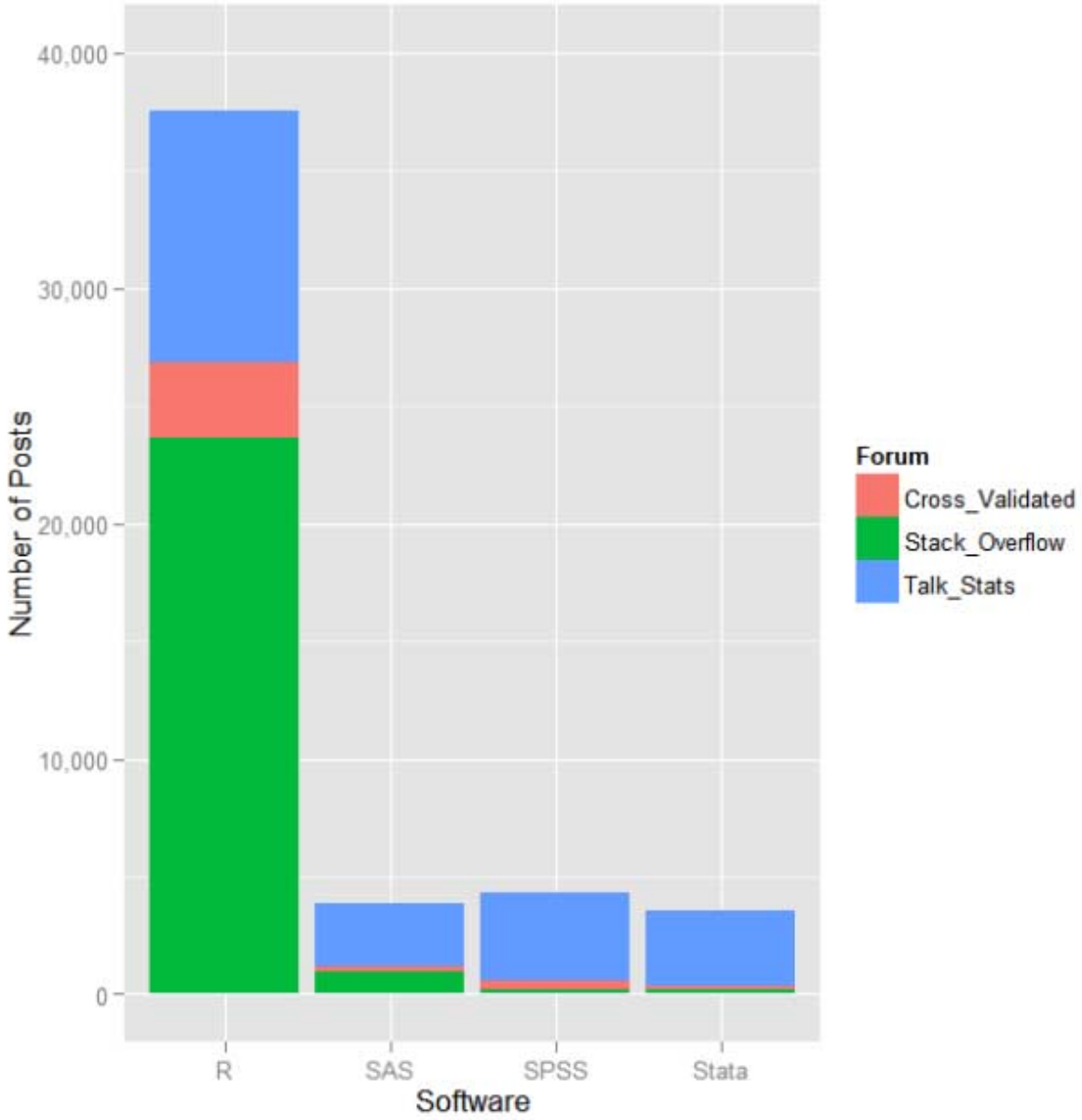


Figure 1b. Number of posts per software on each forum on 2/10/2013.

We can see that the discussion of R is dramatically higher than the other packages, which don't differ very much. Much of this difference is due to the influence of Stack Overflow, reflecting the vastly greater popularity of R as a programming language. However, even removing that effect, it is easy to see that R still dominates the discussions on the more statistically-oriented forums. This data is cumulative, but we can get an historical view of just two of the tags: R and SAS (Figure 1c).

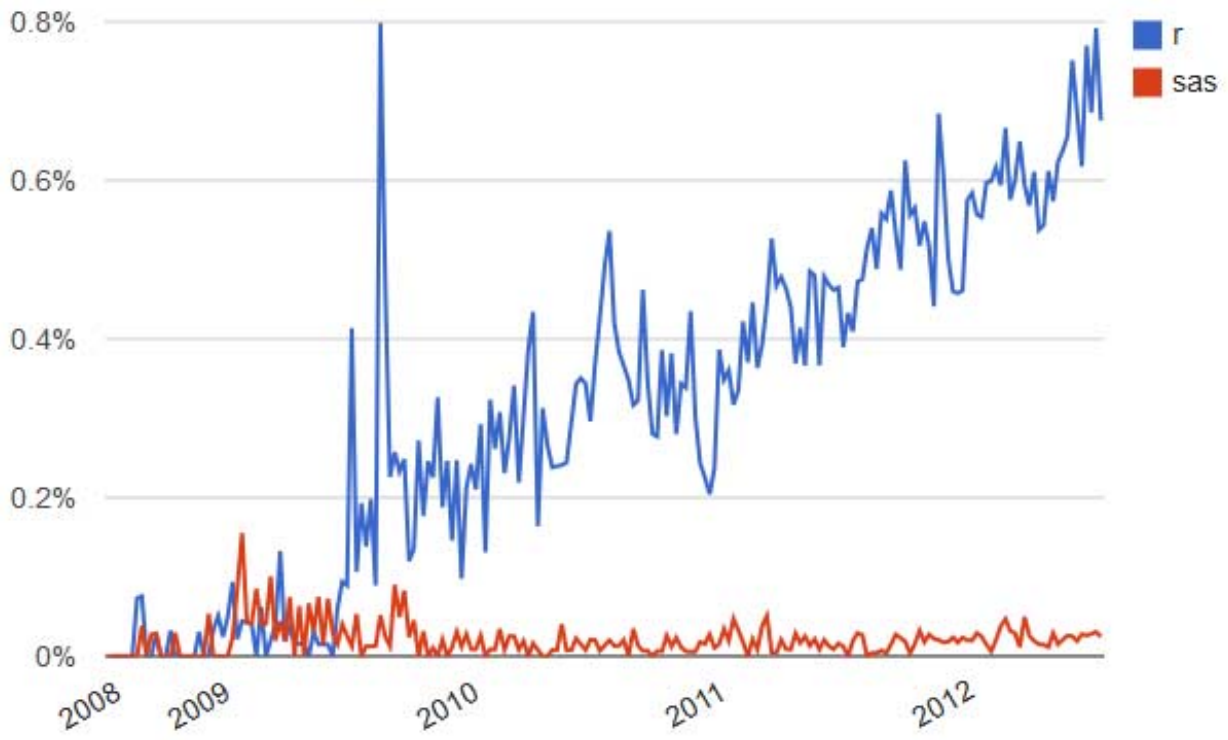


Figure 1c. Number of R- or SAS-related posts to Stack Overflow by week.

We see that discussion of SAS and R were roughly comparable until mid-2009 when the discussion of R began its very rapid climb. The page that provides this data does not display data for SPSS or Stata. The amount of data may be too low; no message provides the reason (see http://hewgill.com/~greg/stackoverflow/stack_overflow/tags).

Other popular discussion forum sites are LinkedIn.com and Quora.com. Neither of these sites make it easy to count number of posts, but they do display the number of people who have joined discussion groups (Figure 1d).

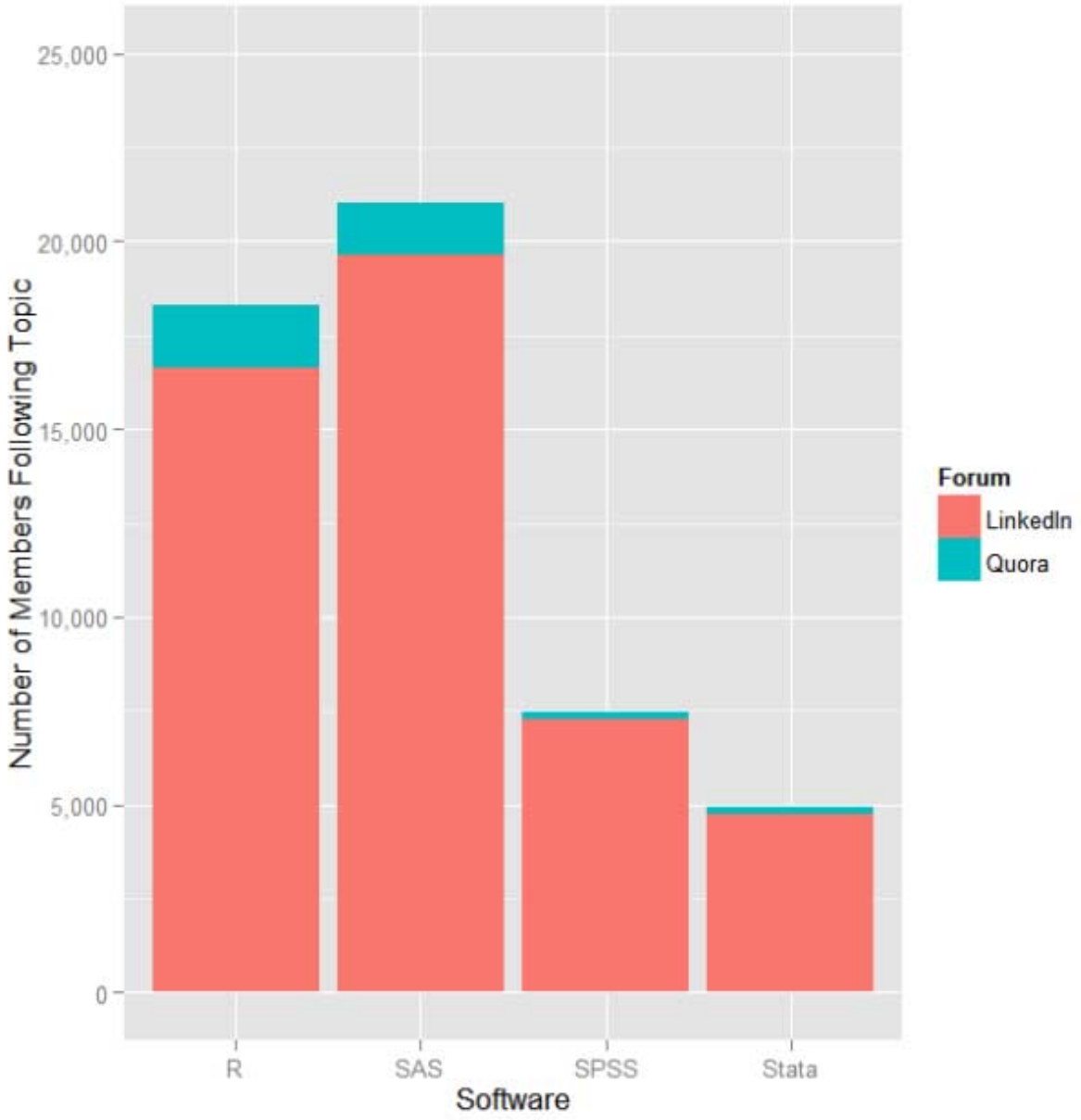


Figure 1d. Number of people registered in the main discussion group for each software.

In Figure 1c we get a better view of corporate software use. I do not know the ratio of corporate to academic use of LinkedIn, but among the academics I do know (quite a few) they use it very little. In this world, SAS is the leader with R close behind. It's interesting to see SPSS with a 50% lead over Stata; it was also slightly higher in Fig. 1b. Remember these are people who have joined a group, not necessary people who are talking as the previous two figures were. Still, group membership should be a reasonable proxy for popularity or market share.

Blogs

On Internet blogs, people write about software that interests them, showing how to solve problems and interpreting events in the field. The more popular a software package is, the more bloggers there are writing about it. Blog consolidators like Tal Galili's R-Bloggers.com and SAS-X.com, and sasCommunity.org Planet combine various blogs into a single location. While any particular blogger may write only an article every week or so, by combining them, the consolidators essentially provide a daily newspaper on various packages. So far only R and SAS are popular enough to have consolidated versions of their blogs (see Table 3).

Software	Number of Blogs
R	365
SAS	40
Stata	8
Others	0-3

Table 3. Number of blogs devoted to each software package on March 13, 2012.

R's 290 blogs put it way out in front of the pack, with SAS coming in at second place with 39. Stata has 7, which are listed here. Each of the other packages have either none or just a few.

Competition Use

Kaggle.com is a web site that sponsors data analysis contests. People post data analysis problems there along the amount of money they are willing pay the person or team who solves their problem the best. As I write this (1/2/2012) there are over 25,000 analysts working on over 72,000 problems. Figure 2 shows the software used by the data analysts working on the problems. R is in the lead by a wide margin. R's dominance is even greater among the contest winners, over 50% of whom used R.A potential source of bias in these figures is that the licenses of most proprietary software prohibits its use for the benefit of outside organizations (universities can help federal grant-providing agencies such as NSF and NIH, but cannot even solve problems for government agencies in general or nonprofits). However, I manage the research software site licenses at the University of Tennessee, and I can attest to the fact that people are often unaware of this limitation.

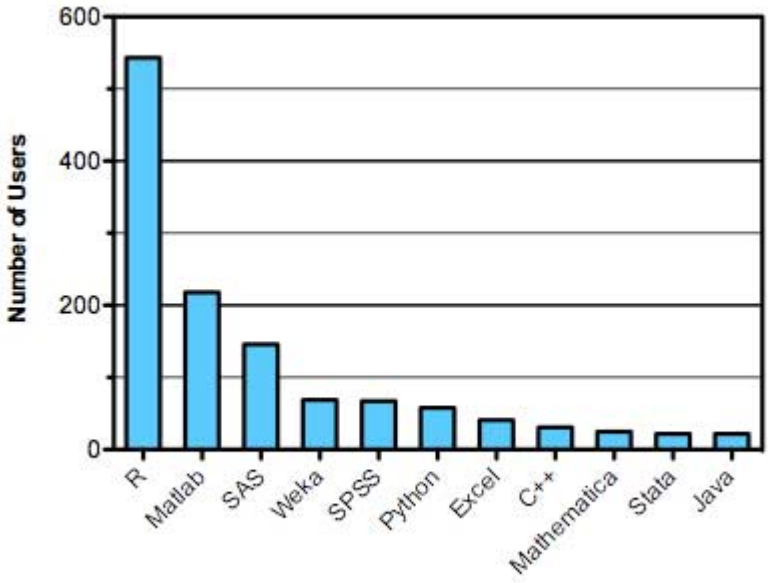


Figure 2. Software used in data analysis competitions in 2011.

Surveys of Use

One way to estimate the relative popularity of data analysis software is though a survey. Rexer Analytics does a survey each year asking about tools used for data mining. The difference between software for classical data analysis software and data mining seems like more of a marketing concept than one based on any actual difference in analytic need. Figure 3 shows the results of just one "check all that apply" type question about the tools that respondents reported using in 2009 (the survey was taken in 2010).

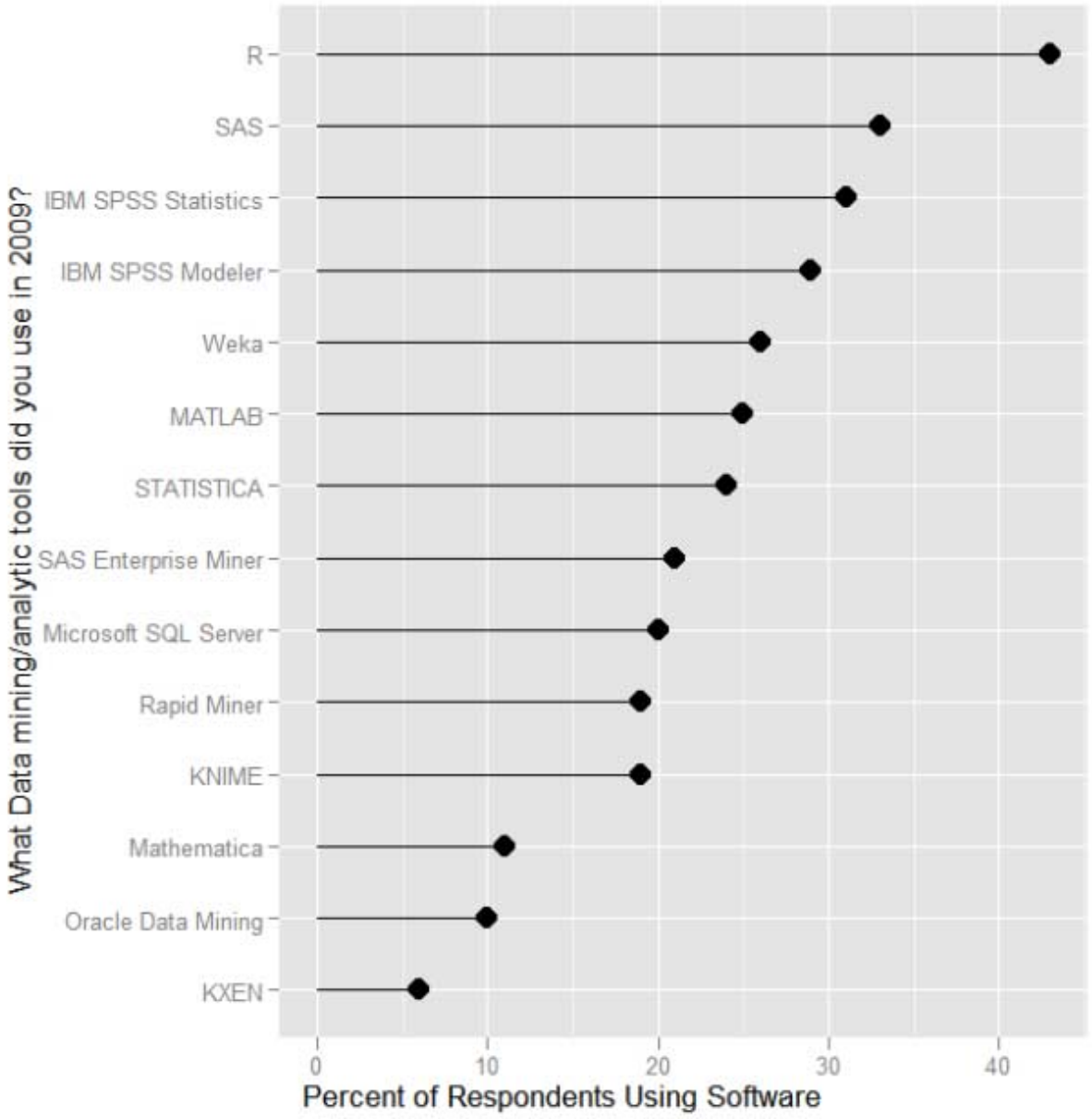


Figure 3. Data mining/analytic tools reported in use on Rexer Analytics survey during 2009.

We see that R comes out on top, followed by SAS and SPSS. The entire report contained over 40 questions on topics such as algorithms used, fields, challenges, data, impact of the economy on the field, and more. More comprehensive results are available here. It's interesting to note that SPSS and SAS are used more often than their more expensive products aimed specifically at data mining, SPSS IBM Modeler (formerly Clementine) and SAS Enterprise Miner. This data is two years old now and due to be updated soon.

The results of a similar survey done by the data mining web site KDnuggets in 2012 are shown in Figure 4. This one shows R in first place with 30.7% of users reporting having used it for a

real project. Excel is almost as popular. It seems out of place among so many more capable packages, but Excel is a tool that almost everyone has and knows how to use.

It's interesting to note that four of the top five packages used were open source. While open source packages are clearly playing a major role in analytics, people still reported using more commercial software (1086) than open source (927).

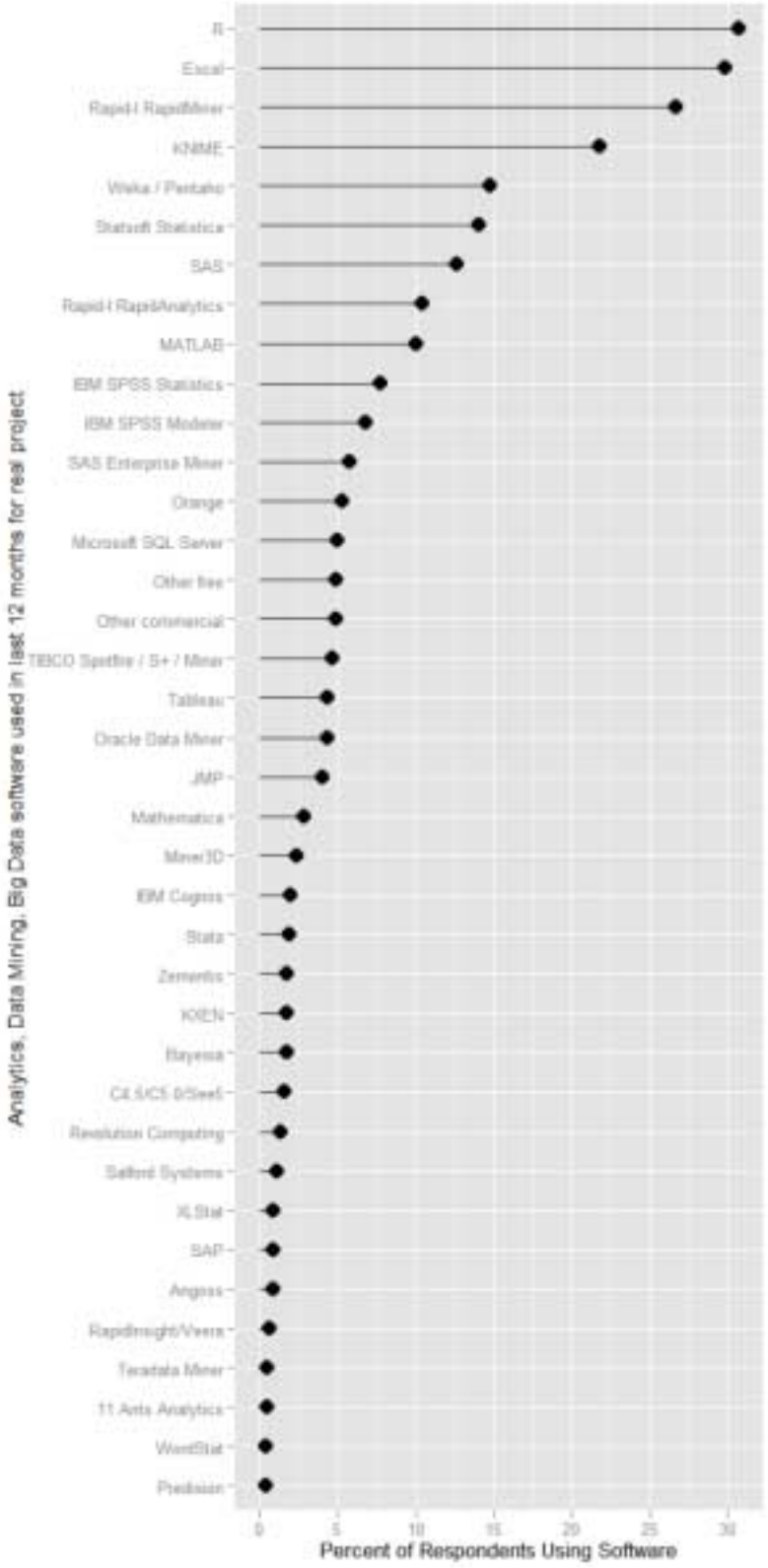
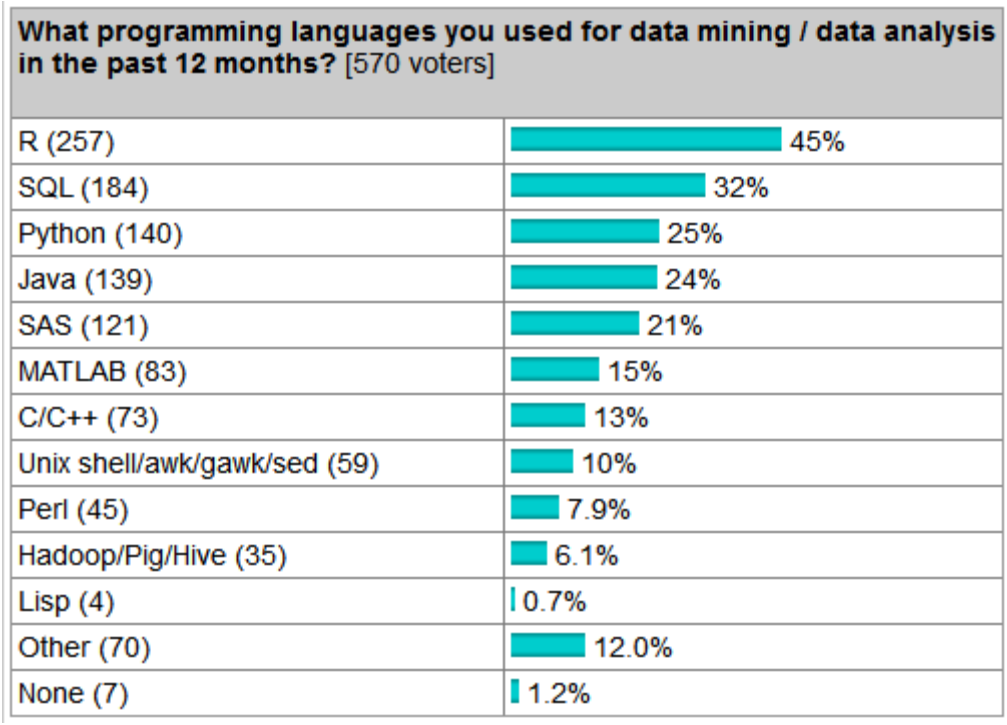


Figure 4. Percent of KDnuggets survey respondents that reported using software for analytics, data mining or big data project for 12 months prior to May 2012.

The KDnuggets site conducted similar poll, this time asking, "What programming languages you used for data mining / data analysis in the past 12 months?" R dominated this poll, as shown in Figure 5.



What programming languages you used for data mining / data analysis in the past 12 months? [570 voters]	
R (257)	45%
SQL (184)	32%
Python (140)	25%
Java (139)	24%
SAS (121)	21%
MATLAB (83)	15%
C/C++ (73)	13%
Unix shell/awk/gawk/sed (59)	10%
Perl (45)	7.9%
Hadoop/Pig/Hive (35)	6.1%
Lisp (4)	0.7%
Other (70)	12.0%
None (7)	1.2%

Figure 5. Languages used in data mining or analysis.

Books

The number of books published on each software reflects their relative popularity. Amazon.com offers an advanced search method which works well for all the software except R. I configured it with the following parameters:

Title: SAS -excerpt -chapter -changes [using SAS as an example]

Subject: Computers & Internet

Condition: New

Format: All formats

Publication Date: After September, 2001 [i.e. 10 years before the search on 10/13/2011]

Since it's difficult to determine how many books use a particular software in its examples, I searched for books that included the software in the *title*. SAS has many manuals for sale as individual chapters or excerpts. Luckily, they contain "chapter" or "excerpt" in their title so I excluded them using the minus sign, e.g. "-excerpt". SAS also has short "changes and

enhancements" booklets that the other packages release only in the form of flyers and/or web pages so I excluded "changes" as well.

SAS and SPSS both have many versions of the same book or manual still for sale. For example, Marija Norusis' 3 books on SPSS appear 20 times for various versions of SPSS released in the last 10 years. The SAS and SPSS numbers are both somewhat inflated as a result. Limiting the search to books published in the last 10 years mitigated this problem somewhat, but the SAS and SPSS figures are probably both still somewhat exaggerated.

The count of R books came from http://www.r-project.org/doc/bib/R-books.html. This list does contain seven books on S that are older but still relevant. Version numbers do not appear in any book titles so R avoids the over-counting problem that plagued my count of SAS and SPSS manuals. The most surprising aspect of the result (Figure 6) was how extremely dominant the top few packages are and that three well known packages had no books at all written about them (BMDP, Statistica, Systat). Revolution R and R-PLUS have no books with their names in the titles, but of course the books on R apply to them as well.

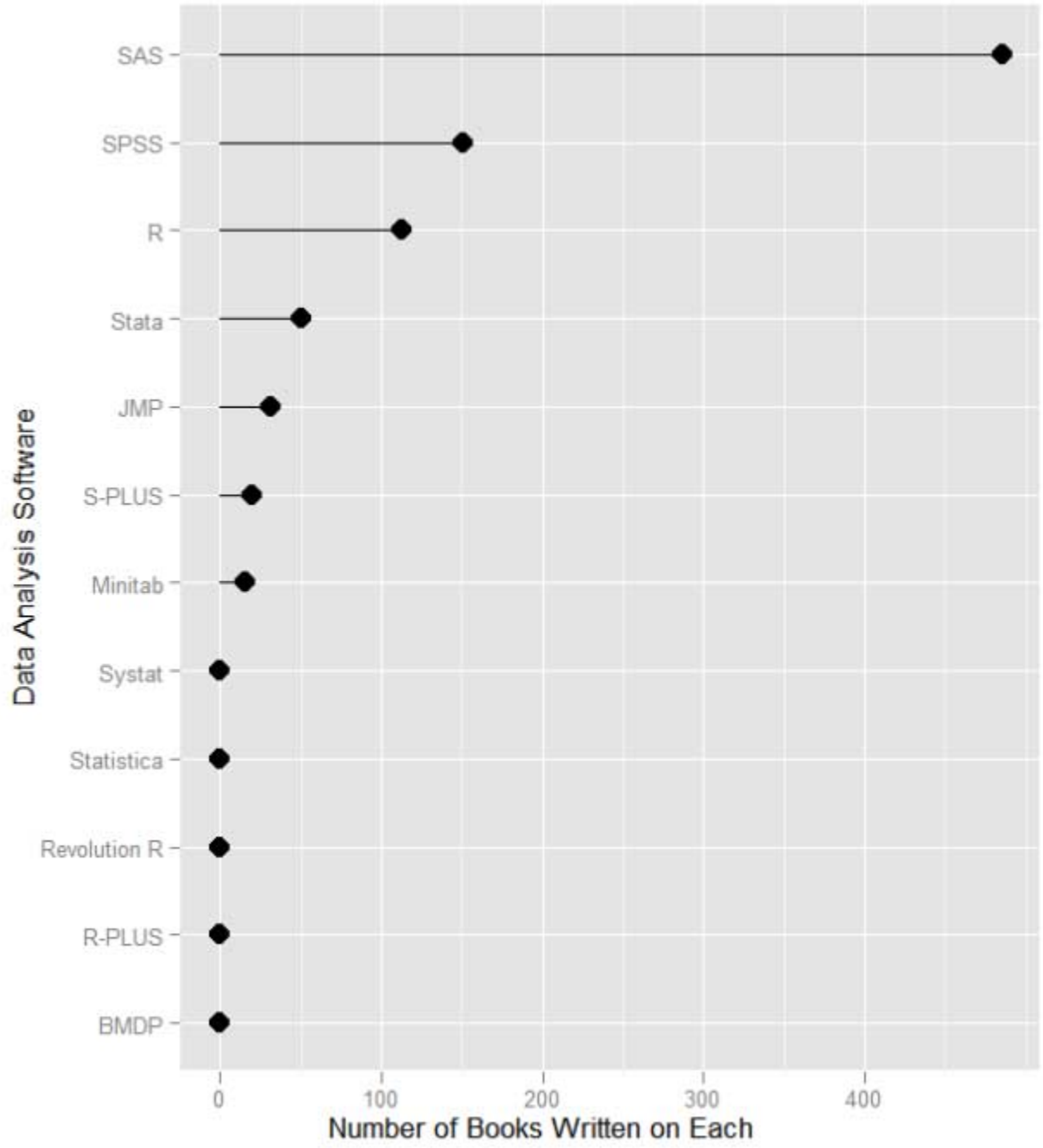


Figure 6. The number of books that contain the name of each software package in their titles.

Impact on Scholarly Activity

While Internet search engines make it very easy to locate information about software, their inclusive nature make it difficult to narrow the search enough to determine the prevalence of various packages. For example, searching for the term “SAS” quickly locates the main web site for the SAS Institute, but it also ends up including many hits regarding a shoe company, an

airline and the British commando group. Even in the realm of scholarly journal articles, S.A.S. stands for over a dozen terms such as Synthetic Aperture Sonar.

The more popular a software package is, the more likely it will appear in scholarly publications as a topic and as a method of analysis. Google Scholar offers a convenient way to measure such activity. No search of this magnitude is perfect and will include some irrelevant articles and reject some relevant ones. The final set of search terms is described at http://librestats.com/2012/04/12/statistical-software-popularity-on-google-scholar/. Figure 7a shows the number of articles for the most popular six statistics packages from 1995 through 2011. SPSS had a surprising advantage over most other package for much of this time. It seems suspiciously large but after fairly extensive study of the result it does not seem to be spurious. Last year's graph did however have a spurious result. Stata apparently means "was" in Italian and so it appeared to follow a similar path to SAS, but exceeding both SAS and SPSS in recent years. Changing that search to "Statacorp", which should be included in the citation for the Stata software yielded what is probably a much more accurate set of data. The Librestats article makes it easy for anyone to try variations on these searches.

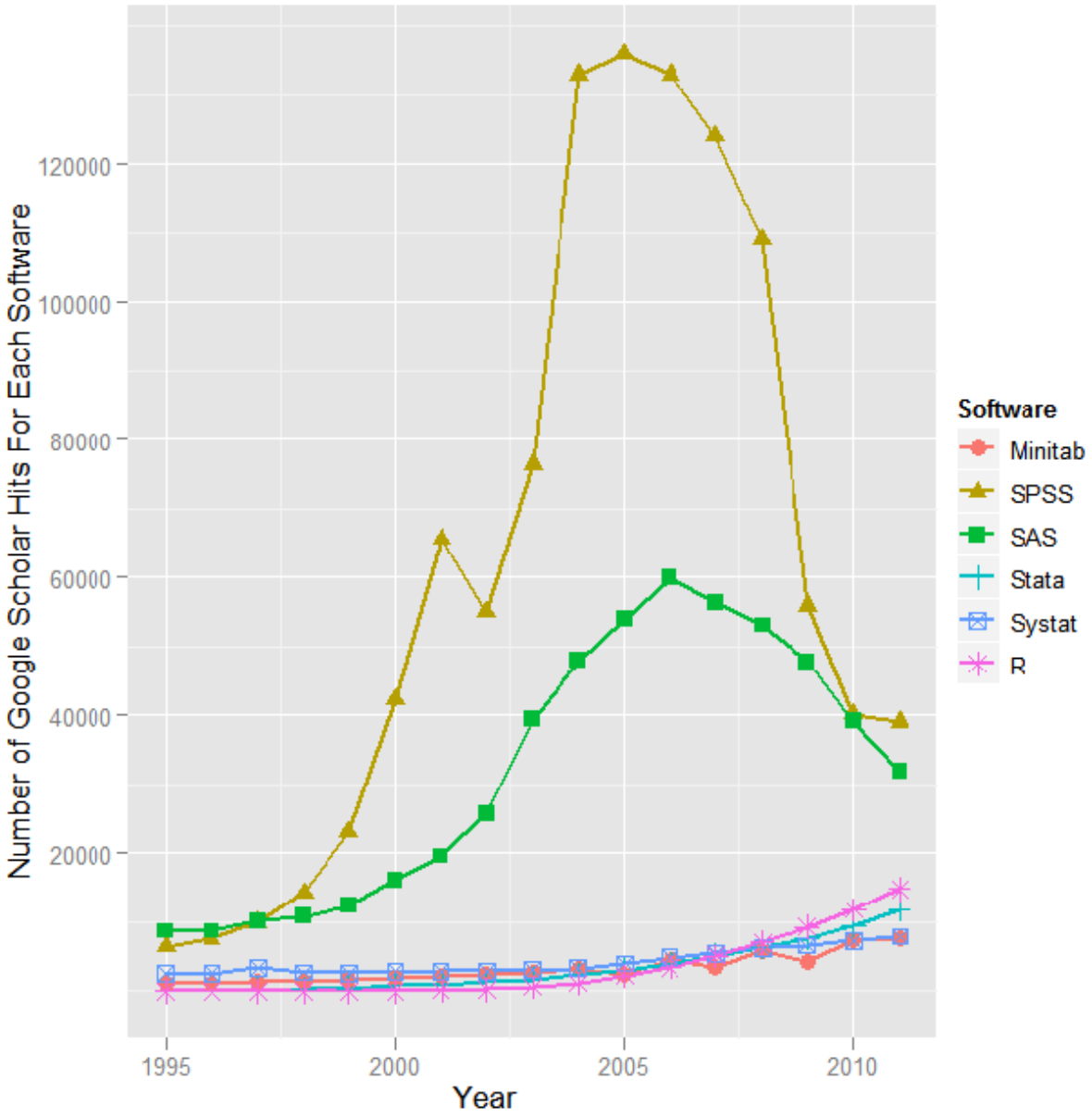


Figure 7a. Use of data analysis software in academic publications as measured by hits on Google Scholar.

Use of SPSS and SAS in scholarly articles peaked in 2005 and 2006, respectively. The decline they have seen since may be due to competition from the other packages. The total of the other packages in 2011 is a similar to the amount of decline that SPSS and SAS have seen since their peak. If the trends for SAS and R were to continue on their current trajectories, scholarly use of R could surpass SAS use in 2015. That's a big "IF" of course! In 2011 the downward trend in SPSS use flattened out quite a bit, making a forecast more difficult. See this blog article for more discussion of this forecast.

Since SAS and SPSS still dominate scholarly use by such a wide margin, I removed those two packages and added JMP and Statistica as shown in Fig. 7b. That figure shows the rapid rise of all software except Statistica. Note that the symbols and colors used in Fig. 7b *do not* match those in 7a. From 2008 on, R reaches the #3 spot (after SPSS and SAS) and extends its lead in consecutive years.

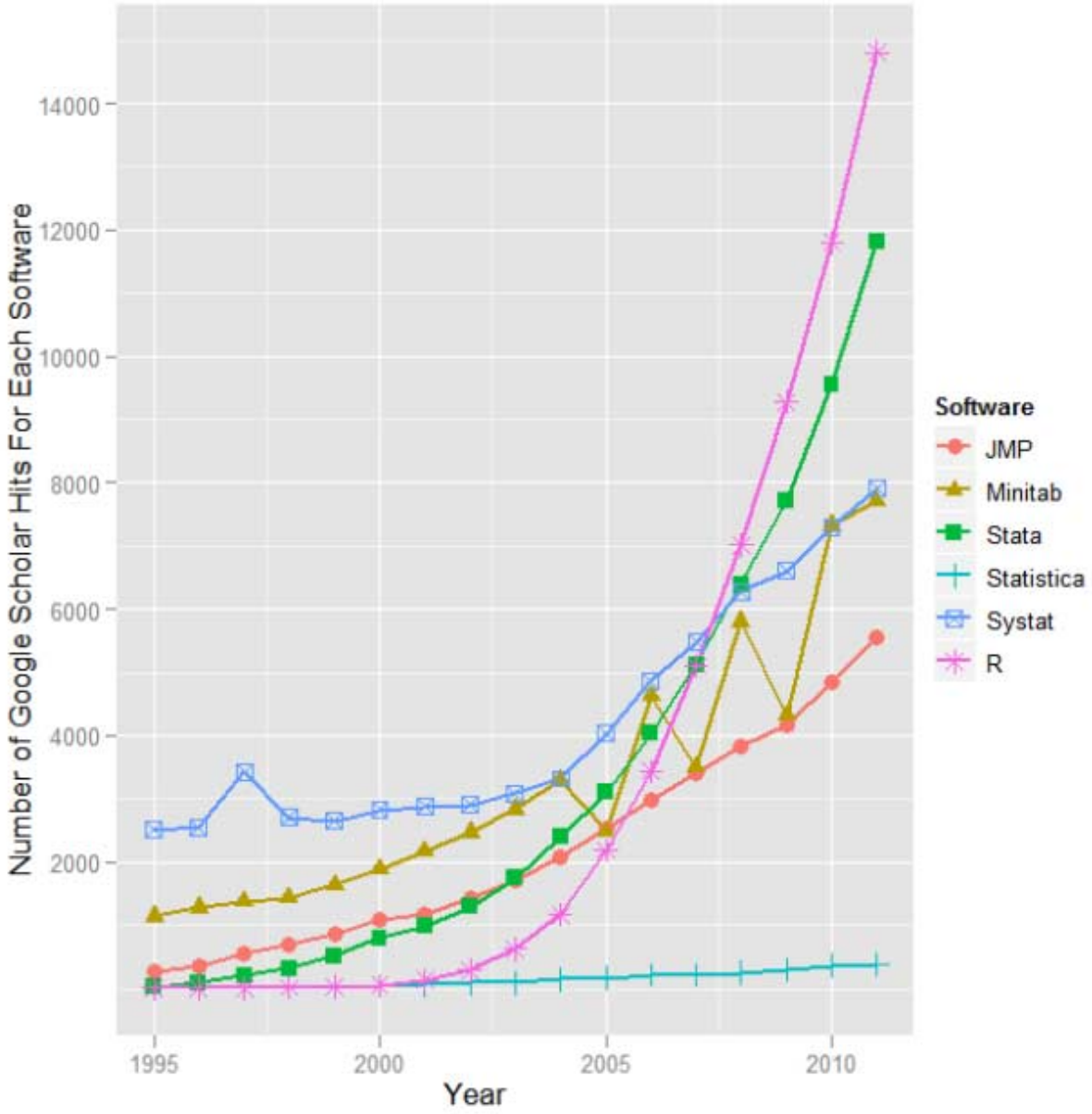


Figure 7b. Use of data analysis software in academic publications as measured by hits on Google Scholar, excluding SAS and SPSS

Web Site Popularity

Another measure of software popularity is the number of other web pages that contain links that point to the software's main web site. Figure 8 provides those numbers, recorded using Google on January 5, 2012.

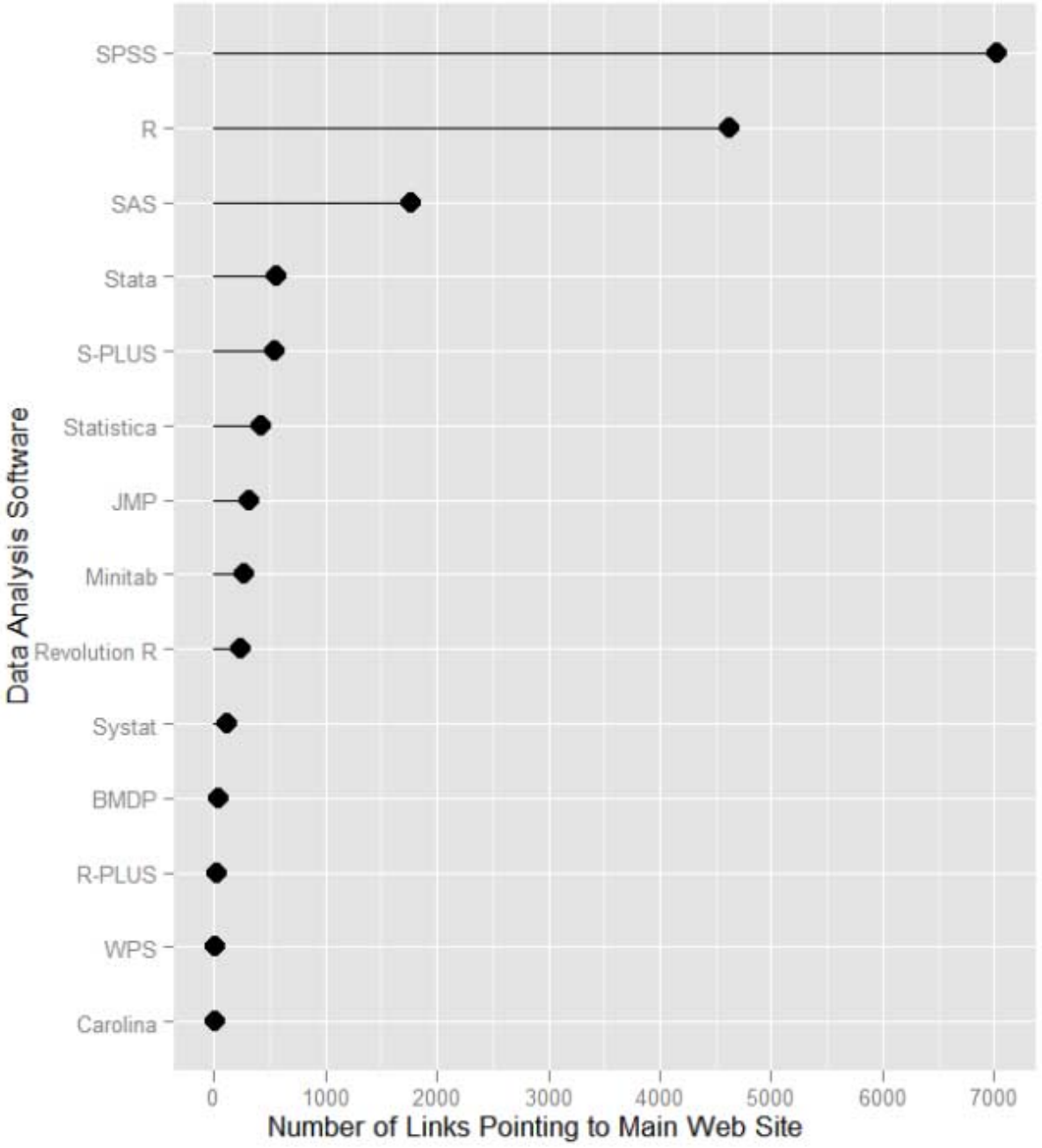


Figure 8. The number of web site links that point to the main web site of each software package.

Now that SPSS is part of IBM, it dominates the results. This reflects the wide range of products that IBM sells, including computer hardware and services that have nothing to do with data analysis. However, the older SPSS.com website no longer shows up early in a web search and

the IBM site that it redirects to has a tiny incoming link measure since it is not meant to be a direct link.

R is next in line with a little over half of IBM's measure, followed by SAS with well less than R's value. The other software follows in the order that I suspect is reflective of their respective market shares. Revolution R Enterprise and R-PLUS are commercial versions of R that are relatively quite new to the market. WPS is an implementation of the SAS Language and Carolina is a SAS-to-Java compiler.

The number of incoming links is an important part of Google's famous PageRank algorithm (http://en.wikipedia.org/wiki/PageRank). PageRank is made more useful for searching by (among other things) weighting the importance of each link. Links from major sites like WikiPedia would carry far more weight than would a link from a professor's course syllabus. The practical range of PageRank is from 1 to 10. Figure 9 plots this data (collected on on January 4, 2012). The software appear in tiers, with the two dominant players, SAS and SPSS (IBM), at the highest, and their well-known alternatives one level down. I find it odd that Stata is not in this level. At the very bottom are the World Programming System (WPS) and Carolina, two companies that use the SAS language. There have been quite a few changes in this ranking since last year, with SAS, SPSS and Revolution Analytics moving up one point and R, Stata and Carolina moving down one point. The R-PLUS site maintained its PageRank of 5 this year, which is a bit surprising given that many of its links are broken, and it is in its fourth year of saying, "Be the first to get R-PLUS 3.3″

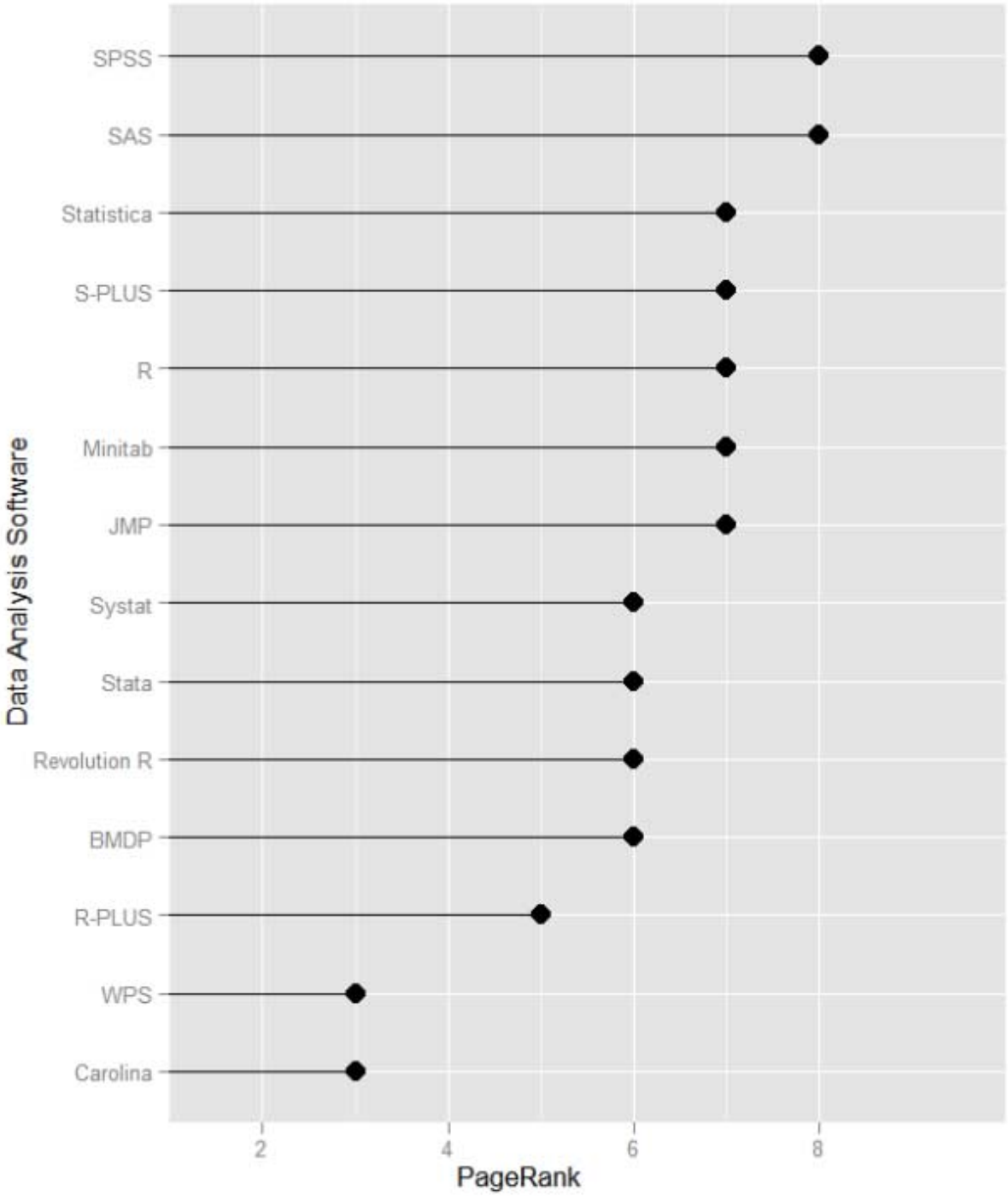


Figure 9. Google PageRanks of each software's web site.

Growth in Capability

The capability of all the software in this article has grown significantly over the years. It would be helpful to be able to plot the growth of each software package's capabilities, but such data is

hard to obtain. John Fox (2009) acquired it for R's main distribution site http://cran.r-project.org/. I collected the data for later versions following his method.

Figure 10 shows that the growth in R packages is following a rapid parabolic arc (quadratic fit with R-squared=.995). Early version numbers of R increase by 0.10 while more recent ones increased by 0.01. To make the x-axis consistent, the graph displays simply the numerical order in which the versions were released. The right-most point is for version 2.15.2, the last version released in 2012.

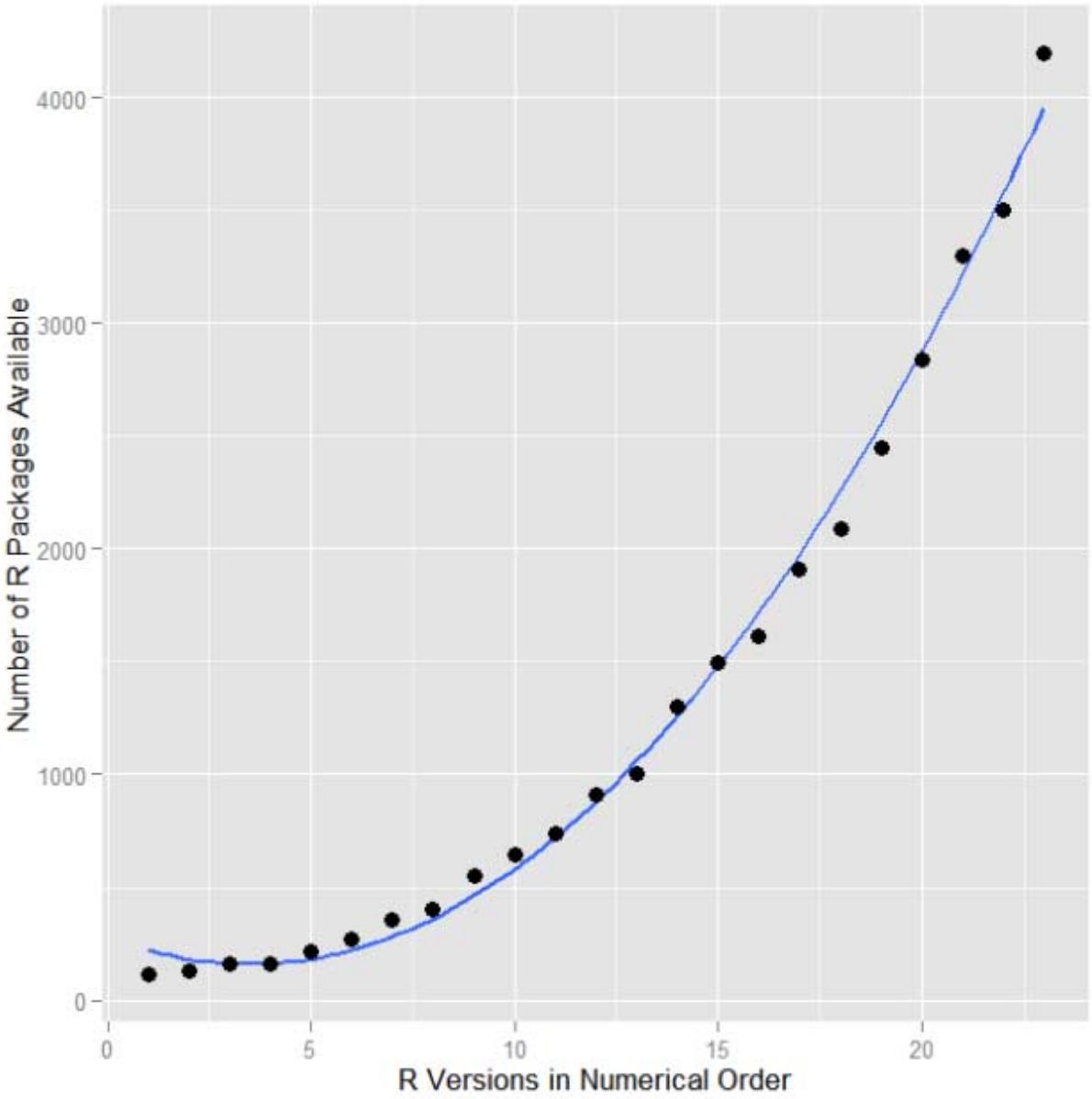


Figure 10. Number of R packages plotted for each major release of R. The last value on the x-axis represents version 2.15.2, the final release in 2012.

As rapid as this growth has been, the data in Figure 10 represents only the main CRAN repository. R does have eight other software repositories, such as the one at http://www.bioconductor.org/ that are not included in this graph. A program run on 3/19/2013 counted 6,275 R packages at all major repositories, 4,315 of which were at CRAN. So the growth curve for the software at all repositories would be roughly 30% higher on the y-axis than the one shown in Figure 10. As with any analysis software, individuals also maintain their own separate collections typically available on their web sites.

To put this astonishing growth in perspective, let us compare it to the most dominant commercial package, SAS. In its most recent version, 9.3, SAS offers around 1,200 commands that are roughly equivalent to R functions (procs, functions etc. in Base, Stat, ETS, HP Forecasting, Graph, IML, Macro, OR, QC). R packages contain a median of 5 functions (Rasmus Bååth, 12/1012 personal communication). Therefore R has approximately 31,375 functions compared to SAS' 1,200. *In fact, during 2012 alone, R added more functions/procs than SAS Institute has written in its entire history!* That's 701 packages, counting only CRAN, or around 3,505 functions. Of course these are not perfectly equivalent. Some SAS procedures have many more options to control their output than R functions do. However, R functions can nest inside one another, creating nearly infinite combinations. While the comparison is not perfect, it is certainly an eye opener.

IT Research Firms

IT research firms study software products and corporate strategies and provide their opinions on each in reports they sell to their clients. Two such reports that focus on data mining tools are here:

Forrester http://www.sas.com/news/analysts/forresterwave-predictive-analytics-dm-104388-0210.pdf

Gartner Group: http://www.spss.com.hk/PDFs/Gartner_Magic_Quadrant.pdf

Both firms rank SAS and SPSS as the top two and also predict greater than 100% annual growth for open source business intelligence software.

Job Market

Employment is important to us all, so what software skills are employers seeking? A thorough answer to this question would require a time consuming content analysis of job descriptions. However we can get a rough idea by searching on job advertising sites. Indeed.com is the most popular job search site in the world, so I went there and searched for jobs that listed data analysis software in its requirements, searching for keywords such as "SPSS" or "Minitab." It turns out that the abbreviation "SAS" is a common abbreviation in computer storage, so I avoided those advertisements by searching for "SAS !SATA !storage !firmware" (the exclamation point represents a logical "not"). I focused on R while avoiding related topics like "R&D" by using "R SAS" or "SAS R", including each package in the graph. In previous years I included S-PLUS, but I gave up on it this year. That package almost never appeared without the combination "R or S-PLUS" and it is very hard to search for since the string "…'s, plus…" appears in many irrelevant ads. The data are presented in Figure 11 (collected January 4, 2012).

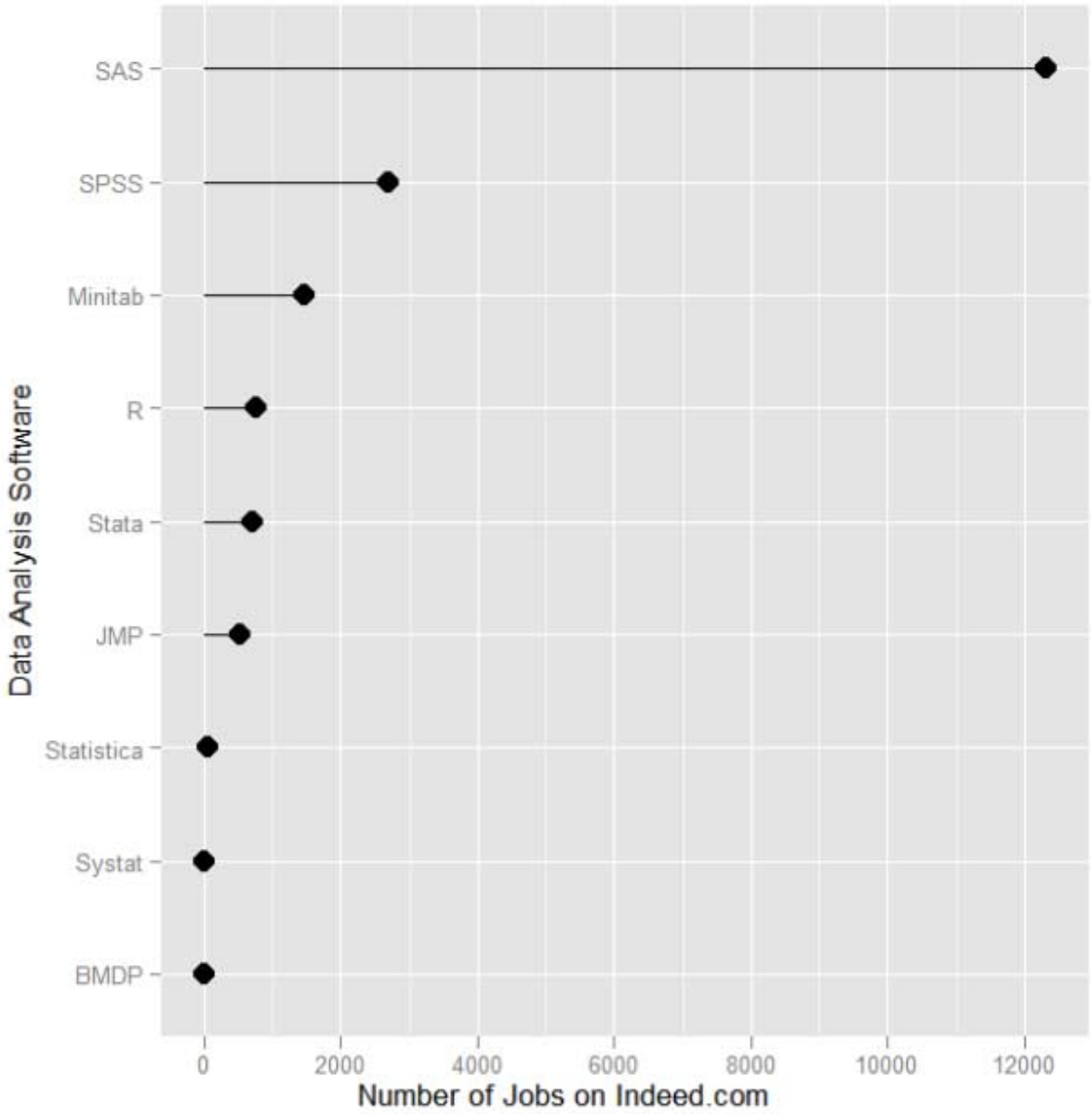


Figure 11. Number of jobs listing each software package in its requirements on Indeed.com.

SAS has a very substantial lead in job openings, with SPSS coming in second with fewer than a quarter of the jobs. Minitab had just over half the SPSS total and R had half again as many as that. A data analyst would do well to know SAS unless he or she were training for field in which one of the other packages is dominant.

What's Missing?

The most frequent question I receive about this paper is why I don't collect data on MATLAB, Mathematica, or similar open source software such as Octave, Scilab and Sage. They are, of course, quite capable of doing data analysis. However, I did not collect data on them because their use is more popular in the fields of general science and engineering, not data analysis in the

statistical or predictive analytics sense. Graphs from other sources, however, occasionally do include them.

The other thing missing is the discussion I previously included on Google Trends. That site tracks not what's actually on the Internet via searches, but rather the keywords and phrases that people are entering into their Google searches. That ended up being so variable as to be essentially worthless. For an interesting discussion of this topic, see this article by Rick Wicklin.

Conclusion

By most of the measures discussed here, R is competing well with the commercial software vendors. However, I advise not over generalizing from this data. SAS and SPSS continue to dominate the corporate world and Stata is doing quite well in the scholarly arena. Each of these packages is dominant in one market or another. I'm interested in other ways to measure software popularity. If you have any ideas on the subject, please contact me at muenchen.bob@gmail.com.

If you are a SAS or SPSS user interested in learning more about R, you might consider my book, *R for SAS and SPSS Users*. Stata users might want to consider reading *R for Stata Users*, which I wrote with Stata guru Joe Hilbe.

Acknowledgments

I am grateful to the following people for their suggestions that improved this article: John Fox (2009) provided the data on R package growth; Marc Schwartz (2009) suggested plotting the amount of activity on e-mail discussion lists; Duncan Murdoch clarified the pitfalls of counting downloads; Martin Weiss pointed out both how to query Statlist for its number of subscribers; Christopher Baum provided information regarding counting Stata downloads; John (Jiangtang) HU suggeseted I add more detail from the TIOBE index; Andre Wielki suggested the addition of SAS Institute's support forums; Kjetil Halvorsen provided the location of the expanded list of Internet R discussions; Dario Solari and Joris Meys suggested how to improve Google Insight searches; Keo Ormsby provded useful suggestions regarding Google Scholar; Karl Rexer provided his data mining survey data; Gregory Piatetsky-Shapiro provided his KDnuggets data mining poll; Tal Galili provided advice on blogs and consolidation, as well as Stack Exchange and Stack Overflow; Patrick Burns provided general advice; Nick Cox clarified the role of Stata's software repositories and of popularity itself; Stas Kolenikov provided the link of known Stata repositories; Rick Wicklin convinced me to stop trying to get anything useful out of Google Insights; Drew Schmidt automated the collection of the data in Figures 7a and 7b; Francois Briatte provided the link that creates Figure 1c; Rasmus Bååth provided the median number of functions in an R package.

Correction

An earlier version of this document listed the number of SAS commmands in just Base and Stat as 647. The figure was later revised up to 1,200 by adding procedures, functions and commands from the SAS products ETS, Graph, HP Forecasting, IML, Macro, OR, and QC.

Bibliography

J. Fox. Aspects of the Social Organization and Trajectory of the R Project. *R Journal*, http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Fox.pdf

R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

R. Muenchen, *R for SAS and SPSS Users*, Springer, 2009

R. Muenchen, J. Hilbe, *R for Stata Users*, Springer, 2010

M. Schwartz, 1/7/2009, http://tolstoy.newcastle.edu.au/R/e6/help/09/01/0517.html

Trademarks