Free Open-Source Forecasting Using R

by Stephan Kolassa and Rob J. Hyndman

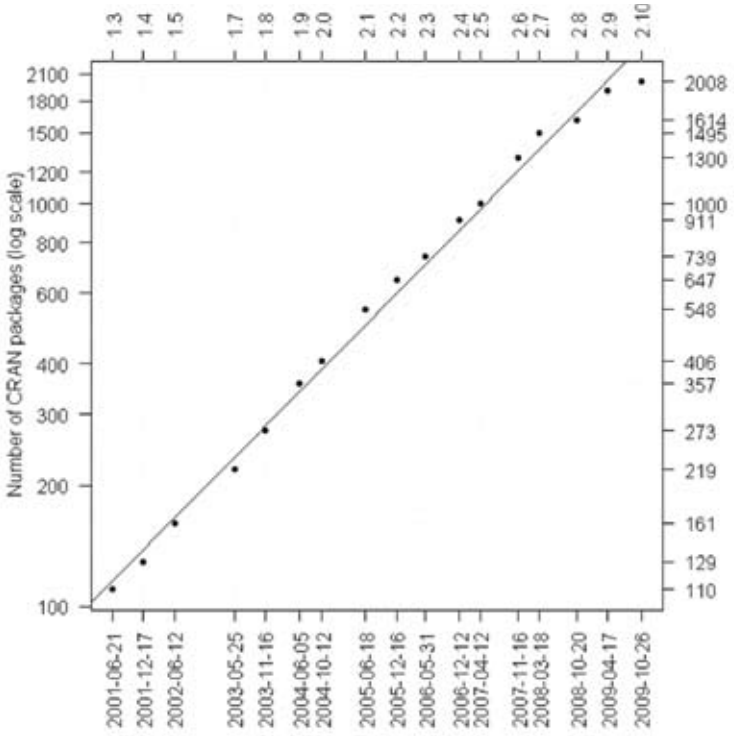
PREVIEW. *Neelie Kroes, while antitrust commissioner for the EU, said, "In the current economic context, all companies are looking for cost-effective IT solutions. Systems based on open-source software are increasingly emerging as viable alternatives to proprietary solutions." Stephan Kolassa and Rob Hyndman provide their evaluation of whether R, an open-source statistical computing environment, can be used for forecasting. They compare it favorably to professionally produced and quality-controlled commercial software.*

WHAT IS R, AND WHY SHOULD FORECASTERS CARE?

R is a statistical programming environment currently being maintained by the R Foundation and freely available under the GNU General Public License at ***www.r-project.org.*** R runs on a large number of platforms, from Microsoft Windows and MacOS to Linux. New versions are released frequently – at this writing, version 2.10.1 is the most current, but this will probably have changed by the time our review sees print. R is mainly command-line based, but Graphical User Interfaces (GUIs) such as Rcmdr (Fox, 2005) exist to make life easier for beginners.

R does all kinds of statistical calculations, and it can be extended by user-contributed "packages," which are published on the Comprehensive R Archive Network (CRAN) at ***http://cran.r-project.org.*** As shown in Figure 1, there are already over

Figure 1. The Number of User-Contributed Packages on the Comprehensive R Archive Network (CRAN) Note the logarithmic scale – the number of packages is growing exponentially.



Key Points

- R is free and available at ***www.r-project.org***.
- Advantages of R (aside from its zero cost) include the large number of user-contributed packages, its production-quality graphics, and the capability to extend its functionality by linking fast-compiled C/C++/Fortran code.
- Disadvantages of R include a steep learning curve and lack of speed when dealing with truly massive amounts of data.
- R can be advantageous to forecasters with a technical background who are comfortable with the command line. This group can use it for business intelligence analyses and graphics beyond forecasting.

2,000 packages in CRAN, covering fields as diverse as genetic microarray analyses, text mining, and inferential statistics for medicine and psychology. If you have any kind of statistics, data-mining or machine learning problem, chances are that someone already had this kind of problem before – and wrote an R package to deal with it.

Consequently, there are also packages for time-series forecasting that extend the native R forecasting capabilities. Examples include the forecast package (containing functions for exponential smoothing and ARIMA models), the Mcomp package (containing the data from the various M-competitions), the expsmooth package (containing the data used in the book by Hyndmanand colleagues, 2008), and the fma package (containing the data used in the book by Makridakisand colleagues, 1998).

A QUICK EXAMPLE

Let's walk through a short example on how to forecast using R. If you like, you can download and install R and follow along by entering the commands as you read. The code and the resulting plot are shown in Figure 2. First, we need to install the forecast package (Hyndman 2009):

```
install.packages("forecast")
```

R will probably prompt you to choose a mirror to download from – just select whatever is close to you. Next, we make the forecast package available to the current R session:

```
library(forecast)
```

And now for some serious forecasting!

The forecast package includes a dataset called "wineind," which contains Australian total monthly wine sales, by wine makers in bottles up to 1 liter, from January 1980 through August 1994. We want to forecast this time series using the exponential-smoothing model, additive Holt-Winters. This model has an additive error term, an additive undamped trend and additive seasonality. The "ets" function (for error, trend and seasonality) fits the necessary model with the state space formulation of Hyndman and colleagues (2008), doing all the parameter estimation for us. We store the resulting fitted model in an object which we will call "fitted.model," using the "<-" operator, which assigns the result of an operation to an object.

```
fitted.model <- ets(wineind, "AAA")
```

Now that fitted.model contains all the model information we need, we can use it to forecast. (If we are interested in the fitted parameters and other model information, we can look at fitted.model by simply typing "fitted.model.") For instance, we can forecast for a horizon of 120 months. In addition, we can calculate 80% prediction intervals for our forecast to get an idea of how much variability there may be in future wine production. We store the forecasts in an object called "fcst":

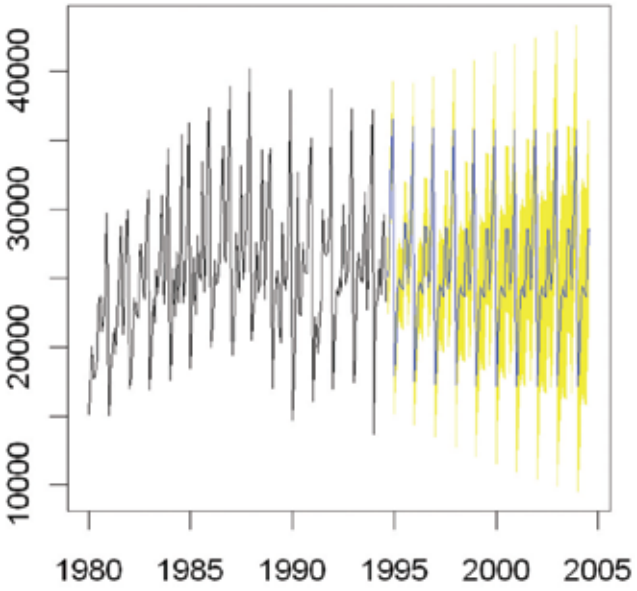
```
fcst <- forecast(fitted.model, h=120, level=80)
```

Figure 2.
R Code in the "Quick Example" and a Resulting Graph with the Time Series and the Quantile Forecast

R Code

```
install.packages("forecast")
library(forecast)
fitted.model <- ets(wineind, "AAA")
fcst <- forecast (fitted.model, h=120,
level=80)
plot.forecast(fcst, main="Quantile
Forecasts")
write.csv(fcst, "forecast.csv")
?plot.forecast
```

Quantile Forecasts



And now we can plot the historical time series and the forecasts. The 80% prediction intervals are given as yellow bands, while the point forecasts are shown in blue. In addition, we specify a nice title for our plot. Figure 2 shows the result:

```
plot.forecast(fcst, main="Quantile Forecasts")
```

Finally, we will probably want to use the forecasts in some other context, so we write them to a file "forecast.csv" in comma-separated format, which can be read by any editor or spreadsheet program.

```
write.csv(fcst, "forecast.csv")
```

Now, maybe you're not happy with yellow bands as confidence intervals – perhaps you prefer red ones. No problem! Every R function has a number of arguments we can use to change the default behavior. To find out about them, we look at the help page for the function, which we can show using "?":

```
?plot.forecast
```

Here, we learn about the "shadecols" parameter, which we can set to a different value, e.g., "shadecols=red". And we can continue tweaking the code to our heart's content!

PROS

What are the advantages of R over other forecasting software and systems?

The first and perhaps most self-evident benefit is price. R can be downloaded free of charge, along with all user-contributed packages. Take note, though, that there are companies specializing in enhancing R functionalities and providing commercial support, such as REvolution Computing (***www.revolution-computing.com/***).

R has a rapidly growing user base. University statistics departments are increasingly using and teaching R to their students. Not only is it much cheaper than other statistics packages, it also allows users to quickly implement new statistical techniques as packages to be disseminated among stat-

isticians. If you hire a recent statistics graduate as a forecaster, chances are that she or he has already worked with R.

One benefit of this exploding number of users is that there are high-traffic mailing lists, such as R-help, where advanced users will quickly answer questions you may have. And a large user base leads to more and more introductory and advanced books on R, a few of which are listed at the end of this review. Also, the more users R has, the more specific problems have already been faced by those users, which results in more packages being posted to CRAN to deal with those problems – leading to exponential growth in the number of user-contributed extensions as seen in Figure 1.

This brings us to the question that's inevitably asked about open-source software: does it compare in quality with professionally produced and quality-controlled commercial software? Legions of satisfied users of operating systems such as Linux and office applications like Open Office provide the answer. Not surprisingly, then, R passes tests on numerical stability and other quality metrics with flying colors, showing accuracy and reliability comparable to that of commercial statistical packages such as SAS or SPSS (e.g., Keeling & Pavur, 2007).

In fact, not only are the numerical and statistical capabilities of R on a level with other programs, the graphics produced by R, such as that in Figure 2, are of the highest caliber; they are production quality, even without tweaking parameters and graphics settings. When you begin to delve into the manifold graphics parameters or look at user-contributed graphics packages, you will very quickly be able to conjure impressive graphical displays to visualize your forecasts and other data. Once again, the large user base means that there are lots of tutorials and examples on the Web, e.g., at the R Graph Gallery (***http://addicted-tor.free.fr/graphiques/***).

While we are focused here on using R for forecasting, an R user will soon discover the additional possibilities offered by the software. Basically, any imaginable kind of statistical or Business Intelligence analysis can be done in R, supported by user-contributed functionality and visualized by R's graphics capabilities. Data can be retrieved from databases and reports can be generated automatically.

R also offers a complete programming environment that is thoroughly customizable. Any user can define new functions, compiled C/C++/Fortran code can be linked at runtime, and power users can implement functions in C to directly manipulate R objects. In addition, because R is open source, we can always look at the source code if we are unsure whether functionality is implemented correctly.

CONS

Is R too good to be true? Is it really the end-all to our forecasting problems, or are there disadvantages that counterbalance the merits detailed above?

One drawback to R is certainly its steep learning curve. New users who are unfamiliar with the command-line interface can quickly feel lost just reading in data. In this case, we recommend using a GUI such as Rcmdr and always having a good introductory text available when taking your first steps in R. Indeed, although R-help and other mailing lists are excellent go-to resources, it is expected that posters will first try to solve problems themselves before asking for advice – laziness will get you flamed.

R can do nearly anything you would ask of statistics and forecasting software; however, this very quality of R's generality means that you will need to write a lot of your own code if, for example, forecasts will be used for subsequent decisions, such as ordering with safety stock calculations and order optimization. Thus, specific requirements may demand specific software – or a lot of work in R.

Also, R as an interpreted language is not as quick as optimized compiled code. If you want to forecast large numbers (i.e., thousands) of time series in a time-critical environment, you could write an extension to R in compiled C or C++, or simply switch to other specialized software.

CONCLUSION – WHO IS R FOR?

R offers a low-cost way of forecasting time series and visualizing forecasts, and experienced users can utilize R for many other analyses. R is best suited for technical people who are comfortable "rolling their own" statistics and modifying default parameter settings. Being skilled in the ways of the command-line interface is definitely helpful.

R is a boon for anyone interested in additional analyses of their data – R makes it easy to forecast product sales, analyze sales data by marketing campaign and price using a regression, and finally segment customers using a classification tree. Take note, however: R may not be the best fit for novices or part-time forecasters with limited technical backgrounds.

RECOMMENDED WEB RESOURCES

- Downloads:
  R itself: ***www.r-project.org/***
  The GUI Rcmdr:***http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/***
- Forecasting packages at CRAN:
  expsmooth
  fma
  forecast
  Mcomp
  tseries
  zoo
- The CRAN Task View on Time Series at ***http://cran.r-project.org/web/views/TimeSeries.html*** has an overview of R commands and packages for time-series modeling and forecasting
- Mailing lists dedicated to R are found here: ***www.r-project.org/mail.html***
- Stack Overflow offers rated and tagged questions and answers on R: ***http://stackoverflow.com/questions/tagged/r***

REFERENCES AND LITERATURE ON R

Cowpertwait, P.S. P. & Metcalfe, A.V. (2009). *Introductory Time Series with R*, Springer Use R! Series, New York.

Data Analysts Captivated by R's Power (2009). *The New York Times*, B6, New York edition on January 7. Online at: ***www.nytimes.com/2009/01/07/technology/business-computing/07program.html***

Fox, J. (2005). The R Commander: A basic-statistics graphical user interface to R, *Journal of Statistical Software*, 14(9): 1-42.

Hyndman, R.J. (2009). Forecasting functions for time series, R package version 2.01. ***http://cran.r-project.org/package=forecast***

Hyndman, R.J., Koehler, A.B., Ord, J.K. & Snyder, R.D. (2008). *Forecasting with Exponential Smoothing – The State Space Approach*. Springer Series in Statistics, New York.

Keeling, K.B. & Pavur, R.J. (2007). A comparative study of the reliability of nine statistical software packages, *Computational Statistics & Data Analysis*, 51: 3811-3831.

Kleiber, C. & Zeileis, A. (2008). *Applied Econometrics with R*, Springer Use R! Series, New York.

Makridakis, S., Wheelwright, S.C. & Hyndman, R.J. (1998). *Forecasting – Methods and Applications (3rd ed.)*, John Wiley & Sons, New York.

R Development Core Team (2009). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Zuur, A.F., Ieno, E.N. & Meesters, E.H.W.G. (2009). *A Beginner's Guide to R*, Springer Use R! Series, New York.

Rob J. Hyndman is Professor of Statistics and Director of the Business and Economic Forecasting Unit at Monash University and Editor in Chief of The International Journal of Forecasting. He has used R and its predecessors for more than 20 years and has authored many R packages, including the forecast package noted in this article.

rob.hyndman@buseco.monash.edu.au

Stephan Kolassa is *Foresight*'s Associate Editor. His day job is Vice President of Corporate Research at SAF AG in Switzerland, where he forecasts sales at major supermarket and drugstore chains. He uses R on a daily basis for data analysis and prototyping for new forecasting and replenishment algorithms.

stephan.kolassa@saf-ag.com