R News

The Newsletter of the R Project Volume 5/2, November 2005

# Editorial

*by Douglas Bates*

One of the strengths of the R Project is its package system and its network of archives of packages. All useRs will be (or at least should be) familiar with CRAN, the Comprehensive R Archive Network, which provides access to more than 600 contributed packages. This wealth of contributed software is a tribute to the useRs and developeRs who have written and contributed those packages and to the CRAN maintainers who have devoted untold hours of effort to establishing, maintaining and continuously improving CRAN.

Kurt Hornik and Fritz Leisch created CRAN and continue to do the lion's share of the work in maintaining and improving it. They were also the founding editors of *R News*. Part of their vision for *R News* was to provide a forum in which to highlight some of the packages available on CRAN, to explain their purpose and to provide an illustration of their use.

This issue of *R News* brings that vision to fruition in that almost all our articles are written about CRAN packages by the authors of those packages.

We begin with an article by Raftery, Painter and Volinsky about the BMA package that uses Bayesian Model Averaging to deal with uncertainty in model selection.

Next Pebesma and Bivand discuss the sp package in the context of providing the underlying classes and methods for working with spatial data in R. Early on they mention the **task view** for spatial statistical analysis that is being maintained on CRAN. Task views are a new feature on CRAN designed to make it easier for those who are interested in a particular topic to navigate their way through the multitude of packages that are available and to find the ones targeted to their particular interests.

To provide the exception that proves the rule, Xie describes the use of condor, a batch process scheduling system, and its DAG (directed acylic graph) capability to coordinate remote execution of long-running R processes. This article is an exception in that it doesn't deal with an R package but it certainly should be of interest to anyone using R for running simulations or for generating MCMC samples, etc. and for anyone who works with others who have such applications. I can tell you from experience that the combination of condor and R, as described by Xie, can help enormously in maintaining civility in a large and active group of statistics researchers who share access to computers.

Continuing in the general area of distributed statistical computing, L'Ecuyer and Leydold describe their rstream package for maintaining multiple reproducible streams of pseudo-random numbers, possibly across simulations that are being run in parallel.

This is followed by Benner's description of his mfp package for multivariable fractional polynomials and Sailer's description of his crossdes package

## Contents of this issue:

for design and randomization in crossover studies.

We round out the issue with regular features that include the R Help Desk, in which Duncan Murdoch joins Uwe Ligges to explain the arcane art of building R packages under Windows, descriptions of changes in the 2.2.0 release of R and recent changes on CRAN and an announcement of an upcoming event, useR!2006. Be sure to read all the way through to the useR! announcement. If this meeting is anything like the previous useR! meetings – and it will be – it will be great! Please do consider attending.

This is my last issue on the *R News* editorial board and I would like to take this opportunity to extend my heartfelt thanks to Paul Murrell and Torsten Hothorn, my associate editors. They have again stepped up and shouldered the majority of the work of preparing this issue while I was distracted by other concerns. Their attitude exemplifies the spirit of volunteerism and helpfulness that makes it so rewarding (and so much fun) to work on the R Project.

*Douglas Bates*
*University of Wisconsin – Madison, U.S.A.*
bates@R-project.org

# BMA: An R package for Bayesian Model Averaging

*by Adrian E. Raftery, Ian S. Painter and Christopher T. Volinsky*

Bayesian model averaging (BMA) is a way of taking account of uncertainty about model form or assumptions and propagating it through to inferences about an unknown quantity of interest such as a population parameter, a future observation, or the future payoff or cost of a course of action. The BMA posterior distribution of the quantity of interest is a weighted average of its posterior distributions under each of the models considered, where a model's weight is equal to the posterior probability that it is correct, given that one of the models considered is correct.

Model uncertainty can be large when observational data are modeled using regression, or its extensions such as generalized linear models or survival (or event history) analysis. There are often many modeling choices that are secondary to the main questions of interest but can still have an important effect on conclusions. These can include which potential confounding variables to control for, how to transform or recode variables whose effects may be nonlinear, and which data points to identify as outliers and exclude. Each combination of choices represents a statistical model, and the number of possible models can be enormous.

The R package `BMA` provides ways of carrying out BMA for linear regression, generalized linear models, and survival or event history analysis using Cox proportional hazards models. The functions `bicreg`, `bic.glm` and `bic.surv`, account for uncertainty about the variables to be included in the model, using the simple BIC (Bayesian Information Criterion) approximation to the posterior model probabilities. They do an exhaustive search over the model space using the fast leaps and bounds algorithm. The function `glib` allows one to specify one's own prior distribution. The function `MC3.REG` does BMA for linear regression using Markov chain Monte Carlo model composition (MC$^3$), and allows one to specify a prior distribution and to make inference about the variables to be included and about possible outliers at the same time.

## Basic ideas

Suppose we want to make inference about an unknown quantity of interest $\Delta$, and we have data $D$. We are considering several possible statistical models for doing this, $M_1, \dots, M_K$. The number of models could be quite large. For example, if we consider only regression models but are unsure about which of $p$ possible predictors to include, there could be as many as $2^p$ models considered. In sociology and epidemiology, for example, values of $p$ on the order of 30 are not uncommon, and this could correspond to around $2^{30}$, or one billion models.

Bayesian statistics expresses all uncertainties in terms of probability, and make all inferences by applying the basic rules of probability calculus. BMA is no more than basic Bayesian statistics in the presence of model uncertainty. By a simple application of the law of total probability, the BMA posterior distribution of $\Delta$ is

$$p(\Delta|D) = \sum_{k=1}^{K} p(\Delta|D, M_k)p(M_k|D), \qquad (1)$$

where $p(\Delta|D, M_k)$ is the posterior distribution of $\Delta$ given the model $M_k$, and $p(M_k|D)$ is the posterior probability that $M_k$ is the correct model, given that one of the models considered is correct. Thus the BMA posterior distribution of $\Delta$ is a weighted average of the posterior distributions of $\Delta$ under each of the models, weighted by their posterior model probabilities. In (1), $p(\Delta|D)$ and $p(\Delta|D, M_k)$ can be prob-