8-1-2014

Support Of Teaching And Research In Hydroinformatics With R

Milan Cisty

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

HYDROINFORMATICS EDUCATION AND RESEARCH WITH R

MILAN CISTY, LUBOMIR CELAR, ANNA BECOVA
Department of Land and Water resources Management, Faculty of Civil Engineering, Slovak University of Technology, 813 68 Bratislava, Slovak Republic
milan.cisty@stuba.sk, lubomir.celar@stuba.sk, anna.becova@stuba.sk

This review paper will deal with the possibilities of applying the R programming language in water resources and hydrologic applications in education and research. The objective of this paper is to present some features and packages that make R a powerful environment for analyzing data from the hydrology and water resources management fields, hydrological modelling, the post-processing of the results of such modelling, and other tasks.

INTRODUCTION

R is an open-source software for statistical computing, which means that R is freely available, so its users are free to see how it is written and improve or extend its possibilities. The last characteristic is particularly important from the point of view of this review, because the possibility of its extension is widely used by R users from many different backgrounds. Consequently, this leads to one of the best things about R, which is the large amount of existing add-ins (so-called packages), which are aimed at solving various tasks in different fields, including hydrology, water resources, climatology, soil science and meteorology. These packages can be optionally loaded into the basic R environment, which permits access to both well-established and experimental computational methods from different fields.

APPLICATIONS OF R IN HYDROLOGY AND RELATED SUBJECTS

Hydrological modellers spend a large amount of time on various data pre-processing and post-processing tasks. R is logically a powerful environment for such tasks, because it is generally oriented towards management of data and their statistical analysis. Because of the possibility of using basic R literature for this subject, we will not go into much detail here, and only “water specific“ tools in the R description follow.

The *waterData* package [1] allows users to import the U.S. Geological Survey (USGS) daily hydrological time series data into R; it cleans, plots and summarizes the imported data and calculates and plots eventual streamflow anomalies.

With regard to missing data in general, many modelling functions in R offer options for dealing with missing values, e.g., for some R functions data should not necessarily have to be complete. Besides this, good missing data functionality can be accessed through various other (non-hydrological) R packages, e.g., *Amelia II* [2], *Mice* [3], *mitools* [4] or *VIM* [5], which offer general missing data functions that are also suitable for hydrological data sets.

Another interesting package, which contains functions to support the processing and exploration of data, is *wq*, which was originally developed for monitoring aquatic ecosystems.

The name of the *wq* [6] package stands for water quality and reflects a focus on time series data describing the physical and chemical properties of water, as well as plankton. However, many of the functions should be useful for a time series analysis regardless of the subject matter. E.g., the function *mannKen* does a Mann-Kendall test of trends in a time series (it includes a seasonal alteration of this function); the *decompTs* accomplishes multiplicative and additive decomposition of time series), etc.

A very interesting R package is *hydroTSM* [7], which provides functions for the management, analysis, interpolation and plotting of time series used in hydrology and related environmental sciences. Various conversion functions are available for obtaining, e.g., monthly, annual or seasonal time series from daily data. Automatic interpolation for a hydrological time series with an optional plot could be accomplished by the *hydrokrige* function from this package.

Hydrological Modelling

There are several steps in hydrological modelling: preprocessing the data, sensitivity analysis (identification of the responsiveness of the model parameters); calibration, e.g., tuning the model parameters by checking the results of the modelling against observations by utilizing graphs and various goodness-of-fit statistics, the modelling itself by various types of models (data driven, conceptual, physically based), the validation of the model, and evaluating the results and their visualization.

A sensitivity analysis of the models could be supported by the *sensitivtity* and *fast* packages. The *sensitivity* R package [8] contains a collection of functions for a global sensitivity analysis of a model's output. *fast* is an implementation of the Fourier Amplitude Sensitivity Test, which is a method used to determine the global sensitivities of a models parameter changes with relatively few runs of the model (which is useful in the case of, e.g., physically-based models). The R package *FME* is a modelling package designed to confront a mathematical model with data. It includes algorithms for sensitivity and Monte Carlo analysis, parameter identifiability and model fitting; it also provides a Markov-chain based method to estimate parameter confidence intervals.

Regarding the modelling itself, great support is available for data-driven models; some tools are available for conceptual models as well as for physically-based hydrological models. Physically-based models solve exact physical equations (differential equations), usually on the basis of spatially distributed inputs. Although it is better to accomplish such a type of modelling, which is usually computationally demanding in compiled languages, there is also some support for this type of modelling in R.

The R package *R-SWAT-FME* [9] is a comprehensive modelling framework that adopts an R package *Flexible Modelling Environment* [10], and Soil and Water Assessment Tool model [11]. This framework provides the functionalities of parameter identifiability, model calibration, sensitivity and uncertainty analysis with instant visualization. The Soil and Water Assessment Tool (SWAT) is a semi-distributed hydrological model jointly developed by the

USDA Agricultural Research Service and the Texas A&M AgriLife Research. SWAT is a small watershed-to-river basin-scale model to simulate the quality and quantity of surface and ground water and predict the environmental impact of land use, land management practices, and climate change.

Many optimization functions in R allow for the interfacing of any computer simulation model with them in the calibration process. E.g., the *genalg* package contains *rbga*, an implementation of a genetic algorithm; the *DEoptim* package provides a global optimizer based on a differential evolution algorithm; the *cmaes* package implements a global optimization procedure using a covariance matrix–adapting evolutionary strategy (CMA-ES). As can be seen from its name, the h*ydroPSO* package [12] is more specific and is mostly intended for the calibration of environmental and hydrological models.

So-called conceptual models describe the main features of an idealized hydrological cycle. The *hydromad* package [13] provides a modelling framework, which supports such a model design (a conceptual model could be composed from the predefined functions of this package), and also offers functions for the simulation, estimation and visualization of the results modeled. Another conceptual hydrological model is available in the *TUWmodel* package. The *TUWmodel* [14] is a lumped conceptual rainfall-runoff model, which follows the structure of the well-known HBV model. The model runs on a daily time step and consists of a snow routine, a soil moisture routine, and a flow routing routine.

Data-driven models analyse and derive results only from the observed input (e.g., in the case of rainfall-runoff modelling, from temperatures, evapotranspiration or rainfall) and output of a modeled system (e.g., a flow in the case of modelling watershed processes. R, as a language for statistics, is particularly competitive in this modelling area in comparison with other software environments. For this reason we will not review in this paper potential full extent of this subject (there is too much functionality which R can offer in this area for one paper.

Spatial Data Manipulation

R's ability to analyze and visualize data makes it a good choice for spatial data analysis. For some spatial analysis projects, using only R may be sufficient. In many cases, however, R can be used in conjunction with GIS software. It is better not to try to substitute GIS with R if it is necessary to do specialized GIS tasks, e.g., an interactive display or the editing of spatial data. The core R engine was not designed specifically for the display and analysis of maps, and the limited interactive facilities it offers have drawbacks in this area [15].

The basic package for the definition of the various types of spatial object structures (e.g., points, lines, polygons or grids) in R is *sp* [15]. Several utility functions are provided in sp, e.g., for conversion between data formats, plotting maps, spatial selection and overlays, as well as methods for retrieving coordinates, or for subsetting, printing, summarizing data, etc.

Various packages serve for accessing vector data, e.g., *RArcInfo* [16] allows *ArcInfo* v.7 binary files and *.e00 files to be read; also the *maptools* [17] and *shapefiles* [18] packages read and write ArcView shapefiles and various other formats. The *rgdal* package [19] provides functions to read and write a lot of grid and vector formats, and it provides access to projection and transformation operations. The *rgdal* package provides an interface to the GDAL/OGR library, which powers the data import and export capabilities of many geospatially aware software applications.

The *raster* package [20] provides access to data in raster formats and includes analytical tools for this type of spatial data. The *raster* package provides, among other things, the creation of raster objects from scratch or from a file, the handling of extremely large raster files, raster algebra and overlay functions, distance functions, polygons, lines and points to raster conversion, summarizing raster values, easy access to raster cell values, plotting, reading and writing various raster file types. The *rasterVis* package complements [21] the *raster* package, which provides a set of methods for enhanced visualization and interaction.

Vector data manipulation, e.g., topology operations on geometries, are accessible with the help of the functions from the *rgeos* package [22]. It contains many functions for handling, combining, and querying points, lines, and polygon types of spatial data. There is more than one hundred of such functions available in the *rgeos* package.

The *gstat* package [23] offers a wide range of univariable and multivariable geostatistical modelling methodologies, prediction and simulation functions, variogram modelling, variogram map plotting, everything from simple global kriging to local universal cokriging, multivariate geostatistics, block kriging, etc.

Web-based services are becoming ever more important channels for exchanging spatial data. The *RgoogleMaps* [24] package provides tools to access Google MapsTM data in an image form using the Google Static Maps API, in order to permit background maps to be used in R. The *ggmap* package allows for the easy visualization of spatial data on top of Google Maps, OpenStreetMaps, Stamen Maps, or CloudMade Maps using *ggplot2*. *ggplot2* [25] is one of the most popular packages in R intended as an alternative for data visualization; it is based on Leland Wilkinson’s grammar of graphics. It provides a scheme for data visualization, which breaks up the creation of graph into a layers. In *ggmap* usage, a basic layer, e.g., from Google Maps, is firstly downloaded, and then its object is created in R. Then other layers with lines, points, polygons, texts and other features from various sources are added to it according to its rules of syntax. Vector data from OpenStreetMap is also available for downloading by using the recently contributed *osmar* package [26].

CONCLUSIONS

In this paper a selection of possibilities offered by the R development core team, as well as by the work of various contributors in the areas of hydrology and water resources management, was described. As can be seen, there are a useful collection of options, even on a level where the user is actually not programming something. It is necessary to know the basic syntax of the language, but what was basically presented in the paper were ready-made functions. The only obstacle for less-experienced users could be that these functions are not managed by a clickable GUI, but through a command line interface. Moreover, in the case of also exploiting all the programming possibilities of R, the effect of using R for water practitioners or scientists will be even more useful.

Although having a tool is not enough for serious work which involves other subject-related theoretical and practical knowledge and skills, a tool such as R is very useful, e.g., in the process of learning some difficult subject related to an analysis of hydrological data (e.g., copulas, to mention one). In R one has the possibility of easily trying corresponding computations, which are otherwise only described by complicated theories. Of course, it is necessary to know the background of the computations, but it is very helpful in the process of learning some intimidating and complicated subject, if one knows that he can do the very thing which he is trying to understand. This is supported by a unified system as to how the so-called S3 and S4 functions of R are written. For the R user it means that there is not a big difference in exploiting R's possibilities if he is interested in the capabilities of GIS or the mentioned copula computations. R is also an interesting social phenomenon with enthusiastic features in its user community; often someone makes such complicated tasks easier by offering appropriate tools, e.g., GUIs, which help to do basic tasks in such areas easily and intuitively (e.g., from the items described here in before, the extremes package for an extreme value analysis is a good example of this point, which is a de facto wizard that uses computational engines from other packages). Many R functions are wrap programs, which are written in FORTRAN, Java, C++, etc., which often makes it easier to use these programs. Moreover, an advantage is obtained when the opportunity arises to use them in a stream with the other software packages available in R. This could produce an interesting synergistic effect, at least in terms of productivity.

Because the packages contributed are the result of voluntary efforts, there is no guarantee which methods based on such a genesis exist in R. As has already been mentioned, users may need to be prepared to do a little work to find what they need, e.g., to be prepared sometimes for a little digging. Besides the disadvantages associated with it (e.g., we will not find something, because no one has done this very thing, or because somebody stopped caring about it), due to the enthusiastic elements in the R community, some existential pleasure, which is associated with treasure hunting, as a result of such digging, is often attained as a reward.

ACKNOWLEDGEMENTS

This work was supported by the Slovak Research and Development Agency under Contract No. APVV-0496-10 and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/1044/11 and 1/0243/11.

REFERENCES

[1] Ryberg K.R. and Vecchia A.V., *"waterData: An R Package for Retrieval, Analysis, and Anomaly Calculation of Daily Hydrologic Time Series Data"*. R package version 1.0.3., (2014), http://CRAN.R-project.org/package=waterData

[2] Honaker J., King G. and Blackwell M., "Amelia II, A Program for Missing Data", *Journal of Statistical Software*, Vol.45, (2011), pp 1–47.

[3] Van Buuren S. and Groothuis-Oudshoorn K., "mice: Multivariate Imputation by Chained Equations in R", *Journal of Statistical Software*, Vol.45, (2011), pp 1–47.

[4] Lumley T., *"mitools: Tools for Multiple Imputation of Missing Data"*, R package version 2.2., (2014), http://CRAN.R-project.org/package=mitools

[5] Templ M., Alfons A., Kowarik A. and Prantner B., *"VIM: Visualization and Imputation of Missing Values,"* R package version 4.0.0., (2014), http://CRAN.R-project.org/package=VIM

[6] Jassby A.D. and Cloern J.E., *"wq: Some Tools for Exploring Water Quality Monitoring Data"*, R package version 0.3-11, (2014), http://CRAN.R-project.org/package=wq

[7] Bigiarini-Zambrano M., *"hydroTSM: Time Series Management, Analysis and Interpolation for Hydrological Modelling"*, R package version 0.4-2-1., (2014), http://CRAN.R-project.org/package=hydroTSM

[8] Pujol G., Iooss B. and Janon A., *"Sensitivity: Sensitivity Analysis"*, R package version 1.7., (2014), http://CRAN.R-project.org/package=sensitivity

[9] Wu Y., and Liu S., "Automating Calibration, Sensitivity and Uncertainty Analysis of Complex Models Using the R Package Flexible Modelling Environment (FME) SWAT as an Example", *Environmental Modelling & Software,* Vol. 31, (2012), pp 99–109.

[10] Soetaert K. and Petzoldt T., "Inverse Modelling, Sensitivity and Monte Carlo Analysis in R using Package FME", *Journal of Statistical Software*, Vol. 33, (2010), pp 1–28.

[11] Arnold J.G., Srinivasan R., Muttiah R.S. and Williams J.R., "Large Area Hydrologic Modelling and Assessment; Part 1 Model Development", *Journal of the American Water Resources Association*, Vol. 34, (1998), pp 73-89.

[12] Bigiarini-Zambrano M. and Rojas R., "A Model-independent Particle Swarm Optimisation Software for Model Calibration", *Environmental Modelling & Software,* Vol. 43, (2013), pp 5–25.

[13] Andrews F.T., Croke B.F.W. and Jakeman A.J., "An Open Software Environment for Hydrological Model Assessment and Development", *Environmental Modelling & Software*, Vol. 26, (2011), pp 1171–1185.

[14] Parajka J., Merz R. and Bloschl G., "Uncertainty and Multiple Objective Calibration in Regional Water Balance Modelling: Case Study in 320 Austrian Catchments", *Hydrological Processes,* Vol. 21, (2007), 435–446

[15] Bivand R., Pebesma E. and Gómez-Rubio V., *"Applied Spatial Data Analysis with R"*, 2nd edition, Springer New York, (2013).
[16] Gómez-Rubio V. and López-Qulez A., "RArcInfo: Using GIS Data with R", *Computers & Geosciences*, Vol. 31, (2005), 1000–1006.
[17] Bivand R. and Lewin-Koh N., Maptools, *"Tools for Reading and Handling Spatial Objects",* R package version 0.8-27., (2014), http://CRAN.R-project.org/package=maptools
[18] Stabler B., *"Shapefiles: Read and Write ESRI Shapefiles",* R package version 0.7., (2014), http://CRAN.R-project.org/package=shapefiles
[19] Bivand R., Keitt T. and Rowlingson B., *"rgdal: Bindings for the Geospatial Data Abstraction Library",* R package version 0.8-14., (2014), http://CRAN. R-project.org/package=rgdal
[20] Hijmans R.J., *"Raster: Geographic Data Analysis and Modelling"*, R package version 2.1-49, (2014), http://CRAN.R-project.org/package=raster
[21] Perpinan O. and Hijmans R.J., *"rasterVis: Visualization Methods for the Raster Package",* R package version 0.27., (2014), http://CRAN.R-project.org/ package=rasterVis
[22] Bivand R. and Rundel C., *"rgeos: Interface to Geometry Engine - Open Source (GEOS)",* R package version 0.3-2., (2014), http://CRAN.R-project.org/package=rgeos
[23] Pebesma E.J., "Multi Variable Geostatistics in S: the gstat Package", *Computers & Geosciences*, Vol. 30, (2004), pp 683–691.
[24] Loecher M., "*Overlays on Google Map Tiles in R*", R package version 1.2.0.5., (2014), http://CRAN.R-project.org/package=RgoogleMaps
[25] Wickham H., "*ggplot2: Elegant Graphics for Data Analysis"*, Springer, New York, (2009).
[26] Eugster M.J.A. and Schlesinger T., *"osmar: OpenStreetMap and R"*, (2014)