Time Series RDM Green Shoots Report

A Research Data Management (RDM) "Green Shoots" Pilots
Project Report by Nick Jones
Imperial College London

This project was funded as part of Imperial College's RDM "Green Shoots" Programme. In 2014, the Vice-Provost, Research, approved an allocation of £100K for academically-driven projects to identify and generate exemplars of best practice in RDM, specifically frameworks and prototypes that would comply with key funder RDM policies and the College position. The call for projects outlined that frameworks could be based either on original ideas or integrating existing solutions into the research process, improving its efficacy or the breadth of its usage. There was an expectation that solutions would support open access for data; solutions that supported Open Innovation were strongly encouraged.

Six projects were funded, covering different disciplines, faculties and research areas. The projects ran for six months, finishing at the end of 2014. Project reports were made available in 2015.

For more information on the programme and projects please visit:

http://www3.imperial.ac.uk/researchsupport/rdm/policy/greenshoots

Imperial College London

Time Series RDM Green Shoots Report

Nick Jones; Imperial College London

What we did in one sentence

We modified our web-resource so that now users can upload time series data, and methods, and compare them in a basic manner.

Details of funding provided

We were provided with 20k. This was used to recruit a scientific web developer, Dr Philip Knaute.

Context

Univariate time series are sequences of measurements of a quantity: from a patient's hourly blood pressure measurements to stock prices recorded each millisecond. As such they underpin huge areas of science, technology and medicine. My research group has developed a method that allows new (time series) data and models/methods to be automatically placed in the context of past efforts [1, 2]. We recently developed a website (http://www.comp-engine.org/timeseries/) that allows people to download the largest interdisciplinary collection of time-series data and time-series analysis code that we know of. We also go further by allowing users to identify sets of time series (and sets of time-series analysis methods) in our database that are functionally similar to one another. In our proposal we sought to make a first attempt at extending this capability, to allow scientists, and the broader public, to automatically determine how their own methods and data are related to our extensive collections.

Our basic approach exploits feature-based comparison of time series and methods. We subject time series to thousands of distinct data analysis methods where each method takes a time series and returns a single output number. This yields a comprehensive feature vector that is informative about a wide range of structures in time series [1, 2]. We can also take time-series analysis algorithms and, by looking at their output on a set of time series, correspondingly represent them by a feature vector with one element for each output. By comparing feature vectors and using tools from machine learning we can compare and organize both time series and their analysis methods according to their empirical properties and behaviour. Importantly, many time series in our database are generated from various mathematical models (e.g., sets of ODEs, iterative maps, stochastic processes). This allows us to take empirical data and identify the model-generated data that it is most similar to. By these means we have the ability to automatically investigate the structure of connections between different empirical time-series data and models, as well as between a broad and extensive range of scientific time-series analysis methods. As our database continues to grow we will be able to place each new piece of time-series data (be it real-world or model-generated), and data analysis method, in its natural scientific context.

Activities supported

Our three major objectives were to make it possible for users to:

1. automatically profile their time series
2. automatically profile their time series algorithms
3. use these profiles to place their work in the context of others

We have delivered on these three aspects: you can explore the upload and analysis capabilities here. Users can now (after installing the free MATLAB runtime environment) analyse their time series using our bundle of analysis methods and upload the output (plus raw data) to our site while supplying metadata. Our site does a search and returns a list of closest matching time series which can then be explored in their turn. Users can also perform a similar set of steps for time series data. The website has explanatory text but we also supply a small read-me.

How does this support best practice in RDM?

An issue with data sharing is that the benefits are substantially larger for the community than for the sharer. We have made the first steps to giving users extra reasons to share their data: we make the data easy to find and we make it easy for the sharer to learn about their data. We thus doubly incentivize upload: uploading allows analysis of the data and the discovery of nearest neighbours and uploading allows the work to be automatically found by others.

Next steps

At the moment we rely on the MATLAB runtime environment. This was suitable for this short project but it is a large and time-consuming download that will likely put off many users. We are thus seeking resource to make our site quicker by moving some of our computations server-side. We will likely have to switch our code over from MATLAB to C in order to take advantage of speed.

References

[1] B. D. Fulcher, M. A. Little, N. S. Jones. Highly comparative time-series analysis: The empirical structure of time series and their methods, J. Roy. Soc. Interface 10, 20130048 (2013).

[2] B. D. Fulcher, N. S. Jones. Highly comparative, feature-based time-series classification. IEEE Transactions in Knowledge and Data Engineering 26, 3026 (2014).

[3] E. Keogh, S. Kasetty. On the need for time series data mining benchmarks: A survey and empirical demonstration. Data Min. Knowl. Disc. 7, 349 (2003).