Institution: University of Bristol

Unit of Assessment: 10 - Mathematical Sciences

Title of case study: Using the data to choose the best model for a statistical analysis, using Reversible Jump Markov chain Monte Carlo: generic model choice for an evidence-informed society

1. Summary of the impact (indicative maximum 100 words)

Reversible Jump Markov chain Monte Carlo, introduced by Peter Green [1] in 1995, was the first generic technique for conducting the computations necessary for joint Bayesian inference about models and their parameters, and it remains by far the most widely used, 18 years after its introduction. The paper has been (by September 2013) cited over 3800 times in the academic literature, according to Google Scholar, the vast majority of the citing articles being outside statistics and mathematics. This case study, however, focusses on substantive applications outside academic research altogether, in the geophysical sciences, ecology and the environment, agriculture, medicine, social science, commerce and engineering.

2. Underpinning research (indicative maximum 500 words)

Statistical analysis of data is a ubiquitously dominant ingredient in evidence-based decision making across virtually all fields of human endeavour; most of such analysis is based on statistical models, and much of this either entails formal choice between models with differing numbers of parameters, or requires models with variable-dimension parameters. The research underpinning this impact case study consists of work carried out from 1993 at the University of Bristol by Peter Green, culminating with the 1995 publication of a paper [1] in *Biometrika*, which introduced Reversible Jump Markov chain Monte Carlo (RJMCMC), a simulation-based methodology for fitting Bayesian statistical models that have variable-dimension parameters. Mathematically, Reversible Jump is formalism for Metropolis-Hastings MCMC on a general state space consisting of a countable union of Euclidean spaces of differing dimensions. The paper included 3 illustrative applications. Over the following few years, Green developed many more substantial applications of RJMCMC in collaborative research projects, and several resulting publications [including 2-3] have themselves all stimulated further research and are well-cited.

3. References to the research (indicative maximum of six references)

*[1] Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711-732. DOI: 10.1093/biomet/82.4.711
*[2] Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). Read to the Royal Statistical Society on 15 January 1997. *Journal of the Royal Statistical Society* (B), **59**, 731-792. DOI: 10.1111/1467-9868.00095
*[3] Giudici, P. and Green, P. J. (1999). Decomposable graphical Gaussian model determination, *Biometrika*, **86**, 785-801. DOI: 10.1093/biomet/86.4.785
* reference that best indicates the quality of the underpinning research

4. Details of the impact (indicative maximum 750 words)

The paragraphs below describe in brief a wide range of non-academic applications of Reversible jump MCMC, all starting or continuing after 2008. Each of these is corroborated by personal communications held on file, and in some cases also by internal or published documents citing [1] (and sometimes [2] or [3]).

A. Applications in Geophysical sciences

1. Geophysical source reconstruction. At *Defence Research and Development, Canada*, key concepts enunciated in [1] have been used to design an innovative Bayesian inference methodology to address the problem of source reconstruction for the difficult case of multiple sources when even the number of sources is unknown a priori. This effort has had an impact within the realm of public safety and security as it addresses a critical capability gap in current emergency and retrospective management efforts, which involve the covert release of chemical, biological, or radiological agents into the atmosphere.
2. Geophysical electrical resistivity. The *US Geological Survey* is using methodology built on [1] to

explore the space of subsurface electrical resistivity models that are consistent with airborne geophysical data. Data is acquired using airborne geophysical instruments that are sensitive to the spatial distribution of electrical resistivity below ground to depths of ~100m that, in turn, can be interpreted in terms of geologic or hydrologic properties. "There is real-world impact to this work- we are using this algorithm to characterize important groundwater aquifer systems and permafrost in various areas of the U.S." [a]
3. Ground flow models. The *Belgian Nuclear Research Centre* (*SCKCEN*) has developed an MCMC simulation of a highly parameterized groundwater flow model, based on [1], for uncertainty quantification of subsurface transport in the context of the Belgian nuclear waste disposal program.
4. Air pollution, greenhouse gases, remote sensing. *Shell Research* uses Bayesian inference, exploiting MCMC techniques including [1], to estimate the characteristics of sources of airborne species (gases, particulate matter, etc.) [b]. The main value of inference from remote sensing of airborne species to Shell, and to society in general, is to be able to quantify contributions to greenhouse gas emissions from specific human activities over time. The technology is also generally useful in detecting unknown or unanticipated sources ('leaks') of species carried on the wind, and can lead to discovery of (e.g.) new hydrocarbon reserves. A key general ingredient of useful statistical solutions to real-world problems is the flexibility and scalability of Bayesian inference using MCMC, e.g. allowing characterisation of parameters previously consigned to the 'too difficult to measure or estimate' box. Reversible jump MCMC extends this flexibility considerably by allowing dimension-jumping.
5. Air pollution, change point models. *Cox Associates* have used [1] "in advocating (to risk analysts and regulators, in various forums) the importance and practicality of applying better statistical methods for causal analysis of health effects of key regulations, such as air pollution regulations in the U.S". They have testified on the importance and practicality of using better methods of causal analysis in air pollution health effects research before the Subcommittee on Energy and Power of the House Energy and Commerce Committee of Congress on health effects of air pollutants (2012) http://energycommerce.house.gov/hearings/hearingdetail.aspx?NewsID=9594. [c]
6. Climate and land models. The *Geophysical Fluid Dynamics Laboratory* (GFDL) of the US *National Oceanic and Atmospheric Administration* (NOAA) uses [1] in development of land components for climate and Earth System models. These models are needed to make climate projections (e.g. Intergovernmental Panel on Climate Change and national assessments) and seasonal climate predictions. This approach has been used to estimate parameters of the phenology module which is incorporated into a new land model. Furthermore, this new parameterization has been incorporated into a new model of forest dynamics for forest management (a collaborative project with the US Forest Service).

B. Applications in Ecology and the Environment
1. Phylogenetics and biodiversity. Modern molecular phylogenetic inference is of central importance to monitoring species diversity within changing environments. Several projects within *Agriculture and Agri-Food Canada* aim to monitor and understand general features of biodiversity, and therefore rely heavily and phylogenetic inference. Such inference, when conducted probabilistically, rests on explicit models of molecular evolution, and the approach in [1] helps choose the most appropriate one, or allow phylogenies to be based on a weighted-averaging over all possible models of a given class. Algorithms based on [1] are implemented in several phylogenetic inference packages and continue to stimulate applied research in their labs.
2. Phylogenetics and biodiversity. The *Morton Arboretum* in Lisle, Illinois, uses [1] to characterize shifts in chromosome number evolution as a way of understanding biodiversity shifts in sedges Chromosome number evolves independently of genome size in a clade with non-localized centromeres, as well as understanding macroevolutionary shifts in decomposition rates on the tree of life. A lead scientist at Morton Arboretum comments that "Both of these have profound implications for management and conservation of biodiversity as well as ecosystem processes".
3. Animal abundance. *NOAA* uses [1] to allow uncertainty in the number of individuals when estimating the abundance of organisms in line transect surveys with imperfect detection. It is "currently using this type of analysis is to estimate the number of seals in the Bering Sea".
4. Wildlife ecology. At the *US Geological Survey* Patuxent Wildlife Research Center, [1] has been applied "to analyses of the North American Breeding Bird Survey, to toxicological studies, to basic ecological work on life histories, and in demographic analyses".

5. Ecology of salmon. The *US Fish and Wildlife Service* uses [1] in comparing models aimed at assessing the effect of transporting (exporting) water from the Sacramento - San Joaquin Rivers Delta on the survival of juvenile salmon as the salmon were out-migrating from freshwater to the ocean. Water is exported from the Delta for agricultural, municipal, and personal needs and is thought to directly affect over 25 million people in California. Coincident with the increase in water exports over the last 50 years or so, there have been sizable declines in the abundance of several fish species in the Sacramento and San Joaquin river systems. Conflicts have arisen between various stakeholders and interest groups regarding how the water is used and divided up. There have been many lawsuits and court cases. This work was discussed at length in the US Federal District Court (Fresno, California) in April 2010.
6. Ecology, conservation, environment. For *Land Care Research NZ,* [1] "plays an important role on analysing data that has a downstream effect on evolution, ecology and conservation biology" - and hence environment protection.

C. Agricultural applications
Quantitative Trait Loci (QTLs) in agriculture. At the *national agricultural research centre of Japan (AFFRC),* "in our … genome analysis of livestock and crops, some useful QTLs affecting traits of agronomical importance were detected with the developed methods implementing RJ-MCMC. A project to produce new cultivars of crop (tomato) or breeds of pig with high genetic performance using the information from the detected QTLs is now in progress". [d]

D. Medical applications
Protein-DNA interactions and medical implications. Projects at the *UC Denver Medical School* utilizing techniques based on [1] include (i) predicting which human variations or mutations are likely to impact protein structure and function, thereby causing human disease. This is particularly important in rare childhood developmental and neurological diseases; (ii) understanding the relationships among humans in order to improve interpretation of genome-wide association studies, finding genes that are components of quantitative disease; (iii) understanding the role of the interaction of T-cell receptors and major histocompatibility complex (MHC molecules) on defending disease, and also on causing autoimmune disease when things go wrong; (iv) using these methods to understand and predict transcription factor binding mutations, which also can lead to disease and disease or drug interaction modifiers; (v) understanding the biology of transposable elements, which are often heavily implicated in novel diseases, particularly neurological disease, and can also be useful for predicting gene regions that are likely disease-causing mutations.

E. Social and commercial applications
Exchange reserves and Criminology in India. (i) Models quantifying sufficiency of foreign exchange reserves: the *Reserve Bank of India* uses [1] for variable selection within a quantile regression model framework for studying adequacy of foreign exchange reserves to meet US$ demand in India under stressful market conditions. Due to the in-house nature of these models, these are not published or shared.
(ii) Models for studying crime rates in different states of India: a project at the Reserve Bank is attempting to determine relevance of socio-economic variables in determining level of crimes in Indian states. "Understanding the relevance of various factors in determining crime rate is very important in controlling crime rate. For instance, lack of toilet and drinking water facilities require women in India to go away from her house/hutment, which increase rape rate. Thus, a positive association between crime against women and lack of toilet/drinking water facilities demand public policy in developing these basic necessities. Probabilistic models [based on [1] are] likely to throw light on such aspects of crimes in India".

F. Applied image analysis and computer vision
1. Computer vision - object tracking. At *SORMEA*, a French company specializing in measurements and surveys, studies and modelling, Geographic Information Systems, acoustics and product development to improve road safety, an automatic vision-based multi-vehicle tracking system for measuring vehicle flows on crossroads & roundabouts, yielding provenance & destination statistics, has been constructed using [1]. These statistics are requested by local

communities in order to decide whether or not to create, extend or modify crossroads & roundabouts. The system currently is commercially exploited:
http://www.sormea.fr/fr/r-d-innovation-anacomda/anacomda-o-d.html
2. Imaging of geosynchronous orbits, managing space debris. The Lawrence Livermore National Laboratory of the *US Department of Energy* conducted a project to understand how conventional astronomical facilities might aid in determining the distribution of space debris in geosynchronous orbit. The principal application of this work is to protect valuable space assets from collisions with debris. [1] was applied "to select different possible pairings of orbital tracks seen in optical telescopes in a Monte Carlo framework". [e]

G. Implementations of RJMCMC in Software
1. Mr Bayes: [1] is used as one of several standard techniques in the software MrBayes [f]. The use of reversible jump MCMC was recently expanded to integration over nucleotide substitution models, and this is quickly becoming a standard procedure in analyses using the software. The software is widely used across the life sciences for comparative genetics and genomics studies, and more generally for studies in evolutionary biology. The software has attracted more than 16,000 citations to date. It is used widely in research but also in a number of applied contexts. One applied context concerns the identification of strains of disease organisms. Another focuses on phylogenetic studies (inference of evolutionary trees). Among other things, the evolutionary trees form the basis for classifications used in natural history museums around the world and in a wide range of applications related to the environment.
2. WinBugs: The WinBugs system [g] is respected software for Bayesian analysis, widely used by applied statisticians in both the private and public sectors, and its scope has been recently extended to support fitting of a wide range of trans-dimensional models, including variable selections, automatic curve-fitting using splines, Bayesian Multivariate adaptive regression splines (MARS) and Classification and regression trees (CART), normal mixture analysis, spatial epidemiology clustering models and variable-order Markov chains. All of these additional functions are based on [1].
3. LIS: NASA's Land Information System (LIS) [h] is a software framework for high performance land surface modelling and data assimilation. LIS is led by the Hydrological Sciences Branch at NASA's Goddard Space Flight Center. LIS software tools are used to develop customized Land Data Assimilation Systems at NASA's Goddard Space Flight Centre, NOAA's National Centres for Environmental Prediction and the Air Force Weather Agency. MCMC methods including [1] are currently being implemented and incorporated into the system.

5. Sources to corroborate the impact (indicative maximum of 10 references)
Major correspondents include:
[a] Research Geophysicist, USGS Crustal Geophysics and Geochemistry Science Center, Denver, Colorado.
Corroborates item A2 in section 4.
[b] Scientist, Statistics & Chemometrics, Shell Technology Centre, Chester, UK.
Corroborates item A4 in section 4.
[c] President of Cox Associates Consulting, Denver, Colorado.
Corroborates item A5 in section 4.
[d] Project leader, Agriculture, Forestry and Fisheries Research Council (Japan).
Corroborates item C in section 4.
[e] Research scientist, Physics division, Lawrence Livermore National Laboratory, California.
Corroborates item F2 in section 4.

Links to software cited:
[f] http://mrbayes.net
Corroborates item G1 in section 4.
[g] http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml
Corroborates item G2 in section 4.
[h] http://lis.gsfc.nasa.gov/
Corroborates item G3 in section 4.