Institution: University of Bath

Unit of Assessment: 10: Mathematical Sciences

Title of case study: Improving data analysis via better statistical infrastructure

1. Summary of the impact

A generalized additive model (GAM) explores the extent to which a single output variable of a complex system in a noisy environment can be described by a sum of smooth functions of several input variables.

Bath research has substantially improved the estimation and formulation of GAMs and hence

- driven the wide uptake, outside academia, of generalized additive models,
- increased the scope of applicability of these models.

This improved statistical infrastructure has resulted in improved data analysis by practitioners in fields such as natural resource management, energy load prediction, environmental impact assessment, climate policy, epidemiology, finance and economics. In REF impact terms, such changes in practice by practitioners leads ultimately to direct economic and societal benefits, health benefits and policy changes. Below, these impacts are illustrated via two specific examples: (1) use of the methods by the energy company EDF for electricity load forecasting and (2) their use in environmental management. The statistical methods are implemented in *R* via the software package mgcv, largely written at Bath. As a 'recommended' *R* package mgcv has also contributed to the global growth of *R,* which currently has an estimated 1.2M business users worldwide [A].

2. Underpinning research

The underpinning research was undertaken by Simon Wood (Professor at Bath since January 2006). The aim of the research programme is to make the use of *generalized additive models* as reliable and routine as the use of *generalized linear models* has long been, in order that these flexible statistical models can routinely be used beyond academic statistics.

A generalized additive model is a regression model that relates a univariate random response variable to one or more predictor variables. A key feature is that the response depends on a sum of smooth functions of the predictor variables. These functions must be estimated from the data. The flexibility to specify models in terms of unknown functions is useful in fields as diverse as fisheries science and finance, but the additional flexibility comes at the cost of decreased numerical stability and the need to estimate the degree of smoothness of the functions.

The primary contributions of the research programme undertaken at Bath are:

1. *Reliable and efficient computational methods*. The major problem is simultaneously to estimate several smoothing parameters in a computationally efficient and robust way [1, 2]. We have developed a numerical scheme for this for which convergence is guaranteed, provided that the GAM penalized likelihood has a well-defined optimum. Before the development of this method, GAM estimation methods did not always converge [B]. Before mgcv, the only software that estimated GAM smoothing parameters had $O(n^3)$ computational cost, limiting its usefulness. With mgcv the cost is about $O(n^{13/9})$.

2. *Improved means of smoothing with respect to several variables*. Smooth interactions are best represented using tensor product smooth constructions. [3] shows how this can be done in general, while maintaining the important property of scale invariance (the results should not depend on the units of measurement, for example). The generality was important as it allowed, for the first time, the routine construction of space-time smoothers for large datasets. [4] provides an alternative general construction which has the advantage of being usable as a component of any generalized linear mixed model, and also of being quite natural when ANOVA decompositions of functions are of interest. Moving to spatial smoothing, [5] uses the physical analogy of a distorted soap film, and some PDE theory, to construct a novel method for smoothing within finite geographic areas, without smoothing across boundary features. The resulting smoothers have a form that allows their full

integration into GAMs.

3. *A monograph on GAMs.* This helps statisticians beyond HE to use the methods [6].

4. *High quality software implementing the methods.* The Bath written mgcv package [7] is supplied with the *R* statistical programme as the default method for generalized additive modelling.

3. References to the research

References that best indicate the quality of the underpinning research are starred.

[1]* S.N. Wood 2008 Fast stable direct fitting and smoothness selection for Generalized Additive Models. *J. Roy. Stat. Soc. B* 70(3), 495-518. http://dx.doi.org/10.1007/s11222-012-9314-z

[2]* S.N. Wood 2011 Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. Roy. Stat. Soc. B*, 73(1), 3-36. http://dx.doi.org/10.1007/s11222-012-9314-z

[3] S.N. Wood 2006 Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* 62, 1025-1036. http://dx.doi.org/10.1007/s11222-012-9314-z

[4] S.N. Wood, F. Scheipl and J.J. Faraway 2013 Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing.* 23:341-360. http://dx.doi.org/10.1007/s11222-012-9314-z

[5] S.N. Wood, M.V. Bravington and S.L. Hedley 2008 Soap film smoothing. *J. Roy. Stat. Soc. B*, 70(5), 931-955. http://dx.doi.org/10.1007/s11222-012-9314-z

[6]* S.N. Wood 2006 Generalized Additive Models: An introduction with R. CRC Press. e.g. http://reseau-mexico.fr/sites/reseau-mexico.fr/files/igam.pdf

[7] http://cran.r-project.org/web/packages/mgcv/index.html (software package, also available in any version of R, by typing library(mgcv)).

4. Details of the impact

The main contribution of the research is to provide numerical-statistical methods that make the practical use of generalized additive models as reliable and routine as the use of generalized linear models has long been, ensuring wide uptake of the methods beyond academic research. The quality of the methods produced by the research has been recognised by the R core team's inclusion of the Bath produced software “mgcv” as one of only a dozen recommended packages (out of thousands) supplied with all versions of the R statistical computing environment.

R is an open source statistical package/environment that is widely used both in academia and beyond [A]. While the wide academic uptake of R is doubtless driven partly by the fact that it is free, this is less likely as a primary driver of business uptake, where reliability and flexibility are overriding concerns. mgcv and its underpinning research are part of providing this reliability and flexibility.

Example 1. The electricity company EDF produces 22% of Europe's electricity consumption, generating 652 TWh per year (at a wholesale price of around £40 per MWh). It is the dominant electricity producer in France, where it is also the monopoly distributor. French grid load varies between 30 and 80 GW, and with daily energy flows of around 1 Million MWh, accurate load prediction is critical to the success of EDF as a company. Moreover, as the dominant producer, substantial wider social benefits accrue to client countries through the provision of a reliable electricity supply [C].

Load prediction is particularly important for EDF because it generates 77% of its electricity from nuclear power plants, which cannot respond rapidly to unforeseen demand. Under-prediction of load leads either to supply failure, or to EDF having to buy in energy at premium prices. Over-prediction leads to unnecessary production and business inefficiencies. The cost to EDF of over-production by 1% for a single day is around £0.5M [C].

EDF have developed methods for electricity grid load forecasting based on the mgcv software and its underlying methods. EDF’s use of GAMs has been built directly on collaboration with Bath on large dataset and autocorrelation issues [D]. The EDF work is, in particular, reliant on the high

degree of numerical reliability provided by the methods developed in Bath [1, 2], on the ability to handle large datasets [D], and on the handling of interaction terms [3].

An EDF representative [E] states that the Bath mgcv work *"has had a number of concrete and important impacts on our work at EDF... These are commercially important for EDF, both in terms of complying with the requirements of the national grid management bodies, and of matching supply to demand in an economically and environmentally efficient manner".* He goes on to list several specific areas:

"*1. The methods encoded in mgcv are used to discover and investigate new effects... A number of such effects have subsequently been incorporated in the parametric models currently used for operational forecasting.*

2. The methods have been successfully employed in pilot studies on EDF subsidiary companies, and are currently being further developed for operational forecasting purposes for these companies.

3. The methods have been used operationally on the French national grid as a tool to help operators when special meteorological events happen (extreme absolute temperatures or rapid temperature variations, for example). In these cases the GAM based models capture the electricity grid load dynamics better than the current operational models, and are used to correct the operational models.

4. EDF uses the methods for forecasting of heat demand for cogeneration plants where it achieves a 20% gain over the existing methods.

5. EDF leads some important research projects around [the methods]. Among them … collaborations with IBM to implement GAM for massive simulation and online forecasting." [E]

Example 2. The methods are widely used in fisheries where they contribute to policy decisions about quota setting, as the following examples illustrate. The enhanced reliability offered by the methods allows CSIRO Tasmania to use GAMs to analyse and design their fisheries independent survey programme, which helps to improve management of the south east of Australian fisheries (estimated annual value AU$700M, 2005/6) [F]. Similarly, models based on [2] and [5] above have been used as part of IFREMER's (French Research Institute for Exploration of the Sea, wwz.ifremer.fr) input to quota setting for the Blue Ling Fishery [G]. The methods have also been used to develop models for catch per unit effort standardization in deep sea fisheries which in turn inform the policy and management advice of ICES (International Council for the Exploration of the Sea), which is used by the EU for quota setting and other management [H]. A separate use of the methods developed models for fish stock indicator indices used by the EU for stock management assessment [J]. To illustrate the breadth of extra academic impact within fisheries, of the 689 publications citing Bath mgcv related papers on Google Scholar from fisheries, 77% had at least one author with an address outside higher education and on average each publication had 1.3 such addresses [K]. The project has been sufficiently successful in making GAM use statistically routine, that many fisheries uses result in no citation [L]. It is the numerical reliability combined with sound smoothness selection methods that has changed practice among many fisheries statisticians involved in assessment so that they now use the Bath/mgcv based methods.

The success of the methods means that they have become part of the `statistical infrastructure', and in combination with their availability as free software, this complicates the process of gathering direct evidence of extra-academic reach. However, indirect evidence is obtainable [K]. By September 2013, there were over 3200 citations to Bath authored mgcv related publications (i.e. to publications from 2006 onwards) on Google Scholar. Approximately 55% of these have at least one author with a non-academic address and the average number of non-academic author addresses per paper is about 0.9. A substantial proportion of these were government institutes charged with natural resource management (fisheries, forestry, agriculture), but there were also private companies, health charities and bodies, Government bodies (e.g. Deutsch Bundesbank) and international bodies (e.g. WHO and UNESCO). Notable topics were fisheries (689), air pollution (425), medicine (730) and energy (391). Further evidence of the extra academic impact of the work is that SAS, *the* major commercial provider of statistical software to industry are currently implementing GAM functionality into their software, based on the Bath work [M], while private

statistics companies run courses in which mgcv is a major component [N].

5. Sources to corroborate the impact

[A] http://prezi.com/s1qrgfm9ko4i/the-r-ecosystem/ based on data from Revolution Analytics (http://www.revolutionanalytics.com/) puts the number of R business users at >1.2M (c.f. ~1M academic). However the figures are very uncertain. A Jan 2009 New York Times online article (http://bits.blogs.nytimes.com/2009/01/08/r-you-ready-for-r/) puts the user figure at 250,000 – 1M. Various other data are at: http://r4stats.com/articles/popularity/

[B] In 2002, problems with an earlier GAM function in S-plus, in the context of air pollution modelling, were reported in the New York Times http://www.nytimes.com/2002/06/05/us/data-revised-on-soot-in-air-and-deaths.html

[C] http://about-us.edf.com/about-us-43666.html provides EDF information.

[D] Wood, Goude and Shaw "Generalized Additive Models for Large Datasets", accepted subject to minor corrections for *Applied Statistics* (JRSSC). Describes collaboration with EDF on grid load problem

[E] Letter from EDF Research Engineer on the use of mgcv methods at EDF.

[F] Peel, Bravington, Kelly, Wood and Knuckey (2012) A Model-Based Approach to Designing a Fishery-Independent Survey. *Journal of Agricultural, Biological and Environmental Statistics* 18(1):1-21 describes application of GAM modelling in design of a fisheries survey for management. http://dx.doi.org/10.1007/s13253-012-0114-x

[G] Email from scientist at IFREMER, describing use of Blue Ling work and referring to official ICES blue ling assessment document (section 3.6 Data analysis: space-time modelling).

[H] Email from same scientist at IFREMER describing CPUE work, with supporting documents.

[J] Email from another IFREMER scientist describing stock indicator work, with supporting documents.

[K] *Report on impact of mgcv project beyond the higher education sector.* University of Bath, internal report.

[L] Example: the FAO Tuna working group papers include a paper by Haritz Arrizabalaga (2009) on "Estimation of Tuna fishing capacity from stock assessment related information" (http://www.fao.org/docrep/012/i1212e/i1212e.pdf). This makes considerable use of mgcv based GAMS, but you can only tell this by looking up the citation to Retrepos, V.R. (2007) "Estimates of large scale purse seine...", (FAO Fish. Proceedings 8:51-62), which also contains no citation, but contains figures clearly plotted from mgcv, making it clear that it is the basis of the GAM analysis.

[M] Email from SAS.

[N] For example http://www.highstat.com/.