Institution: University of Bath		
Unit of Assessment: B10 Mathematical Sciences		
Title of case study: Improving data analysis through statistical methodology and software		
Period when the underpinning research was undertaken: 2006 - 2015		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Simon Wood	Professor	January 2006 – November 2015
Nicole Augustin	Senior Lecturer, previously Lecturer	September 2005 – January 2020
Period when the claimed impact occurred: 2014 - 2020		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact

A generalized additive model (GAM) describes data through the sum of smooth functions of several input variables. Research at University of Bath has substantially improved the estimation and formulation of GAMs and underpins their implementation in the R package mgcv and the SAS statistical software package. Our research has

- increased the scope of applicability of GAMs and facilitated their use through direct development of open source software and influence on commercial software
- driven their impact through uptake in modelling data in a wide variety of areas including farming and UK public health monitoring.

The examples we present include an ag-tech company, Farmers Business Network, using mgcv that demonstrates gains of around USD100,000 per year for each farmer who applies the methods based on our research; a major energy supplier, EDF modelling demand by GAMs; and Office for National Statistics in their reporting of COVID-19.

2. Underpinning research

The underpinning research was undertaken by Simon Wood (Professor at Bath from 2006 to 2015) and Nicole Augustin (Senior Lecturer at Bath between 2005 and 2020). The aim of the research programme is to make the use of generalized additive models (GAMs) as reliable and routine as the use of generalized linear models has long been, in order that these flexible statistical models can be used as standard practice beyond academic statistics.

A GAM relates a random response variable to a sum of smooth functions of one or more predictor variables. These functions are 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 to do 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 and 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. *Inference for GAMs.* The distinctive nature of GAMs and the data dependent smoothness of model components necessitate new approaches to inference. We have developed methods to compute p-values [3] and the Akaike Information Criterion [4] for GAMS.

3. *GAMs for very large data.* We have developed methods [5] to fit GAMs to very large data sets, with the order of 10^4 coefficients and up to 10^8 data points.

4. *High quality software implementing the methods.* The mgcv package [A], written at Bath, is supplied with R as the default method for generalized additive modelling. Moreover, mgcv has the functionality (through mgcv::jagam) to link to the packages JAGS and STAN for Bayesian analysis of GAMs [6].

3. References to the research

[1] Wood, SN 2008, 'Fast stable direct fitting and smoothness selection for generalized additive models', *Journal of the Royal Statistical Society, Series B (Statistical Methodology),* vol. 70, no. 3, pp. 495-518. https://doi.org/10.1111/j.1467-9868.2007.00646.x

[2] Wood, SN 2011, 'Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models', *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 73, no. 1, pp. 3-36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

[3] Wood, SN 2013, 'On p-values for smooth components of an extended generalized additive model', *Biometrika*, vol. 100, no. 1, pp. 221-228. https://doi.org/10.1093/biomet/ass048

[4] Wood, SN, Pya, N & Säfken, B 2016, 'Smoothing Parameter and Model Selection for General Smooth Models', *Journal of the American Statistical Association*, vol. 111, no. 516, pp. 1548-1563. https://doi.org/10.1080/01621459.2016.1180986 (submitted 2015)

[5] Wood, SN, Li, Z, Shaddick, G & Augustin, NH 2017, 'Generalized additive models for gigadata: modelling the UK black smoke network daily data', *Journal of the American Statistical Association*, vol. 112, no. 519, pp. 1199-1210. https://doi.org/10.1080/01621459.2016.1195744 (submitted 2015)

[6] Wood, SN 2016, 'Just another Gibbs additive Modeler: Interfacing JAGS and mgcv', *Journal of Statistical Software*, vol. 75, no. 1, pp. 1-15. https://doi.org/10.18637/jss.v075.i07 (submitted 2014)

The submission dates for papers [4], [5] and [6] and also the presence of Augustin as an author of [5] show that this research was conducted at Bath.

4. Details of the impact

By providing numerical methods and software implementations, we have facilitated a wide uptake of generalized additive models (GAMs) in a broad range of applications.

(i) Implementation of GAMs in R and SAS

The R package mgcv [A], written at Bath, enables users to fit GAMs with automatic estimation of smoothing parameters. Furthermore, mgcv facilitates Bayes analysis of a GAM through the function mgcv::jagam which generates the JAGS model code for inference about the model via Gibbs sampling. The methods reach a vast audience through R and their importance is recognised by the inclusion of mgcv as one of only a few recommended packages (out of thousands) supplied with all versions of the R statistical computing environment.

In 2015, SAS enhanced their statistical software package by incorporating methods to fit GAMs as their procedure PROC GAMPL. SAS documentation states that the “Outer iteration” method used in model fitting [B, p. 101] is an implementation of Wood’s work [1 and 2], while tests for the presence of smooth component [B, p. 104] are based on results in [3].

The R&D Senior Manager at SAS writes [C]:

"SAS is a major provider of statistical software globally, with software installed in more than 83,000 business, government and university sites in 147 countries. In many industries, for example pharmaceuticals, it sets the standard for statistical software. … industrial users began to request generalized additive modeling functionality, based on the penalized regression spline approach of Wood, as implemented in the R package mgcv. … In response to this demand, SAS/STAT® released PROC GAMPL in 2015 … making extensive use of the methods in Wood (2008, 2011, …) [1, 2]. Based on the same technology, SAS Viya® also released PROC GAMMOD".

The letter gives examples of the pervasiveness of GAM methodology [C]:

" … a user in the insurance industry is using PROC GAMPL to obtain better models to determine insurance premium (Frigo and Osterloo 2016) [D]. The mortgage industry also uses generalized additive models to analyze mortgage claim rates (Rodriguez and Cai 2018) [E]".

and gives the emphatic endorsement [C]:

"We are confident that the technology based on Simon Wood's 2006 book and subsequent papers is essential to every data analyst and modeler".

(ii) Example 1: EDF

EDF is a French multinational company and the world's third largest electricity provider, serving over 5,000,000 customers in the UK alone. Accurate predictions of demand for gas and electricity are crucial to EDF since it generates a high proportion of its electricity from nuclear power plants, which cannot respond rapidly to unforeseen demand: under-prediction of load leads to supply failure, or to EDF having to buy in energy at premium prices; over-prediction leads to unnecessary production and business inefficiencies. Modelling demand by GAMs has become an integral part of EDF's strategy.

The Head of the OSIRIS Department of EDF R&D writes [F]:

"I am writing this letter on behalf of EDF R&D to summarize our use of the methods developed by the team of Professor Simon Wood … As a major electricity and gas provider, EDF faces many statistical and machine learning problems, particularly in the field of demand modelling and forecasting. … In collaboration with Professor Wood, we have developed the proper methodologies to solve our industrial problems … Since 2014, this work has produced significant improvements in the models used in EDF commercial operations.

GAMs and their mgcv implementation are used for EDF's national real time electricity and gas demand forecasting. … Recent advances such as big additive model fitting (Wood et al, JASA, 2017) [5], multi-resolution effects and recent Generalized Extreme Values model families for peak forecasting (Wood, Pya and Saefken, JASA, 2016) [4] are used by EDF R&D to deliver these operational models in a context of constant evolution of demand/production".

(iii) Example 2: Farmers Business Network

Another example of the use of GAMs is in tools for optimizing decision making and for risk management in farming provided by Farmers Business Network (FBN), an ag-tech SME with a network of 16,000 farmers with 45,000,000 acres of land across the US, Canada and Australia [G1].

The Data Scientist at the FBM has written [G2] about two FBN projects referred to as "Fred" and "Wilma". He describes the decision optimization tool, Fred, noting that while the earlier Fred v2 used to take around 40 hours to run its computations, in the recently developed Fred v3 the same process takes less than 10 minutes. The letter [G2] enlarges on this point:

"Thanks in part to the discretization methods invented and implemented by Prof. Wood, the entire model-fitting process now takes less than 10 minutes … Furthermore, thanks to

some unique features of GAMs, Fred v3 has a much greater ability to generalize from the training set, allowing it to make accurate predictions …

How much could these improvements help with a farmer's bottom line? ... one way to measure this is to ask how much a farmer would have gained if, between two choices that they made in different parts of their operation, they picked the one that we predicted to be the superior choice. According to that measure, mgcv-based Fred v3 uncovered $23 dollars [USD23] per acre in additional revenues over Fred v2 on average, for a total of about $80,000 [USD80,000] per year for an average 3,500 acre farm".

Regarding the risk-management tool, Wilma, which uses mgcv, [G2] states:

"Wilma delivers the highest quality risk-management in its target market at a lower cost than any competitors can provide. In some areas, we have customers who could save almost $9 [USD9] per acre on the highest quality product, or about $30,000 [USD30,000] per year for a typical FBN farm".

The impact of FBN's projects on farmers in the USA, Canada and Australia is clear when the additional revenues of USD80,000 per year and savings of USD30,000 per year are ascribed to even a small percentage of the FBNs 16,000 farmers.

(iv) Example 3: Office for National Statistics Reports on Covid19

The Office for National Statistics (ONS) has published a weekly "Coronavirus (COVID-19) Infection Survey, UK" since July 2020 [H1]. This survey reports the percentage of people testing positive for COVID-19 and the data are analysed by region, by age, and in the most recent reports by the proportion of cases with new variant COVID-19. The analysis is key to understanding trends over time and has been shared with the UK Government and Scientific Advisory Group for Emergencies (SAGE) to inform decisions on COVID alerts and the relaxation of restrictions.

The ONS report [H1] refers to ONS report [H2], which in turn refers to the article [H3] where the data analysis methods are described. The R package rstanarm was used to fit a Bayesian multilevel GAM with age, sex, time and region as covariates. This package calls the function mgcv::jagam [6] in carrying out its analyses.

(v) The widespread use of GAMs

The relevance of GAMs along with the availability of mgcv in R and the GAMPL and GAMMOD Procedures in SAS means these methods have become part of the "statistical infrastructure".

To investigate the use of GAMs in applied data analysis and modelling, we carried out a survey [I] of publications citing Wood's papers with publication dates between 2006 and 2016 that are referenced in the mgcv manual [A] (these include references [1] to [4] and [6]). Google Scholar reported 13,100 such citations between 2014 and February 2020 and, of these, 2,700 citations related to fisheries, 1,370 to air pollution, 2,300 to medicine, 1,440 to power production and usage, and 600 to finance.

In a sample of 172 citing articles, 43% had at least one author with an address outside academia, at: institutes charged with natural resource management (e.g., Institute of Marine Research, Norway; International Seafood Sustainability Foundation, Washington DC; EC Institute for Environment and Sustainability, Italy); private companies (e.g., MRAG, London; SOSECALI, Ecuador); health charities and bodies (e.g., Pan American Health Organisation; National Cancer Center, Japan; Public Health Service, Reggio Emilia, Italy); and international bodies (e.g., World Bank; World Health Organisation; EcoHealth Alliance).

5. Sources to corroborate the impact

[A] The package 'mgcv', 27 August 2020.

https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

See also the R web page http://cran.r-project.org/web/packages/mgcv/index.html

[B] SAS/STAT® 14.2 User's Guide for High Performance Procedures (2016), Chapter 4. https://support.sas.com/documentation/onlinedoc/stat/142/stathpug.pdf

[C] Testimonial Letter from R&D Senior Manager, SAS, 5 January 2021.

[D] *"exSPLINE That: Explaining Geographic Variation in Insurance Pricing"* Frigo and Osterloo, 2016, Proceedings of SAS Global Forum 2016.

[E] *"Regression Model Building for Large, Complex Data with SAS® Viya® Procedures"* Rodriguez and Cai, 2018. For Youtube video presentation see: https://www.youtube.com/watch?v=gNic_MFOPsw

[F] Testimonial Letter from Head of the OSIRIS Department, EDF R&D, 9 February 2021.

[G] Farmers Business Network evidence

[G1] Farmers Business Network website, accessed 20 January 2021. https://www.fbn.com/

[G2] Testimonial Letter from Data Scientist, Farmers Business Network, 10 July 2020.

[H] Office of National Statistics COVID-19 Infections Survey evidence

[H1] *"Coronavirus (COVID-19) Infection Survey, UK: 24 December 2020",* Office for National Statistics https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/24december2020

[H2] *"COVID-19 Infection Survey (Pilot): methods and further information"*, Office for National Statistics, 21 September 2020.

https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/methodologies/covid19infectionsurveypilotmethodsandfurtherinformation

[H3] *"Community prevalence of SARS-CoV-2 in England: Results from the ONS Coronavirus Infection Survey Pilot",* K. B. Pouwels et al., 2020, https://www.medrxiv.org/content/10.1101/2020.07.06.20147348v1.full.pdf

[I] Report on impact of mgcv project beyond the higher education sector. University of Bath, internal report, March 2020.