CBCS
HTTP://AGROBIOL.SGGW.WAW.PL/CBCS

COMMUNICATIONS IN BIOMETRY AND CROP SCIENCE
VOL. 10, NO. 1, 2015, PP. 41–57

INTERNATIONAL JOURNAL OF THE FACULTY OF AGRICULTURE AND BIOLOGY,
WARSAW UNIVERSITY OF LIFE SCIENCES – SGGW, POLAND

TEACHING CORNER

A guide to generalized additive models in crop science using SAS and R

Josefine Liew [1,2], Johannes Forkman [1*]

[1] Swedish University of Agricultural Sciences, Department of Crop Production Ecology, Box 7043, SE-750 07 Uppsala, Sweden.
[2] Valiguard AB, SAI Global Scandinavia, Box 5609, SE-114 86 Stockholm, Sweden.
*Corresponding author: Johannes Forkman; E-mail: johannes.forkman@slu.se

CITATION: Liew J., Forkman J. (2015). A guide to generalized additive models in crop science using SAS and R. *Communications in Biometry and Crop Science* 10, 41–57.

Received: 4 December 2014, Accepted: 14 April 2015, Published online: 17 June 2015

ABSTRACT
Linear models and generalized linear models are well known and are used extensively in crop science. Generalized additive models (GAMs) are less well known. GAMs extend generalized linear models through inclusion of smoothing functions of explanatory variables, e.g., spline functions, allowing the curves to bend to better describe the observed data. This article provides an introduction to GAMs in the context of crop science experiments. This is exemplified using a dataset consisting of four populations of perennial sow-thistle (*Sonchus arvensis* L.), originating from two regions, for which emergence of shoots over time was compared.

Key Words: *Generalized additive mode; polynomial regression; Sonchus arvensis; GAM.*

INTRODUCTION

Generalized additive models (GAMs), introduced by Hastie and Tibshirani (1986), relax and extend generalized linear models. GAMs are useful for applications where polynomials fail to describe the observed curvature in the data because they include splines or local regression smoothing (LOESS) functions (Cleveland 1979). The output of a GAM analysis is mainly graphical; the explicit resulting smoothing function is complicated since it is not a function of any easily interpreted parameters (Venables and Dichmont 2004). Although primarily used for exploratory purposes, GAMs can also be used for approximate statistical inference, e.g., hypothesis testing. GAMs are said to be non-parametric (Yee and Mitchell 1991) or semi-parametric (Guisan et al. 2002), which refers here to the lack of a particular functional form of the relationship between the dependent variable Y and the explanatory variable X. GAMs have been available in some software since the early 1990s (Hastie 1992, Hilbe 1993), are available in, for example, the SAS System (SAS Institute 2008) and the open source environment R (Wood 2006a).

GAMs are rarely used in agronomy, although they could be useful in various applications. This article offers a starting point for further study of GAMs for readers who have basic knowledge of statistics as a complement to other fields of research. We specifically describe how GAMs can be used for comparing effects of experimental treatments, since this information is often lacking in other reviews. An example illustrates how GAMs can be fitted using the SAS System and R. Fitting GAMs is compared with fitting polynomials.

GAMs can be explained with reference to linear models (LMs), generalized linear models (GLMs), and additive models (AMs). The LM, with a single response variable, y, and p explanatory variables takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e \,, \quad (1)$$

where β_0 is the intercept and $\beta_1,\ \beta_2, \ldots, \beta_p$ are the regression coefficients, i.e., the slopes of the explanatory variables. In many applications, there is only a single explanatory variable, for example time. In this case, Eq. 1 reduces to a simple linear regression model $y = \beta_0 + \beta_1 x_1 + e$. For statistical inference, it is assumed that the error terms, e, are (i) normally distributed with expected value 0; (ii) have the same variance; and (iii) are independent. It is also assumed that (iv) the values $x_1,\ x_2, \ldots, x_p$ are fixed, i.e., have no variance, and are known.

In LMs, the mean, μ, can be written as: $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$, i.e., it is a direct linear function of the parameters. In GLMs, a transformation, $g(\mu)$, of the mean is a linear function of the parameters. Thus, a transformation of the mean is modeled instead of the observations y. GLMs extend LMs to link functions other than the identity. Assumptions (i) and (ii) for LMs are relaxed. GLMs consist of three components: the random component, which is the response variable and its probability distribution; the systematic component, which is a linear function of the explanatory variables; and the link function, which links the random and the systematic components. With p explanatory variables, a GLM can be written as:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \,, \quad (2)$$

where $g(\mu)$ is the link function and $\beta_0,\ \beta_1, \ldots, \beta_p$ are the parameters to be estimated. The log link, $g(\mu) = \log \mu$, and the logit link, $g(\mu) = \log[\mu/(1-\mu)]$, are common links and are usually chosen when observations are Poisson and binomially distributed, respectively. GLMs can be used for any distribution from the exponential family of distributions, which also includes for example the gamma and negative binomial distributions (McCullagh and Nelder 1989).

In AMs, the terms $\beta_1 x_1,\ \beta_2 x_2, \ldots \beta_p x_p$ in Eq. 1 are replaced by smoothing functions $s_1(x_1),\ s_2(x_2), \ldots, s_p(x_p)$, which can be splines or LOESS functions. These are smooth curves that estimate the functional relationship in noisy data. The AM can be written as:

$$y = \beta_0 + s_1(x_1) + s_2(x_2) + \ldots + s_p(x_p) + e \,. \quad (3)$$

Combination of GLMs and AMs gives a GAM:

$$g(\mu) = \beta_0 + s_1(x_1) + s_2(x_2) + \ldots + s_p(x_p) \,. \quad (4)$$

In contrast to AMs, GAMs are not restricted to the normal distribution, but can be used for any probability distribution from the exponential family. Just as with the other models, GAMs require explanatory variables to be measured without errors. Additionally, high correlation of explanatory variables causes inferential problems because individual effects of the explanatory variables cannot be separated.

SMOOTHING FUNCTIONS

LMs partition data into "model + error", while smoothing functions separate data into "smooth + rough" and strive to minimize the rough part as much as possible (Hastie and Tibshirani 1990). Many types of smoothing functions are available, but all rely on the same principles:

- Each observation in the dataset is predicted from a regression model fitted to the surrounding observations.
- The resulting graph, i.e., the curve of the smoothing function, is smooth.
- The smoothness of the curve is controlled by a smoothing parameter λ.

LOESS functions work as follows: For each value x_i of the explanatory variable X, a neighborhood (also called window, band or span) is defined. Within these neighborhoods, low-order polynomials are fitted using weighted regression. First- or second-order polynomials are generally used. To emphasize that weighted regression is used rather than ordinary regression, LOESS is sometimes called LOWESS. Values close to x_i are weighed higher than values far from x_i. This fitting process may be repeated several times, during which observations with large residuals are down-weighted. Given the polynomial degree, the smoothness is controlled by the band width, which functions as the smoothing parameter. The band width is usually defined as the proportion of the data used in the local regression fit. For example, this proportion may be set to $\lambda = 0.25$, indicating that 25% of the data are used in each local regression. Values larger than $\lambda = 0.25$ are often preferred, but the choice depends on the data. A higher value of the smoothing parameter λ gives a smoother fit than a lower value. In practice, the smoothing parameter is frequently determined by trial and error based on visual inspection of the fitted curve compared with the observations. Undesirable patterns, as results of inadequate fits, can be detected in plots of residuals against fitted values. The obstacle to straightforward application of GAMs is that the same set of data can give rise to different interpretations because the result will depend on the degree of smoothing, which is chosen by the analyst.

Cubic splines are third-degree polynomials that are fitted in adjacent neighborhoods connected at so-called knots. At these knots, the polynomials are constrained to have the same derivatives, so that the polynomials join smoothly. The degree of smoothness can be controlled by the number of knots, with a large number of knots giving a smoother curve than a small number. The knots need not be equally spaced on the x-axis. In intervals with many knots, the curve becomes more flexible than in intervals with few knots. Cubic splines can be chosen on the basis of visual inspection of fitted curves and residuals.

Nowadays, cubic splines are usually fitted by optimizing a penalized least squares criterion. In the case of a single explanatory variable, this criterion, which should be minimized, can be written as:

$$\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda \int (s''(x_i))^2 dx \,. \qquad (5)$$

Here, $s''(x_i)$ is the second order derivative of the smoothing function s. The penalized least squares criterion (Eq. 5) consists of two terms. The first one measures closeness to observations while the second one penalizes high curvature, giving the curve a smooth appearance. If $\lambda = 0$, then Eq. 5 reduces to the standard least squares criterion, according to which the smoothing function, $s(x)$, should be chosen such that the error sum of squares is minimized. It would result in a highly curvaceous graph that joins all observations. Since a smooth curve is preferred, a positive λ value is always used. The second-order derivative in Eq. 5, which measures curvature, is squared and integrated over the x-axis. Thus, a high value of the smoothing parameter λ penalizes curvature more than a lower value and results in a smoother curve. At one extreme, i.e., when $\lambda \to \infty$, the penalty term dominates, so that the graph of the optimal function is a straight line. At the other extreme, i.e., when $\lambda \to 0$, the penalty term becomes unimportant and the optimal curve wriggles up and down, tracing the observations.

The smoothing parameter λ is usually determined indirectly through the choice of effective degrees of freedom, which can be thought of as the "tune" that determines curviness (Jones and Almond 1992). The number of effective degrees of freedom is analogous

to the number of degrees of freedom of a LM, which is the number of linear constraints or, for the error, the difference between the number of observations and the number of linear constraints. Most computer software has options for specifying the preferred number of effective degrees of freedom. The more effective degrees of freedom the smoothing function uses, the rougher and more complex the curve becomes. The choice of the appropriate level of smoothing, by specifying the effective degrees of freedom, is among the most crucial steps in fitting GAMs. Smoothness and fit must be balanced. When the model includes several smoothing functions, they can have varying numbers of effective degrees of freedom.

The number of effective degrees of freedom can also be chosen automatically, using cross validation. Through cross validation, points (x_i, y_i) are left out, one at a time, and the smoothing function at x_i is estimated based on the remaining $n-1$ data points. Cross validation sum of squares for the smoothing parameter λ is computed for a range of λ. The preferred estimate of λ is that which minimizes this sum of squares. However, cross validation may be time consuming, even with fast computers. Instead, a weighted version of the full cross validation procedure may be used (SAS Institute 2008). The performance of automatic procedures for choosing the smoothing parameter is sometimes poor, which in many cases is evident, taking into account the biological interpretation when inspecting the resulting plots. This will be further described in the sections "How to assess model fit" and "Example". In practice, a smoothing parameter is commonly chosen so that the number of effective degrees of freedom is around 3–5.

Smoothing functions are easily fitted for models with one or two explanatory variables. However, models with several explanatory variables, sometimes in interaction (e.g. the effect of the dose of a herbicide may depend on the seed rate of the crop), are common. Smoothing is possible in such cases too. Consider a model with p explanatory variables and one response variable. If $p = 1$, smoothing is done in two dimensions, plotting the response variable against the explanatory variable. If $p = 2$, there are three dimensions, and the graph of the smoothing function is a two-dimensional curved surface. If $p > 2$, it is difficult to visualize the graph. Variants of surface smoothers have been developed for $p > 1$ explanatory variables. The thin-plate regression spline (Wood 2003) is a p-dimensional version of the cubic spline fitted through penalized least squares. Thin-plate regression splines are isotropic, i.e., they are not affected if the co-ordinate system of the explanatory variables is rotated. The assumption of isotropy (uniformity in all directions) is reasonable when the variables are on the same scale, for example when two explanatory variables measure geographical location in two dimensions. However, isotropic smoothing functions are less suitable when the explanatory variables measure different things, for example when one variable measures distance and the other time. Using an isotropic smoothing function, the result will depend on the units chosen, for example whether the explanatory variables are measured in hours and meters, or weeks and inches (Wood 2006a). In such situations , anisotropic smoothing functions may be preferable. In particular, tensor-product smoothing functions (Wood 2006b), constructed from univariate smoothing functions, are frequently used.

STEPS IN GENERALIZED ADDITIVE MODELLING

HOW TO SPECIFY MODELS

For controlled experiments, pure GAMs (such as $g(\mu) = \beta_0 + s(x)$ for a single explanatory variable) are rarely adequate. Since experiments usually aim at comparing treatments, treatment factors must be included in the model. A GAM will then comprise more parameters and several non-parametric smoothing functions.

If $g(\mu)$ as a function of x is generally increasing or decreasing in x, not as a perfect straight line, but as a bending curve around a straight line, this can be modeled as a combination of a

straight line and a smoothing function. With a single explanatory variable, this model is written as:

$$g(\mu) = \beta_0 + \beta_1 x + s(x) . \quad (6)$$

Here, the parametric linear effect $\beta_1 x$ is separated from the effect of the smoothing function $s(x)$. This makes it possible to test the null hypothesis that the model is linear against the alternative that the model is non-linear. By definition, the smoothing function is centered around 0, so the resulting flexible curve becomes centered around the estimated line. If the term $\beta_1 x$ were omitted from Eq. 6, the smoothing function $s(x)$ would describe the sum of all linear and non-linear effects of x. Higher-order polynomials, for example a quadratic curve or a cubic curve, can be fitted combined with a smoothing function only when the dataset is large.

Categorical explanatory variables can be included as factors in GAMs. It is convenient to use dummy variables for coding the factor levels. Thus, let d_j denote a dummy variable that takes the value 1 when the observation belongs to the *j*th level of the factor, and 0 otherwise. For an experiment with two treatments (i.e., a treatment factor with two levels), Eq. 6 can be augmented into

$$g(\mu) = \beta_1 d_1 + \beta_2 d_2 + \beta_3 x + s(x) , \quad (7)$$

which is a model with treatment-specific intercepts and a smooth curve around a straight line. According to Eq. 6, the curve of the first treatment is parallel to the curve of the second treatment. This model is analogous to the linear model for analysis of covariance assuming parallel slopes. Eq. 7 can easily be extended to a model for experiments with $J > 2$ treatments, through the use of J dummy variables and $J + 1$ parameters: $\beta_1, \beta_2, \ldots, \beta_{J+1}$. It is usually not necessary to construct a data set with dummy variables, because statistical software packages do this work automatically if informed that the explanatory variable is a factor. However, for interpretation of parameter estimates it is essential to know exactly how the specific statistical software package codes the factor levels.

A model with two treatment-specific curves, i.e., a model with two curves that are not parallel, can be specified as:

$$g(\mu) = \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_1 x + \beta_4 d_2 x + s_1(x) d_1 + s_2(x) d_2 . \quad (8)$$

Eq. 8 describes a varying-coefficient model (Hastie and Tibshirani 1993). This model is in effect similar to an analysis of covariance model with dissimilar slopes. Note that the smoothing functions $s_1$and s_2 should be functions of x, and not of the products xd_1 and xd_2.This is because when d_1 is zero, the observation should not be included in the fit of the smoothing function s_1 (and similar, when d_2 is zero, the observation should not be included in the fit of s_2).

We can use two types of models with two continuous explanatory variables. The first uses two separate smoothing functions for each variable, as here:

$$g(\mu) = \beta_0 + s_1(x_1) + s_2(x_2) , \quad (9)$$

where x_1 and x_2 are two continuous explanatory variables, for example longitude and latitude. The second one uses a single smoothing function for both variables:

$$g(\mu) = \beta_0 + s(x_1, x_2) . \quad (10)$$

While Eq. 9 does not allow for interaction between x_1 and x_2, Eq. 10 does. Such interaction means that the effect of one variable depends on the value of the other. For example, grain yield of a certain wheat variety depends on nitrogen fertilization, but also on the available soil water content: since water is needed for uptake of nutrients, the effect of fertilization depends on the soil water content. So, nitrogen fertilization interacts with soil water content in terms of affecting grain yield. Eqs 9 and 10 can easily be extended to any number of continuous explanatory variables. For models including many variables, however, fitting may be difficult.

In agricultural science, it is common to perform comparative experiments, in which treatments (e.g. soil preparation methods) are compared with each other. The treatments can be represented in the model through the use of dummy variables. The experiment might also

include some continuous explanatory variables (e.g. clay and soil water content) that could possibly interact with each other. One of the scientific questions might be whether a smoothing function of these two continuous variables is different for the two soil preparation methods. Using dummy variables d_1 and d_2 for the two soil preparation methods, the appropriate statistical model is:

$$g(\mu) = \beta_1 d_1 + \beta_2 d_2 + s_1(x_1, x_2)d_1 + s_2(x_1, x_2)d_2. \quad (11)$$

Under the null hypothesis that there is no difference between the two treatments in the smoothing functions s_1 and s_2, Eq. 11 reduces to Eq. 10. If the experiment compares t treatments, then the model from Eq. 11 will be expanded to include 2 t terms.

HOW TO FIT MODELS

> "*The aim in a good smooth is to capture fully the underlying trend, while not paying too much attention to underlying noise.*" (Jones and Almond 1992, p. 437)

Shipley and Hunt (1996) compared LOESS and cubic spline smoothing functions. They concluded that both methods "represent versatile and accurate ways of form-free curve fitting" that "capture main trends". In their study, LOESS gave slightly better results. However, nowadays cubic spline smoothing functions are more popular, since their accessibility and flexibility make them "the best option" (Zuur et al. 2009).

The *gam* procedure of the SAS System and the *gam* function of the *gam* package in R fit LOESS and cubic spline smoothing functions, using the backfitting algorithm (Hastie and Tibshirani 1986, 1990). The *gam* function of the *mgcv* package in R fits cubic spline smoothing functions using direct penalized likelihood maximization (Wood 2006). With the *gam* procedure of SAS and the *gam* function in the *mgcv* package in R, the number of effective degrees of freedom can be determined through cross validation. The *gam* function of the *mgcv* package can use cross validation when the scale parameter is unknown, and a procedure that minimizes an estimate of the expected mean square error when the scale parameter is known (Wood 2006a). This function can also fit the smoothing function through maximum likelihood or restricted maximum likelihood.

Treatment-specific curves of varying-coefficient models can be fitted one at a time. Thus, to fit Eq. 8, $g(\mu)$ is divided into two parts: $g(\mu) = g_1(\mu) + g_2(\mu)$, where:

$$g_1(\mu) = \beta_1 d_1 + \beta_3 d_1 x + s_1(x)d_1 \; ; \quad (12)$$

$$g_2(\mu) = \beta_2 d_2 + \beta_4 d_2 x + s_2(x)d_2 \; . \quad (13)$$

Eqs. 12 and 13 can be fitted to the observations of the first and second treatment, respectively. The varying-coefficient model can be divided into as many parts as the treatment has levels. Thus, with t levels, also t smoothing functions ($s_1, s_2, \ldots, s_t$) must be fitted. These smoothing functions might need different degrees of smoothness. Users of the *mgcv* package in R can conveniently use the *by* option of the *gam* function for fitting a varying-coefficient model in a single step. In the *gam* procedure of SAS, isotropic thin-plate splines ("*spline2*") can be fitted for $p = 2$ explanatory variables. In the *gam* function of the *mgcv* package in R, isotropic thin-plate regression splines and anisotropic tensor product smoothing functions can be fitted for $p > 1$ explanatory variables (using functions *s* and *te*, respectively). Note that it might not be necessary to fit smoothing functions for all of the model's explanatory variables. Some of the explanatory variables can be included in parametric terms while some other in the smoothing functions. Since we prefer simple models than complex ones when the fits are approximately the same, the number of smoothing functions should as a rule be kept as small as possible.

HOW TO ASSESS MODEL FIT

In LMs and GLMs, parameter estimates constitute a concise description of the fitted linear function. But in GAMs, the fitted curve cannot be described by a small set of

parameter estimates. Instead, it is described graphically by plotting the fitted values against the values of the explanatory variable. To show how well the curve fits the observed values, they can be included in such a plot. It is easy for models with one explanatory variable, but is more complicated for models with several explanatory variables. Partial residuals are useful for investigation of explanatory variables (Guisan et al. 2002, Wood 2006). The partial residual is the sum of the residual for the whole model and the fitted smoothing function for the explanatory variable of interest. In partial residual plots, partial residuals are plotted against the values of the explanatory variable. For example, the partial residual plot for the first explanatory variable x_1 is a plot of partial residuals, i.e., $s(x_1) + e$, against x_1. Such a diagram illustrates the relationship between the dependent variable and the explanatory variable, similarly to a scatter plot. When the fit is good, partial residuals are randomly distributed around the curve that is described by $s(x_1)$.

Deviance (D) is twice the difference between the log likelihood ($\log L$) of the saturated model, which corresponds to a perfect fit, and the log likelihood of the fitted model. Deviance can be viewed as a generalization of sum of squares, since for a normal distribution, deviance equals the error sum of squares. For varying-coefficient models such as Eq. 8, deviance can be computed as the sum of the deviances of the treatment-specific parts. For example, the deviance of the fit of Eq. 8 is the sum of the deviance of the fits of Eqs. 12 and 13.

For GAMs, as for GLMs, it is possible to determine approximately whether a particular explanatory variable is significant or not, by comparing deviances (Hastie and Tibshirani 1990). The difference in deviance between a model omitting the explanatory variable to be tested (i.e., the "reduced model") and a model including the explanatory variable to be tested (i.e., the "full model") is compared with a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between the two models. Thus:

$$\chi^2 = D(Reduced) - D(Full) \quad (14)$$

is approximately chi-square distributed with degrees of freedom equal to the difference between the degrees of freedom (i.e., the sum of effective degrees of freedom and the number of all regression coefficients) of the full and the reduced model, i.e., $df(Full) - df(Reduced)$. Wood (2006) justified the use of this method for penalized regression spline smoothing functions. When performing these tests, the two models should be nested (i.e., the terms of one model should be a subset of the terms of the other model). Consequently, to test for interaction between two continuous explanatory variables, the model described by Eq. 10 cannot be directly compared with the model described by Eq. 9. For hypothesis testing, the model including the interaction term must also include the main effects terms. Thus, the model

$$g(\mu) = \beta_0 + s_1(x_1) + s_2(x_2) + s(x_1, x_2) \quad (15)$$

can be compared with the model specified by Eq. 9. When doing this, one must ensure that the two models are indeed nested. This requirement is satisfied if a tensor product smoothing function is used for the bivariate term in Eq. 15, provided that the marginal smoothing functions of the tensor product smoothing function are of the same type as the univariate smoothing functions in Eq. 9 (Wood 2006).

OVERDISPERSION

Overdispersion occurs when the variance in the observations is greater than expected according to the assumed distribution (Olsson 2002). For Poisson distribution, variance theoretically equals mean, μ, while for Binomial distribution, variance theoretically equals $n\mu(1-\mu)$, where n is the number of independent Bernoulli (i.e., yes/no) trials. In practice, deviance or sum of squared Pearson residuals should be compared with a chi-square distribution with degrees of freedom equal to the number of error degrees of freedom. Usually, these quantities agree as to whether overdispersion is present or not.

The most used methods for taking overdispersion into account are the same as for GLMs (Thurston et al. 2000). In the quasi-likelihood method for Poisson and binomial data, the dispersion parameter ϕ, which for these distributions theoretically equals 1, is estimated from the data. The dispersion parameter can be estimated as either the deviance divided by the number of degrees of freedom for error or the sum of squared Pearson residuals divided by the number of degrees of freedom for error (McCullagh and Nelder 1989, p. 328). Statistical tests are adjusted using the estimated dispersion parameter. If $\phi^\star$ is the estimator of the dispersion parameter ϕ, then the statistic

$$F = \frac{D(Reduced) - D(Full)}{[df(Full) - df(Reduced)]\phi^\star} \quad (16)$$

can be compared with an F-distribution. In the numerator, the number of degrees of freedom equals $df(Full) - df(Reduced)$, as previously defined. In the denominator, the number of degrees of freedom equals the number of degrees of freedom for error of the full model. In a GAM, the number of degrees of freedom for error equals the number of observations minus the sum of the number of regression coefficients and the number of effective degrees of freedom for the whole model. Zuur et al. (2007) suggest F-tests for comparing models when overdispersion is present and provide an example of the use of a quasi-Poisson GAM.

The quasi-likelihood method cannot correct a pattern of residuals (Thurston et al. 2000). When a plot of deviance residuals against fitted values indicates that variation increases with the mean, a negative binomial distribution can be tried instead of the Poisson distribution.

BAYESIAN METHODS

Bayesian methods for smoothing spline functions were introduced by Wahba (1983) and Silverman (1985), and are commonly used for assessing uncertainty in fitted GAM curves. The penalty term of the penalized least squares criterion (Eq. 5) is used as the Bayesian prior distribution, but with a minus sign in front of it so that that smooth curves are believed to be more likely than wiggly curves. Given this prior distribution for the model and the distribution for the data, the Bayesian posterior distribution for the model is the penalized least squares criterion (Eq. 5). The fitted smoothing spline function is the posterior mean in this Bayesian formulation. The confidence in the fitted curve can be expressed by Bayesian credible bands, which are Bayesian analogs to confidence bands, plotted around the fitted curve (Wood 2006a). This option is provided by the *gam* procedure of the SAS System and the *gam* function of the *mgcv* package in R. Marra and Wood (2012) studied the performance of this Bayesian method and showed, through simulation, that Bayesian credible bands have coverage close to nominal levels. By default, the *gam* function of the *mgcv* package in R uses the Bayesian covariance matrix for hypothesis testing of smoothing functions and parametric terms, but the frequentist covariance matrix is also suitable for testing model terms for equality to zero (Wood 2006).

EXAMPLE

The dataset used in the following example is a sub-sample from a dataset previously used by Andersson et al. (2013) to study the seasonal emergence pattern of defoliated plants of perennial sow-thistle (*Sonchus arvensis* L.) with intact root systems. Similar experiments have previously been analyzed using categorical data analysis (Brandsæter et al. 2010), which models time as a classification factor. The Andersson et al. (2013) dataset consists of 128 observations from two northern populations (N1, N3) and two southern populations (S1, S3) of perennial sow-thistle, made on 12 test dates during one year (2008), as follows:

25 observations from N1 (from only nine test dates)
32 observations from N3
36 observations from S1
35 observations from S3

At each test date, three plants (if available) per population were exhumed, defoliated and placed under forcing conditions in a climate room for four weeks. Then, the numbers of

emerged shoots were counted. The objective was to study (i) whether the emergence of shoots varies over time, (ii) whether the two regions differ in this variation over time, and (iii) whether the four populations differ in this variation. The plants were grown in a split-plot design. The experimental field was divided into three blocks with four rows per block, each row constituting a main plot in a split-plot design. Populations were randomized to main plots (i.e. rows) within blocks, and test dates were randomized to subplots within main plots. The observational unit was plant, and there was no repeated measurement on the same sampling unit. In the climate room, the plants were placed in a completely randomized manner and then moved around during the four weeks. The complete experiment in Andersson et al. (2008) was analysed in SAS version 9.2 (SAS Insitutute, 2008), using the *gam* procedure. In the example provided here, new analyses have been done on the subset of 128 observations (i.e., the dataset) and complemented with analyses in R.

ANALYSES IN SAS

The analyses in SAS were performed in version 9.4 of the SAS System (SAS Institute 2008). To assess the importance of the main plot error variance, a split-plot model was fitted:

$$\log(y_{ijk} + 1) = \alpha + b_i + \beta_j + \gamma_k + \delta_{jk} + a_{ij} + e_{ijk}\ , \qquad (17)$$

where α is an intercept, b_i a random block effect (i = 1, 2, 3), β_j a fixed population effect (j = 1, 2, 3, 4), γ_k a fixed date effect (k = 1, 2, ..., 12), δ_{jk} a fixed effect of the population-by-date interaction, a_{ij} a random main plot effect, and e_{ijk} a random residual error. The random effects and the residual errors were assumed to be independent and normally distributed with expected value zero. This model was fitted in SAS using the *mixed* procedure.

To fit polynomials, we used the *genmod* procedure for generalized linear models. The Poisson distribution was used with log link. To adjust for overdispersion, we used the quasi-likelihood method. The dispersion parameter was estimated as the sum of squared Pearson residuals divided by the number of degrees of freedom for error. We investigated three objective procedures for determining the best number of terms for the polynomials: (i) forward selection, (ii) backward selection, and (iii) the Akaike information criterion. To find the best polynomial model using forward selection, a polynomial of degree zero (i.e., a horizontal straight line) was fitted first. Next, a polynomial of degree one, i.e., a straight line, was fitted. The slope of this line being non-significant, the polynomial of degree zero was selected; otherwise a polynomial of degree two was fitted and the significance of the second-order term of that polynomial was tested. This second-order term being non-significant, the polynomial of degree one was chosen; otherwise a polynomial of degree three was tested. In this way, a new term was added to the polynomial until accepting the model because of a non-significant last term. To find the best polynomial model using backward selection, a polynomial of degree seven was initially fitted. When the *genmod* procedure did not converge or the term of the highest order was not significant according to the adjusted F-test, this term was removed from the model. This step-wise reduction of the model was repeated until the model converged and the term of the highest order was significant. The *genmod* procedure in SAS offers the Akaike criterion, a popular tool for model selection. The Akaike criterion value, which should be as small as possible, is twice the difference between the number of parameters and the log likelihood ($\log L$). To find the best polynomial according to the Akaike criterion, all polynomials of degree smaller than eight were investigated. The selected polynomial had the smallest value of the criterion.

GAMs were also fitted, using the *gam* procedure, with generalized cross-validation and with four effective degrees of freedom. In SAS, it is the default option and the program automatically includes a parameter expressing the linear effect of the explanatory variable, like in Eq. 6. Since this parameter uses one degree of freedom, the smoothing function is fitted with three degrees of freedom. To investigate effects of time, region and population, four models were compared: (i) a model with an intercept (i.e., with a mean only), (ii) a GAM with a single spline term (Eq. 6); (iii) a GAM with two spline terms, one per region (Eq. 8);

and (iv) a GAM with four spline terms, one per population. For these analyses, we used SAS with the default four effective degrees of freedom method. Box 1 presents SAS code for population-specific curves. Due to overdispersion, we needed to employ adjusted approximate F-tests (Eq. 16) for inference. Because the *gam* function in SAS does not provide them, we calculated them in Microsoft Excel 2010, based on the output of the GAM analysis in SAS (Tables 1 and 2). The *p*-values were computed using the *fdist* function in Excel.

Table 1. Deviance output in SAS for n=128 observations analyzed with GAM using smoothing functions with four effective degrees of freedom. The complexity of the models, i.e., the number of degrees of freedom (df), increases downwards in the table.

Model	Deviance	Df
Only intercept	1963.44	1
Time	575.928	5
Time x Region	544.574	10
Time x Population	443.193	20

Table 2. Results of approximate F-tests (Eq. 16) based on the deviance output in Table 1.

	Deviance	df	MS	F-ratio	*p*-value
Only intercept (R)	1387.511	4	346.878	74.08	<0.001
Time (F)	575.928	123	4.682		
Time (R)	31.355	5	6.271	1.36	0.244
Time x Region (F)	544.574	118	4.615		
Time x Region (R)	101.381	10	10.138	2.47	0.011
Time x Population (F)	443.193	108	4.104		
Time (R)	132.737	15	8.849	2.16	0.012
Time x Population (F)	443.193	108	4.104		

The table includes four comparisons of reduced models (R) with full models (F). For the reduced models, the Deviance column displays the difference in deviance (Table 1) between the two fitted models. For full models, the Deviance column displays the deviance of the fitted model. For the reduced models, the degrees of freedom (df) column displays df(full model) – df(reduced model). For the full models, df = n – df(full model), where n = 128. Mean Squares (MS) = Deviance/df, and F-ratio = MS (reduced model)/MS (full model). The F-ratio is compared with an F-distribution with degrees of freedom according to the df column, providing an approximate *p*-value

ANALYSES IN R

The analyses in SAS were complemented and compared with some analyses in version 3.1.1 of R (R Core Team 2012). In R, GAMs were fitted with the *gam* function of the *mgcv* package, both with generalized cross-validation and with four effective degrees of freedom. The results were compared to the same analyses in SAS. For populations N3 and S3, the model according to Eq. 6, with three effective degrees of freedom for the smoothing function, did not fit the data well in R. For this reason, models without any parameters expressing linear functions (available in R, but not in SAS) were fitted in R (i.e., using Eq. 5 with $\beta_1 = 0$). In these analyses, the smoothing function was fitted with four effective degrees of freedom. It was obtained by giving the options k = 5 and *fx* = TRUE in the *s* function, which is used within the *gam* function. Box 1 provides R code for fitting of population-specific curves with four effective degrees of freedom.

RESULTS

In the split-plot analysis, the block variance, the main plot error variance, and the subplot error variance were estimated as 0.0020, 0.0072 and 0.3016, respectively. The main plot error variance was small compared with the subplot error variance.

Figure 1 shows the results of the population-wise analyses to find the best polynomial model in SAS. Using forward selection, polynomials of degree one were chosen for populations N3, S1, and S3. However, using backward selection, polynomials of degree six were needed for populations N3 and S3, and a polynomial of degree four was needed for population S1. Use of forward and backward selection resulted in the same conclusion only for population N1, for which a third-order polynomial fitted the data best. For populations N3 and S3, backward selection gave predictions far from real observations. Obviously, these curves did not fit the data well. The AIC criterion resulted in polynomials of small orders: degree one for population N1, degree two for populations N3 and S3, and degree zero for population S1. Because of the logarithmic link function, the polynomials of degree one did not look like straight lines when presented on the original scale (Figure 1). Figure 2 shows the GAM fits using the generalized cross-validation methods of SAS and R. The two software packages produced the same curve only for population N1. For the other populations, SAS gave curves with much larger fluctuations than R. Curves fitted with generalized cross-validation in SAS failed to represent the biological phenomena in the studied example. The generalized cross-validation in SAS resulted in curves that showed pronounced peaks in areas without any observations.

Box 1. R and SAS code for fitting population-specific curves with four effective degrees of freedom.

R code

```
Model.Pop <- gam(shoot ~ factor(Pop) + s(skord, k = 5, fx = T,
by = factor(Pop)), family = poisson, data = Four.pop)
summary(Model.Pop)
Model.Pop$deviance
```

SAS code

```
proc gam data = PopS1 ;
      model Shoot = spline(Skord) / dist = Poisson   ;
      output out = Pred ;
run ; * Deviance =   170.05406604, 5 df, 36 obs ;

proc gam data = PopS3 ;
      model Shoot =  spline(Skord) / dist = Poisson   ;
      output out = Pred ;
run ; * Deviance =   119.79814787, 5 df, 35 obs ;

proc gam data = PopN1 ;
      model Shoot =  spline(Skord) / dist = Poisson   ;
      output out = Pred ;
run ; * Deviance =   57.236398516, 5 df, 25 obs ;

proc gam data = PopN3 ;
      model Shoot =  spline(Skord) / dist = Poisson   ;
      output out = Pred ;
run ; * Deviance =   96.104344253, 5 df, 32 obs ;

      * Total deviance = 170.05406604 + 119.79814787 + 57.236398516 +
                        96.104344253 = 443.193, 20 df. ;
```

Figure 3 shows polynomials of degree four and GAMs using four effective degrees of freedom. Besides population N1, SAS and R gave different results. For populations N3, S1, and S3, the SAS curves obtained using four effective degrees of freedom (Figure 3) fitted the observations better than the corresponding SAS curves fitted using generalized cross-validation (Figure 2). Differences between fits based on four effective degrees of freedom and generalized cross-validation were less pronounced in R. All models illustrated in Figure 3 captured the biologically interesting decrease in emergence in September, followed by a period of low emergence until the rate recovered again in November–December. For population N3, the polynomial seemed to predict too sharp a bend towards the end of the period, at about 45 emerged shoots. The curves produced by R were similar to the polynomials, but with less pronounced fluctuations. In our opinion, the curves obtained using SAS were more realistic from the biological point of view.

Deviance output and the approximate F-tests resulting from the GAM analysis in SAS are shown in Tables 1 and 2. The effect of time was clear because a model with a single curve fitted the data significantly better ($P < 0.001$) than a model with only an intercept. A model with region-specific curves did not fit the data significantly better ($P = 0.244$) than the model with a single curve. However, a model with population-specific curves fitted the observations significantly better ($P = 0.011$) than a model with region-specific curves. The model with population-specific curves was significantly better ($P = 0.012$) than the model with only a single curve. Thus, the best model, according to this analysis in SAS, can be described by the four solid curves shown in Figure 3. It can be noted that the corresponding analysis using R led to the same conclusion, i.e., population-specific curves fitted the observations significantly better than a single or two region-specific curves.

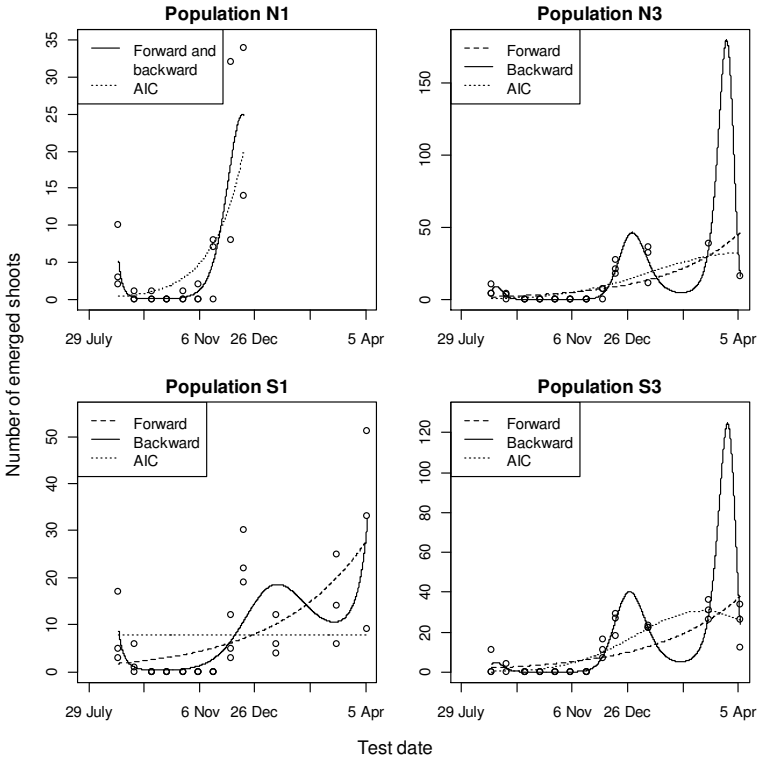

Figure 1. Best polynomials determined by forward and backward selection and using the Akaike information criterion for perennial sow-thistle populations N1, N3, S1, and S3.

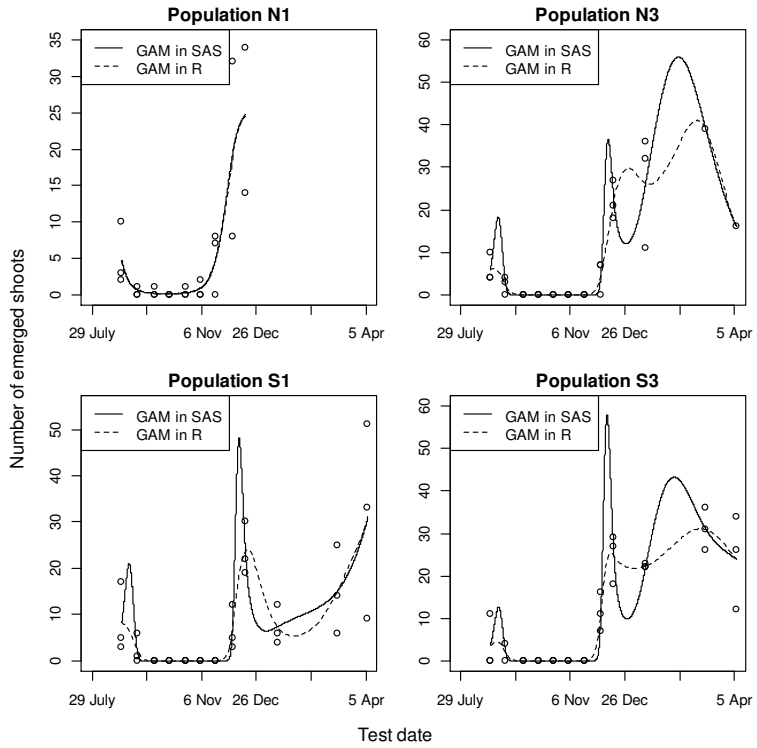

Figure 2. Population-specific GAMs using the generalized cross-validation method of SAS procedure *gam* for perennial sow-thistle populations N1, N3, S1, and S3.

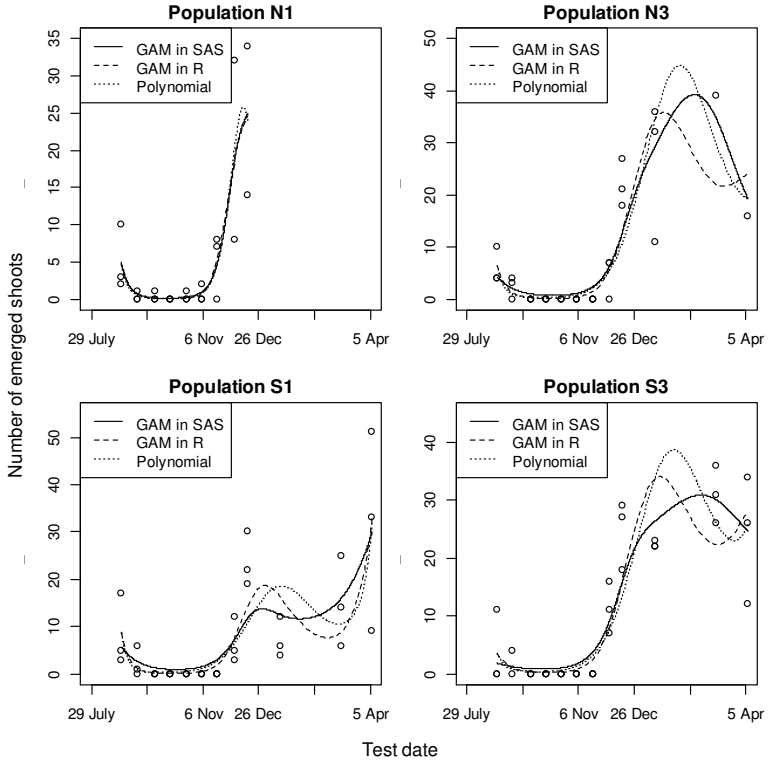

Figure 3. Fourth-degree polynomials compared with GAMs fitted in SAS and R with four effective degrees of freedom for perennial sow-thistle populations N1, N3, S1, and S3.

CONCLUSIONS

This paper deals with GAMs used for comparing experimental treatments when the response variable is a function of one or more explanatory variables. Although there are good introductions to GAMs and other statistical models using the open source software R (such as Zuur et al. 2009 and Everitt and Hothorn 2010), there is, to our knowledge, nothing similar for SAS (apart from SAS Institute 2008). Publications using GAMs for statistical analysis of field experiments are rare. The present paper added a comparison of GAM analyses performed in SAS and R.

The GAM curves fitted in SAS and R can differ. In the example provided here, generalized cross validation performed poorly in SAS, since it produced curves with sharp bends in areas without observations. This is worrying. With a fixed number of effective degrees of freedom, however, SAS produced curves that were slightly more realistic than those produced by R. Based on a small example we will not generalize our conclusions on which of the software program is better. However, it should be pointed out that the functions in R have many more options than the *gam* procedure of the SAS System.

The example provided here illustrated that the procedure of selecting the degree of smoothness is not well established yet. A similar problem sometimes arises when fitting linear models. For example, a third-degree polynomial may fit data significantly better than a second-degree polynomial although the second-degree polynomial does not fit the data significantly better than a straight line. Furthermore, information criteria such as the Akaike information criterion may lead to other conclusions on the model choice than hypothesis testing using forward or backward selection. In the example provided here, the analysis supported the recommendation of Hastie and Tibshirani (1990) to use 3–5 effective degrees of freedom to estimate the smoothing parameter. We believe that in general the number of effective degrees of freedom must be determined from data. The fit of the curve should always be investigated graphically. Cross-validation methods should not be relied upon blindly.

Statistical tests for comparing GAMs are the same as for GLMs. In GLMs, the chi-square test (Eq. 14) and the F-test (Eq. 16) are approximate. However, the performance of the tests for GAMs is more questionable, since the number of degrees of freedom is not exact. There is a risk that the tests are too generous in providing significant results, leading to an over-fitted final model. The curves fitted with penalized maximum likelihood are not maximum likelihood solutions. To overcome these problems, Wood (2006) proposed testing hypotheses using un-penalized fits of functions with fixed smoothing parameters. At the same time, however, fitted curves based on cross-validation and penalized likelihood method are reported. Since this mismatch can mislead, in our view it is better to report approximate *p*-values, but interpret them with care. In particular, one should remember that *p*-values slightly below the significance level might not indicate significant effects. More research is needed on the performance of significance tests for GAMs.

In the example, random effects of rows were not included in the model, while overdispersion was accounted for through the use of approximate F-tests. As was noted, the main plot error variance was small compared with the subplot error variance. For this reason, the loss by ignoring random effects of main plots was probably small, even negligible. Generalized additive mixed models, i.e., GAMs with random effects, can be fitted in SAS (procedure *glimmix*) and R (function *gamm*). Within these procedures, one can also specify correlation structures for modeling temporal or spatial covariance. However, simple models are recommended, because numerical estimation problems can occur (Zuur et al. 2009, p. 323). Verbyla et al. (1999) showed how to analyze designed experiments using generalized additive mixed models and the software ASReml.

Data transformations can sometimes linearize relationships. This method should be preferred to GAMs when applicable because it makes inference easier (Yee and Mitchell

1991). For the same reason, we believe that polynomials should be preferred to GAMs when the polynomials fit the data well. However, biological data are often complex, and fitted polynomials often describe them poorly, typically with amplified bends outside the scatter of observations. Despite all the problems connected with GAMs mentioned above, we encourage increased use of these powerful models. We recommend the use of GAMs when the primary aim is to illustrate biological processes while statistical testing is secondary. For the ease of the beginner, Box 2 offers a short practical guide for GAM analyses.

Box 2. For non-statisticians active in biological research, terms used in statistical text books and journals are sometimes difficult to understand. The following list highlights some basic steps and considerations when evaluating biological data with GAM. The list is not complete, but may work as a starting point for the beginner.

The beginner's guide to GAM analyses

Plot your data! As always, this will give you a first idea of what kind of statistical model to use. If possible, and if statistical testing is primary, use polynomials or other linear or generalized linear models. If such models fit the data poorly and the aim is to illustrate a complex biological process, consider using a GAM.

Choose statistical software. Many software programs cannot perform GAM analyses.

Choose a distribution. Continuous variables are usually modelled using the normal distribution, but the gamma distribution is another option when the data can take only positive values. Counts are usually assumed to be Poisson distributed. The binomial distribution is usually assumed for proportions and when two outcomes (such as dead/alive, or yes/no) are possible.

Choose a link. The most common choices are the identity link for the normal distribution, the log link for the Poisson distribution, and the logit link for the binomial distribution.

Choose the type of smoothing function. This choice may be restricted by the options available in the chosen software. Nowadays, smoothing splines based on penalized likelihood are often preferred.

Choose the smoothing parameter. Using smoothing splines, the smoothing parameter is chosen through the choice of effective degrees of freedom. Use a software integrated smoothness estimator, e.g., cross validation, or find an appropriate number through graphical examination.

Check the residuals. This can be done in the same manner as for GLMs. Plots of deviance residuals vs fitted values should be free from patterns.

Check for collinearity and concurvity. Plot the explanatory variables against each other and study the correlation matrix. When some variables are highly correlated (i.e. collinearity), consider using only one of them. If one of the variables can be well described by a smooth function of another (i.e., concurvity occurs), then consider using only the other variable.

Look for overdispersion. Oversdispersion takes place when the variance of the observations is larger than expected according to the assumed distribution. In case of severe overdispersion, consider to use another model or distribution.

Hypothesis testing and inference. A reduced model can be compared with a full model, using an approximate chi-square test (Eq. 14). For overdispersed data, the approximate F-test (Eq. 16) can be used instead. Keep in mind that these tests are approximate.

ACKNOWLEDGEMENTS

We thank the editors and the reviewers for suggestions that improved the manuscript.

REFERENCES

Andersson L., Boström U., Forkman J., Hakman I., Liew J., Magnuski E. (2013). Sprouting capacity from intact root systems of *Cirsium arvense* and *Sonchus arvensis* decrease in autumn. *Weed Research* 53, 183–191.

Brandsæter, L. O., Fogelfors, H., Fykse, H., Graglia, E., Jensen, R. K., Melander, B., Salonen, J., Vanhala, P. (2010). Seasonal restrictions of bud growth on roots of *Cirsium arvense* and *Sonchus arvensis* and rhizomes of *Elymus repens*. *Weed Research* 50, 102–109.

Cleveland W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74, 829–836.

Elith J., Leathwick J.R., Hastie T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology* 77, 802–813.

Everitt S. E., Hothorn T. (2010). *A Handbook of Statistical Analyses Using R*. 2nd ed. Chapman & Hall/CRC, Boca Raton, FL.

Guisan A., Edwards T.C., Hastie T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* 157, 89–100.

Hastie T.J. (1992). Generalized additive models. In: Chambers J.M., Hastie T.J. (Eds.) *Statistical Models in S*. Chapman & Hall/CRC, Boca Raton, FL.

Hastie T., Tibshirani R. (1986). Generalized additive models (with discussion). *Statistical Science* 1, 297–318.

Hastie T., Tibshirani R. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Hastie T., Tibshirani R. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society, Series B* 55, 757–796.

Hilbe J. M. (1993). Generalized additive models software. *The American Statistician* 47, 59–64.

Jones K., Almond S. (1992). Moving out the linear rut: the possibilities of generalized additive models. *Transactions of the Institute of British Geographers* 17, 434–447.

Marra G., Wood S.N. (2012) Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics* 39, 53–74.

McCullagh P., Nelder J.A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, Cambridge, UK.

Olsson U. (2002). *Generalized Linear Models: An Applied Approach*. Studentlitteratur, Lund, Sweden.

Quinn G.P., Keough M.J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

Ramsay T.O., Burnett R.T., Krewski D. (2003). The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology* 14, 18–23.

R Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

SAS Institute (2008). *SAS/STAT 9.2 User's Guide*. Chapter 36: The GAM Procedure. SAS Institute, Cary, NC.

Shipley B., Hunt R. (1996). Regression smoothers for estimating parameters of growth analyses. *Annals of Botany* 78, 569–576.

Silverman B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, Series B* 47, 1–53.

Thurston S.W., Wand M.P., Wiencke J.K. (2000). Negative binomial additive models. *Biometrics* 56, 139–144.

Venables W.N., Dichmont C.M. (2004). GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. *Fisheries Research* 70, 319–337.

Verbyla A.P., Cullis B.R., Kenward M.G., Welham S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society. Series C* 48, 269–311.

Wahba G. (1983). Bayesian confidence intervals for the cross validated smoothing spline. *Journal of the Royal Statistical Society, Series B* 45, 133–150.

Wood S.N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B* 65, 95–114.

Wood S.N. (2006a). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton.

Wood S.N. (2006b). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* 62, 1025–1036.

Yee T.W., Mitchell N.D. (1991). Generalized additive models in plant biology. *Journal of Vegetation Science* 2, 587–602.

Zuur A. F., Ieno E.L., Smith G.M. (2007). *Analysing Ecological Data*. Springer, New York.

Zuur A.F., Ieno E.L., Walker N.J., Saveliev A.A., Smith G.M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York.