Tellus (2006), 58B, 305–309
Printed in Singapore. All rights reserved

TELLUS

Propagating data uncertainty through smoothing spline fits

By I. G. ENTING[1], C. M. TRUDINGER[2*] and D. M. ETHERIDGE[2], [1]*MASCOS, 139 Barry St, The University of Melbourne, Vic 3010, Australia;* [2]*CSIRO Marine and Atmospheric Research, Private Bag 1, Aspendale, Vic 3195, Australia*

(Manuscript received 28 March 2006; in final form 7 June 2006)

ABSTRACT

Smoothing splines have been used extensively in the analysis of gas concentration data from bubbles in ice cores. Since the fit is a linear projection of the data, propagation of data uncertainty through the fitting process is formally straightforward and, as we demonstrate, readily achievable from pre-existing spline-fitting procedures. The uncertainty propagation can be extended to determining both uncertainties in derivatives and uncertainties in quantities that reflect rates of input to the atmosphere. As an example, we apply the technique to 1000 yr of methane data from a Law Dome ice core.

1. Introduction

Smoothing splines have been used extensively in trace gas studies, particularly for ice-core data (e.g. Joos et al., 1999; Trudinger et al., 1999). Among the advantages of smoothing splines for such studies is the ability to handle unequally spaced data. In addition, as smoothing splines are differentiable functions, the derivative can be used to infer information about sources.

This note describes the process of propagating data uncertainty through the spline-fitting process. As a side benefit, we also show how to propagate uncertainty ranges for the derivatives. For concentrations, $f(t)$, of compounds with atmospheric lifetime τ, uncertainty ranges can also be calculated for combinations such as $\dot{f}(t) + f(t)/\tau$ that reflect source strengths.

The properties of splines have been widely studied but expressed using a range of different normalizations (Cox, 1983; Silverman, 1984, 1985). Enting (1987) summarized some of the most important results, using a normalization consistent with the computer routines given by de Boor (1978).

For a set of N data points taking values z_j at times t_j, the smoothing spline fit, $\hat{f}(t)$, is defined as the instance of the function $F(t)$ that minimizes

$$\Theta = \sum_{j=1}^{N} [F(t_j) - z_j]^2 + \lambda \int_{t_1}^{t_N} [\ddot{F}(t)]^2 \, dt \tag{1a}$$

or more generally with data weights, v_j:

$$\Theta = \sum_{j=1}^{N} [F(t_j) - z_j]^2 / v_j^2 + \lambda \int_{t_1}^{t_N} [\ddot{F}(t)]^2 \, dt \tag{1b}$$

or formally, treating Θ as a functional,

$$\Theta[\hat{f}(t)] \leq \Theta[F(t)] \quad \forall \text{ twice differentiable } F(t). \tag{1c}$$

The solution is a cubic spline, that is, a piecewise cubic with nodes at the t_j and with $\hat{f}(t)$ and its first and second derivatives, $\dot{\hat{f}}(t)$ and $\ddot{\hat{f}}(t)$, continuous everywhere. We use the notation $\hat{f}(t)$ to indicate that we are treating the spline as an estimate of an ideal smooth variation, $f(t)$, which we are trying to estimate in the presence of data uncertainty.

Note that the de Boor code requires distinct times, t_j. Therefore, data sets containing replicate times need to be converted into an equivalent form without such replicates. This can be done, without changing the value of (1b), by using the mean of data values at each replicate time, with a weight of $1/\sqrt{m}$ when m times coincide, [or more generally $(\sum 1/v_j^2)^{-1/2}$].

Note also that for large N, (typically between 1000 and 2000 when using 32-bit arithmetic), the de Boor algorithm becomes unstable. In such cases, the use of splines with fewer nodes (Granek, 1995) should be considered. For the reasons given by Enting (1986), the suggestion by de Boor that the problem be addressed by choosing a node-spacing that gives the requisite smoothing with $\lambda = 0$ is likely to be unsuitable for many applications.

*Corresponding author.
e-mail: cathy.trudinger@csiro.au
DOI: 10.1111/j.1600-0889.2006.00193.x

2. Cases

Various forms of spline fitting can be described by writing (1a) and (1b) as

$$\Theta = S + \lambda Q. \tag{2}$$

These forms are as follows.

(1) Choose λ on the basis of desired filtering. This is the main case that has been used in earlier work by ourselves and our CSIRO collaborators. For case (1a), spline smoothing acts as a low-pass filter with an effective frequency response $\phi(\theta)$ given by

$$\phi(\theta) = 1/\left[1 + (\theta/\theta_{0.5})^4\right], \tag{3a}$$

with the frequency for 50% attenuation given by

$$\theta_{0.5} = (\lambda \Delta t)^{-1/4}, \tag{3b}$$

or equivalently the period for 50% attenuation:

$$P_{0.5} = 2\pi/\theta_{0.5} = 2\pi(\lambda \Delta t)^{1/4}, \tag{3c}$$

where Δt is the mean data spacing. For the generalization (1b), Δt needs to be replaced by an effective data spacing given by the mean of $\Delta t \times v_j^2$.

(2) Choose λ on the basis of data fit, that is, minimize Q given S. Enting (1987) noted that this is the form given by de Boor (1978) and other workers. It is implemented by using case (1), iterating over λ until the condition on S was satisfied. The code used in our studies was derived from that given by de Boor by removing the iterative loop [and converting to the parameter, $p = 1/(1 + \lambda)$, actually used in de Boor's code].

(3) Choose λ on the basis of smoothness, that is, minimize S given Q. Enting (1987) noted early applications of this approach in CO_2 studies.

(4) Generalized cross-validation (GCV) (Craven and Wahba, 1979), which corresponds to choosing λ so that the fit is consistent with the residuals being white noise.

For equispaced data, the uncertainty associated with using filtering to separate a signal, $s(t)$, from noise can be expressed as a mean-square error in the estimate $\hat{s}$ obtained by digital filtering as:

$$\mathrm{E}\{[\hat{s}(t) - s(t)]^2\} = \int_{-\pi}^{\pi} |\phi(\theta)|^2 h_e(\theta)\, d\theta + \int_{-\pi}^{\pi} |1 - \phi(\theta)|^2 h_s(\theta)\, d\theta \tag{4}$$

where $h_s(\theta)$ and $h_e(\theta)$ are the power spectra of signal and noise, respectively [see Box 4.1 of Enting, 2002 for a discussion of normalisation]. The first term represents a variance due to the filter failing to remove all noise. The second term represents a bias due to the filter distorting the signal (most commonly by removing high-frequency variation). From this perspective, the present note is concerned only with the first term, addressing the issue of how much the data uncertainty degrades the smoothed curve. Specifically, our uncertainty calculations reflect the uncertainty in the slowly varying components of concentrations, growth rates and sources. The issue, encapsulated in the second term of (4), of how much pointwise difference there is between an ideal smooth curve and an actual instantaneous value is beyond the reach of this type of data analysis.

3. Error propagation

The solution to the spline-fitting problem can be simplified by writing a general spline function, $F(t)$, as a sum of cubic B-splines (de Boor, 1978), functions $B_j^{[4]}(t)$ that are zero outside the range $t_j < t < t_{j+4}$,

$$F(t) = \sum_j a_j B_j^{[4]}(t). \tag{5}$$

(Since the minimizer of Θ is known to be a spline, the actual minimization calculations can be restricted to the relevant set of splines.)

The minimization of either (1a) or (1b) leads to a set of linear equations for the a_j, written in vector form as:

$$\mathbf{z} = \mathbf{A}\hat{\mathbf{a}} \tag{6a}$$

whence

$$\hat{\mathbf{a}} = \mathbf{A}^{-1}\mathbf{z}, \tag{6b}$$

where the notation $\hat{\mathbf{a}}$ indicates that we treat these as estimates of some true $\mathbf{a}$, the difference, $\hat{\mathbf{a}} - \mathbf{a}$, reflecting uncertainties in the data, z_j. Since $\mathbf{A}^{-1}$ is the inverse of a sparse matrix, rather than being sparse itself, $\mathbf{A}^{-1}$ is never calculated explicitly. Its operation on vectors is implemented by solving sparse matrix equations in the smoothing spline procedures.

These estimated B-spline coefficients lead to estimates, $\hat{f}(t)$, expressed as a vector

$$\hat{\mathbf{f}} = \mathbf{B}\mathbf{A}^{-1}\mathbf{z}, \tag{7a}$$

where $\mathbf{B}$ is a matrix of B-spline values, evaluated at a set of K times, T_k, which need not correspond to nodes, t_j, that is,

$$B_{kj} = B_j^{[4]}(T_k). \tag{7b}$$

Similarly, we can calculate derivatives as

$$\dot{\hat{\mathbf{f}}} = \dot{\mathbf{B}}\mathbf{A}^{-1}\mathbf{z}, \tag{8}$$

where $\dot{\mathbf{B}}$ is a matrix of values of derivatives of B-splines. Note that what we actually calculate is $\dot{\hat{\mathbf{f}}}$, derivatives of the estimate, and we use these as $\hat{\mathbf{f}}$, estimates of the derivative of the ideal $f(t)$.

Formally, data uncertainties, expressed in terms of a covariance matrix $\mathbf{R}$ (whose elements R_{jk} are the covariances of observations z_j and z_k) can be propagated as

$$\mathrm{cov}[\hat{\mathbf{f}} - \mathbf{f}] = \mathbf{B}\mathbf{A}^{-1}\mathbf{R}(\mathbf{A}^{-1})^{\mathsf{T}}\mathbf{B}^{\mathsf{T}}. \tag{9}$$

However, since, as noted above, $\mathbf{A}^{-1}$ is the inverse of a banded matrix, it is computationally more efficient to put

$$\mathrm{cov}[\hat{\mathbf{a}} - \mathbf{a}] = \mathbf{V} = \left[\sum_{j,k} R_{jk} \mathbf{b}^{[j]} (\mathbf{b}^{[k]})^{\mathsf{T}} \right], \tag{10a}$$

where $\mathbf{b}^{[j]}$ is defined by

$$\mathbf{b}^{[j]} = \mathbf{A}^{-1} \mathbf{y}^{[j]} \qquad \text{for } j = 1 \text{ to } N, \tag{10b}$$

where $\mathbf{y}^{[j]}$ has zero elements except for the jth component which is 1. Comparison of (6b) and (10b) shows that the $\mathbf{b}^{[j]}$ can be obtained by applying the smoothing spline fitting procedure to data $\mathbf{y}^{[j]}$. (Note that these N calculations all have to use the same value of λ and the same weights, v_j).

We can then put

$$\mathrm{cov}[\hat{\mathbf{f}} - \mathbf{f}] = \mathbf{B}\mathbf{V}\mathbf{B}^{\mathsf{T}} \tag{11a}$$

and, using the derivative of the estimate as an estimate of the derivative,

$$\mathrm{cov}[\dot{\hat{\mathbf{f}}} - \dot{\mathbf{f}}] = \dot{\mathbf{B}}\mathbf{V}\dot{\mathbf{B}}^{\mathsf{T}}. \tag{11b}$$

Similarly, combinations associated with estimating source strengths, such as $\dot{\mathbf{f}} + \mathbf{f}/\tau$ for a constituent with lifetime τ, can be expressed as

$$\mathrm{cov}[\dot{\hat{\mathbf{f}}} + \hat{\mathbf{f}}/\tau - \dot{\mathbf{f}} - \mathbf{f}/\tau] = [\dot{\mathbf{B}} + \mathbf{B}/\tau]\mathbf{V}[\dot{\mathbf{B}} + \mathbf{B}/\tau]^{\mathsf{T}}. \tag{11c}$$

Note that, in general,

$$\mathrm{cov}[\dot{\hat{\mathbf{f}}} + \hat{\mathbf{f}}/\tau - \dot{\mathbf{f}} - \mathbf{f}/\tau] \neq \mathrm{cov}[\dot{\hat{\mathbf{f}}} - \dot{\mathbf{f}}] + \mathrm{cov}[\hat{\mathbf{f}} - \mathbf{f}]/\tau^2. \tag{12}$$

The formalism for propagating uncertainties becomes particularly simple when $\mathbf{R}$ is diagonal with non-zero elements $R_{jj} = u_j^2$ and only pointwise uncertainties are required for the $f(t)$ [and/or $\dot{f}(t)$ or combinations of these]. One has

$$\mathrm{var}[\hat{f}(T_k) - f(T_k)] = \sum_j R_{jj} \sum_i \left[B_i^{[4]}(T_k) b_i^{[j]} \right]^2, \tag{13}$$

where, for any particular k, the sum over i is restricted due to the finite range for which $B_i^{[4]}(T_k)$ is non-zero.

The weights v_j used in the spline fitting (1b) need not have any special relation to the data uncertainties, u_j. Indeed, the standard definition of smoothing splines, (1a), uses $v_j \equiv 1$. The most useful cases would seem to be the following.

(1) $v_j = 1$ and the u_j are set to actual data uncertainties;
(2) both the v_j and the u_j are set to actual data uncertainties;
(3) it may also be appropriate to use cases where the u_j are various fixed multiples of data uncertainty, in order to produce bands of ranges for several confidence levels;
(4) there is also the possibility, noted in Section 5, of using weighting to perform adjustment of the degree of smoothing within a record.

In each case, these weights would need to be modified, as described above, to remove any replicate times, t_j.

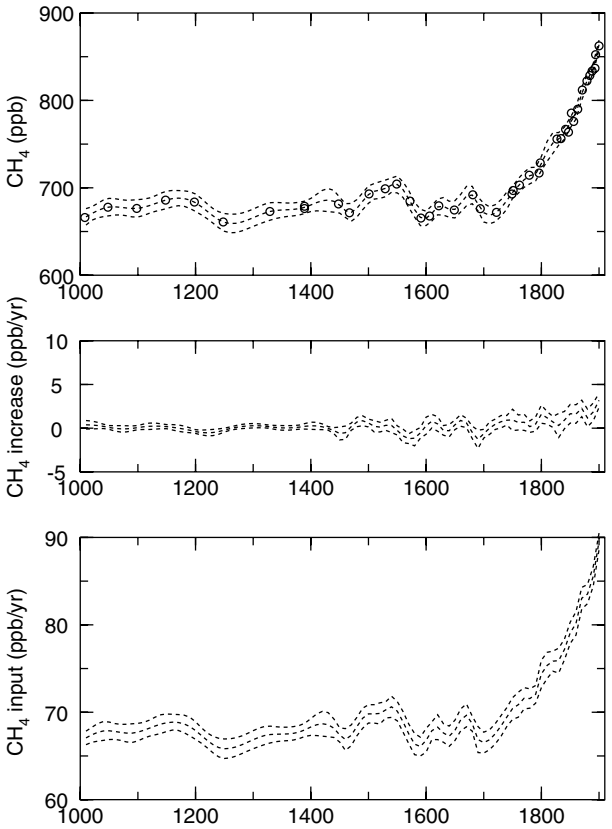

Fig. 1. Illustrative example using data from Etheridge *et al*. (1998) (shown as open circles). Top to bottom panel: splines as estimates (with ±2 s.d. uncertainties) of concentration, f (in ppb); rate of change, $\dot{f}$ (in ppb/yr); rate of input $\dot{f} + f/10.0$ (also in ppb/yr). Fits from a smoothing spline with average 50% attenuation at periods of 26 yr, and ±2 s.d. uncertainties derived from ±10 ppb 2 s.d. data error.

4. Applications

We have implemented the procedure above for calculating uncertainties (11a–c). A Fortran-90 program makes repeated calls to the subroutines from de Boor (1978) to obtain the $\mathbf{b}^{[j]}$ as notional spline fits to data sets, each with a single non-zero value set to one. These $\mathbf{b}^{[j]}$ are either used to construct uncertainty ranges using the special case (13), or used to construct $\mathbf{V}$ the full covariance matrix for the B-spline coefficients (9) which is used in (11a–c) to obtain the full autocovariance matrix for the spline, its derivative and the source function. Information about the availability of the code can be obtained from http://ms.unimelb.edu.au/~enting/spline.html.

As an example, we present CH_4 concentration data from Law Dome ice cores (Etheridge et al., 1998). Figure 1 shows the results of a spline with $\lambda = 30$. From top to bottom, the panels show the spline fit $\hat{f}(t)$, its derivative $\dot{\hat{f}}$, and the combination $\dot{\hat{f}}(t) + \hat{f}(t)/\tau$ which, for a lifetime of $\tau = 10$ yr, corresponds to the estimated rate of input of methane to the atmosphere.

Since the calculation is using total methane, we use the turnover time, and not the perturbation lifetime. [This distinction was introduced by Prather (1994) and illustrated in figure 15.1 of Enting (2002)]. Inspection of (11c) shows that generalization of the formalism to include an explicitly time-varying lifetime, $\tau(t)$, is straightforward. The value of $\tau = 10$ yr is chosen here to aid visual comparisons between sections of Fig. 1, and is somewhat higher than current best estimates.

The ranges correspond to a ± 10 ppb data uncertainty which Etheridge *et al.* (1998) gave as corresponding to two standard deviations in measurement uncertainty. Although the spline fit is calculated using the full data set, which extends to 1980, the plots are terminated at 1900. The CH_4 concentration increase over the 20th century is so large compared to the uncertainty ranges that plotting the full fit at this size would have the uncertainty ranges comparable to the line thickness. The ± 2 s.d. range in data becomes, by virtue of the linear relations, ± 2 s.d. ranges in the $\hat{\mathbf{a}}$ and the $\hat{f}(t)$, with (in each case) 5% of individual cases expected to lie outside these ranges. However, individual cases will be correlated, with the correlations described by (11a,b,c) for splines, growth rates and sources.

The panels are drawn so that the scales for $\hat{f}(t)/10$, $\dot{\hat{f}}$, and $\dot{\hat{f}}(t) + \hat{f}(t)/10$ are equal, allowing ready comparison of the uncertainties. In our example, the derivative has smaller uncertainty before 1400 when data density is lower, than after 1400, whereas the expectation would have been that more data should reduce the uncertainty. Conversely, the uncertainty in methane input, $\dot{\hat{f}}(t) + \hat{f}(t)/10$, is relatively constant over the whole period. These and some other counterintuitive features of the results are explained in the next section.

5. Discussion

An understanding of the results in Fig. 1 comes from careful consideration of what questions are actually being answered. Wider discussion in the context of CO_2 time-series is given by Enting (2000). In particular, as noted above, what is calculated is the effect of data uncertainty on uncertainties in smooth trends in atmospheric concentrations and growth rate, not in instantaneous concentrations and growth rates. A critical point is that different choices of smoothing imply different uncertainties. More smoothing means that more actual variability is removed, and its uncertainty is, by definition, being excluded.

We mentioned earlier that the spline has smaller uncertainty before 1400 than after. The degree of smoothing of a spline depends on the data spacing (for example, this is shown by equation (3c) which can be applied to the data density Δt_{local} for a particular part of the record). The lower data density before 1400 leads to a smoother spline for the earlier part of the record. The uncertainty in the spline before 1400 is less than the uncertainty in the spline after 1400 mainly because the spline before 1400 has a higher degree of smoothing due to the lower data density.

The derivative is far more sensitive to the cutoff frequency than the spline itself, since higher frequencies contribute most to the derivative. A smoother spline implies less uncertainty in the derivative *of the spline*. As noted in connection with equation (4), information about higher frequency contributions to the actual atmospheric growth rate is inaccessible due to lack of data. The effect becomes particularly marked when, as here, the notional smoothing cutoff time, $P_{0.5}$ (resulting from the choice of λ) is comparable to the data spacing. In cases, such as our example, where there is a distinct change in data spacing, a more uniform filtering could be achieved by using weighted splines. The cutoff can be obtained from (3c) with λ replaced by $\lambda \times v_j^2$, so that to keep $P_{0.5}$ constant the v_j^2 should be set proportional to $1/\Delta t_{\text{local}}$ (see for example Trudinger *et al.,* (1999). The use of constant weights in our example illustrates both the variation of smoothing with data density and the related dependence of uncertainty on chosen smoothing, emphasizing the need to be clear about what is being estimated. In our example, the change in data density has little effect on the uncertainty range for methane input, $\dot{\hat{f}}(t) + \hat{f}(t)/10$.

Another somewhat surprising characteristic of the fit is that the uncertainty range in the derivative only excludes zero at a time well after the apparent beginning of the increase after 1700. Careful inspection of the concentration fit suggests that, at the smoothing time-scale of around 30 yr, such a large uncertainty is appropriate. The derivative of the spline fit to concentration is actually quite variable.

6. Conclusion

With the increasing importance of paleodata in establishing the various natural and anthropogenic contributions to global climate change, tools that can assist in quantifying uncertainties have a vital role to play. We have presented a method for estimating the uncertainty in smoothing splines due to data uncertainty. We emphasize in particular that the uncertainty range relates to the chosen degree of smoothing for the spline and refer to uncertainties in the smoothed component.

With modern computing power, exact propagation of data uncertainty through spline fits becomes more convenient, and in this case more accurate, than analytic approximations based on asymptotic results.

7. Acknowledgments

The Center of Excellence for Mathematics and Statistics of Complex Systems (MASCOS) is funded by the Australian Research Council. The authors wish to thank F. Joos and an anonymous referee for useful comments and C. de Boor for permission to make the code available. IGE's fellowship at MASCOS is

supported in part by CSIRO. This work was created using the Tellus LaTeX2e class file.

Notation

a_j Coefficient of jth B-spline.

A Matrix defining the mapping from B-spline coefficients to data.

$B_j^{[4]}(t)$ The jth B-spline: a cubic spline that is zero outside the range $t_j < t < t_{j+4}$ (notation from de Boor (1978)—the [4] indicates the order of the spline).

$\hat{f}(t)$ Spline fit, with first and second derivatives $\dot{\hat{f}}(t)$ and $\ddot{\hat{f}}(t)$.

$f(t)$ Ideal 'error-free' smooth function of which the spline fit, $\hat{f}(t)$, is regarded as an estimate.

$F(t)$ Arbitrary twice differentiable function in set of functions over which the objective Θ is (formally) minimised.

K Number of times at which splines, derivatives and their linear combinations are evaluated.

N Number of nodes.

$P_{0.5}$ Period of variations subject to 50% attenuation by spline fit.

Q Characteristic smoothness, defined as integral of square of second derivative.

S Sum of squares of deviations between spline and data.

t Time.

t_j Time of jth node, $j = 1$ to N.

T_k The kth time for which the spline is to be evaluated, $k = 1$ to K, with default in code as $T_k = T_0 + k\Delta T$.

u_j Data uncertainties used in error propagation calculations.

v_j Weights used in spline fit.

$\mathbf{y}^{[j]}$ Vector, with element j equal to one and all other elements zero.

z_j The jth data value (occurring at time t_j).

Δt Mean data spacing.

ΔT Spacing of output times.

$\phi(\theta)$ Filter response function, in frequency domain.

λ Regularization parameter, defining trade-off between data-fit and smoothness.

θ Angular frequency.

$\theta_{0.5}$ Angular frequency for 50% attenuation by spline fit.

Θ Objective function whose minimum defines smoothing spline. $\Theta = S + \lambda Q$.

τ Atmospheric lifetime.

References

Cox, D. D. 1983. Asymptotics for M-type smoothing splines. *Annals of Statistics*, **22**, 530–551.

Craven, P. and Wahba, G. 1979. Smoothing noisy data with spline functions. *Numer. Math.*, **31**, 377–403.

de Boor, C. 1978. *A Practical Guide to Splines*. Springer-Verlag, New York.

Enting, I. G. 1986. Potential problems with the use of least squares spline fits to filter CO_2 data. *J. Geophys. Res.*, **91D**, 6668–6670.

Enting, I. G. 1987. On the use of smoothing splines to filter CO_2 data. *J. Geophys. Res.*, **92D**, 10977–10984.

Enting, I. G. 2000. *Characterising the temporal variability of the global carbon cycle*. CSIRO Atmospheric Research Technical Paper no. 40. CSIRO, Australia. http://www.cmar.csiro.au/e-print/open/enting_2000a.pdf

Enting, I. G. 2002. *Inverse Problems in Atmospheric Constituent Transport*. CUP, Cambridge, UK.

Etheridge, D. M., Steele, L. P., Francey, R. J., and Langenfelds, R. L. 1998. Atmospheric methane between 1000 A.D. and present: Evidence of anthropogenic emissions and climatic variability. *J. Geophys. Res.*, **103D**, 15979–15993.

Granek, H. 1995. Generalized smoothing splines in CO_2 analysis. *J. Geophys. Res.*, **100D**, 16857–16865.

Joos, F., Meyer, R., Bruno, M., and Leuenberger, M. 1999. The variability in the carbon sinks as reconstructed for the last 1000 years. *Geophys. Res. Lett.*, **26**(10), 1437–1440.

Prather, M. J. 1994. Lifetimes and eigenstates in atmospheric chemistry. *Geophys. Res. Lett.*, **21**, 801–804.

Silverman, B. W. 1984. Spline smoothing: The equivalent variable kernel method. *Annals of Statistics*, **12**, 989–916.

Silverman, B. W. 1985. Some aspects of the spline: Smoothing approach to non-parametric regression curve fitting. (with discussion). *J. R. Statist. Soc.*, **47**, 1–52.

Trudinger, C. M., Enting, I. G., Francey, R. J., Etheridge, D. M., and Rayner, P. J. 1999. Long-term variability in the global carbon cycle inferred from a high precision CO_2 and $\delta^{13}C$ ice core record. *Tellus*, **51B**, 233–248.