# FITTING POPULATION DYNAMIC MODELS TO TIME SERIES DATA BY GRADIENT MATCHING: APPENDIX.

STEPHEN P. ELLNER, YODIT SEIFU, AND ROBERT H. SMITH

## APPENDIX A: DETAILS OF METHODS

*Bandwidth selection for estimating the gradient.*

The choice of bandwidth $h$ for local polynomial regression is usually a compromise between the bias and the variance of the fitted curve. At a given point $x$ the mean squared error (MSE) between the estimate $\hat{f}(x)$ and the true value $f(x)$ can be decomposed as $MSE = bias^2 + variance$, where $bias = E[\hat{f}(x)] - f(x)$, $variance = Var[\hat{f}(x)]$. For large $h$ the fitted value at any point is based on a large fraction of the total data set, hence the variance is low. However a low-order polynomial is usually a poor approximation over a long span of time, so the bias is high. A smaller bandwidth reduces the bias, but increases the variance. The optimal bandwidth for minimizing MSE is therefore an intermediate value where neither the bias nor the variance is too large.

For our purposes this reasoning about bandwidths does not apply because estimating *dx/dt* is just an intermediate step towards estimating rate equations by regressing *dx/dt* on the state variables. In that latter step, the estimates of *dx/dt* appear as the dependent variable. As sample size $n$ (length of the time series) increases, the variance of estimated regression parameters due to variance in estimates of *dx/dt* decreases at rate $1/n$. But the bias in regression parameters due to bias in estimates of *dx/dt* does not decrease at all, because one is just adding more and more data points with the same systematic distortions. An "optimal" bandwidth in this situation is therefore far smaller than the usual optimality criteria would suggest. As a rule of thumb we suggest that $h$ should be roughly 1-2 times the time interval between population counts, with higher measurement error variance favoring larger $h$. In a specific application the value of $h$ could be fine-tuned by using simulated data from the fitted model to produce a plot like our Figure 1 that reveals the amount of bias in the gradient estimates. The use of small $h$ is a significant difference between this paper and Ellner et al. (1997). Small $h$ drastically reduces the fraction of pairwise-correlated gradient residuals. This allows fitting and model selection to proceed as if gradient residuals were uncorrelated (as explained below: *Generalized or Ordinary Least Squares?*), eliminating the need for the *ad hoc* methods used by Ellner et al. (1997).

Gradient estimates can be improved by estimating and correcting for the bias at extreme gradient values. By passing an interpolating cubic spline through the time series, we obtain a series very similar to the original data, for which the exact value and gradient are known at all times. We can therefore simulate the process of sampling the interpolated series (at the same frequency as the original time series) and estimating the gradient at sampling times. The relationship between the estimated and true gradients for the interpolated series is then fitted by a regression spline constrained to be monotonic, and the fitted spline is applied to the original gradient estimates. The net result is essentially a smoothed version of estimating the gradient by interpolation ($h$=0). Interpolation minimizes the bias but it is unstable (especially with finely sampled noisy data) and prone to occasional spectacular errors. Using interpolation indirectly, to estimate and remove the bias of a stable estimate, eliminates the instability of interpolation without greatly increasing the bias.

*Penalized regression splines*

The additive model (6) was fitted by linear regression onto the set of regression spline basis functions, each of which is a function of $X$ alone, or of $Y$ alone (Ruppert and Carroll 1997). The basis functions for a variable $X$ on an interval $[a,b]$ are defined by a set of *knots* $\{\kappa_j, j = 1,2,\cdots m\}$, $a < \kappa_j < b$, and the regression spline estimate of a function *f(X)* is

$$\hat{f}(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \sum_{j=1}^{m} \beta_{3+j} \max(X - \kappa_j, 0)^3 \qquad (.1)$$

Equation $(.1)$ is the cubic regression spline, and quadratic or linear regression splines are defined similarly. To fit the additive model (6) we simply add a second "copy" to the right-hand side of $(.1)$, with $X$ replaced by $Y$. The resulting model has the form

$$\begin{aligned} B(X) - D(Y) = {} & \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \sum_{j=1}^{m_x} \beta_{3+j} \max(X - \kappa_j, 0)^3 \\ & + \delta_1 Y + \delta_2 Y^2 + \delta_3 Y^3 + \sum_{j=1}^{m_y} \delta_{3+j} \max(Y - \lambda_j, 0)^3 \end{aligned} \qquad (.2)$$

where the $\lambda_j$ are the knots for $Y$. Note that in $(.2)$ the intercept term has been omitted, reflecting the fact that no births or deaths occur at zero population density. In all analyses reported here we used $m$=20 knots for both $B$ and $D$.

The fitting criterion is the sum of squared errors (**SSE**) penalized by the sum of squared spline coefficients,

$$\sum_i \left(\hat{Z}_i - Z_i\right)^2 + \alpha_X \sum_{j=1}^{m_X} \beta_{3+j}^2 + \alpha_Y \sum_{j=1}^{m_Y} \delta_{3+j}^2 , \qquad (.3)$$

with the "smoothing parameters" ($\alpha_X$, $\alpha_Y$) determining the complexity of the fitted curve as described in the text. The **SSE** in equation (.3) is an ordinary least squares criterion, which ignores the correlations in the errors (in *Z*) that result from the gradient estimation step. However, as we discuss below, trying to take account of the correlations by using a generalized least squares criterion is actually counterproductive.

When the true death rate functions are polynomials, as in the simulation models used to test the methods, it might appear that the spline terms in the fitted model (the *Y* terms with *j*>3 in equation (.2)) are superfluous. However fits with the spline terms omitted were much more strongly biased. The spline terms allow the fitted function to err in one spot (for example, the occasional spurious curvature when the death rate is actually linear) without the error propagating to other parts of the fitted curve. Even a reduction to 5 knots (instead of 20) caused a visible reduction of accuracy in the fits.

*Fitting a Penalized Regression Spline with Constraints*

Our procedures require fitting penalized regression splines to data $\{x_i, y_i\}_{i=1}^N$ with the constraint of the fitted spline being positive, monotonic increasing in *x*, or both. Following Wood (1997), our numerical methods applied the looser (but in practice equivalent) constraint that the function values at a set of points $\{\chi_i\}_{i=1}^M$ satisfy the constraint(s); we refer to these as the *constraining grid points*. For positivity constraints the default in our code (and throughout the paper) is that the $\chi_i$ consist of the knots supplemented by the endpoints of the interval on which the spline is defined, while for monotonicity constraints a finer grid (generally 50 points equally spaced between the endpoints) was required.

Let *X* be the matrix of size *N*×(*m*+4) whose columns consists of the spline basis functions $1, x, x^2, x^3, \{\max(x-\kappa_j, 0)^3\}_{j=1}^m$ evaluated at the *N* data points $x_i$, *Y* the vector of corresponding $y_i$ values, and $\theta = (\beta_0, \beta_1, \cdots \beta_{3+m})$. We can assume that the *x*'s are sorted in increasing order.

Given the smoothing parameter $\alpha$, let $D$ be the square diagonal matrix of linear size ($m$+4) with diagonal entries $(0,0,0,0,\alpha,\cdots,\alpha)$. The unconstrained fitting problem is to minimize

$$F(\theta) = (Y - X\theta)^T (Y - X\theta) + \theta^T D\theta .$$

To apply the constraints, let $Z$ be the the $M\times(m+4)$ matrix whose columns are the basis functions evaluated at the $M$ constraining grid points, and let $C$ be the matrix whose $j$th row is the difference between ($j$+1)st and $j$th rows of $Z$. Thus, $Z\theta$ is the vector of regression spline values at the constraining grid points, and $C\theta$ is the vector of differences between the values at successive grid points. The constrained fitting problem then is to maximize $F(\theta)$ subject to $Z\theta \geq 0$, $C\theta \geq 0$, or both. As Wood (1997) observed, these are standard quadratic programming problems; we obtained numerical solutions using the **quadprog** library in **R** (**www.R-project.org**). The function in the supplied code also returns the GCV score for the fitted model, so the value of $\alpha$ can then be selected by minimizing GCV($\alpha$).

*Generalized or Ordinary Least Squares?*

The differences between gradient values estimated by smoothing the time series, and the expected gradients from the model (equation 2), will generally be autocorrelated because the $j^{th}$ data value affects the gradient estimates at sampling times $j \pm 1, j \pm 2, \cdots$, out to several multiples of the bandwidth $h$. This suggests that it might be preferable to fit $B$ and $D$ using a generalized least squares (GLS) fitting criterion that takes account of correlated errors, rather than the ordinary least squares (OLS) criterion ( .3). OLS estimates are still unbiased when errors are correlated, but GLS estimates can have a higher efficiency (i.e. a lower variance) if the variance-covariance matrix of the errors is known or estimated (Draper and Smith 1981). However we recommend using OLS, for the following two reasons.

First, the error correlation is weak enough that OLS is nearly as efficient as GLS, due to the "whitening by windowing principle" (Hart 1996) for nonparametric time series models in the presence of autocorrelated errors. The principle derives from the fact that for large sample size, temporally adjacent points in the original time series usually are not nearest neighbors in the space of the model's independent variables – in our case, $(x(t), x(t-\tau))$. Rather, most points in any small neighborhood in $(x(t), x(t-\tau))$ space come from widely separated times, so the

associated errors are nearly uncorrelated (i.e., the errors in a spatial window are nearly "white noise"). For the model in Figure 1 the temporal autocorrelation of errors at lag 1 is significant at about +0.5 (depending on the particular realization of the model and measurement errors), but the autocorrelation is below 0.2 in absolute value at all higher lags because of the small bandwidth used for the gradient estimates. Consequently the spatial autocorrelation of residuals (Moran's $I$ statistic) in $(x(t), x(t-\tau))$ space is significantly positive but small. In 25 replicate realizations of the model, the maximum spatial autocorrelation (for neighborhood areas between 5% and 50% of the smallest square covering the data) never exceeded +0.20, and only one $I$ value exceeded +0.15.

The second reason is that the cure is often worse than the disease, in the presence of measurement errors. Measurement error in population density can cause appreciable bias in both OLS and GLS regression coefficients, and (as discussed in the main text) this bias is the most problematic source of estimation error in our procedures. Analytic large-sample approximations and simulations (Y. Seifu, *unpublished*) show that the bias is larger for GLS than for OLS. The higher bias of GLS estimates often outweighs their slightly smaller variance, such that GLS estimates have a higher mean-square error than OLS estimates.

*Poisson sampling model*

Measurement errors for simulated data were generated by a mechanistic model for the sampling process. We assume that the population is sampled such that each individual has some small probability $p$ of being counted, and we refer to $p$ as the *capture probability*. When there are $N$ individuals in the population, the sample is therefore distributed as $Binomial(N,p)$, which we approximated as $Poisson(Np)$. For plotting and model fitting, these values were re-scaled to estimate the actual population size. Thus a "data with measurement error" value for a simulated data point with $N$ individuals is distributed as $(1/p) \times Poisson(Np)$, which has mean $N$ and standard deviation $\sqrt{N/p}$.

*SIMEX bias reduction*

Given a model for the measurement or sampling errors, the bias in regression estimates can be reduced by the SIMEX procedure (Cook and Stefanski 1994, Stefanski and Cook 1995). For the case of nonparametric regression (Carroll et al. 1995, 1999), the general procedure is to

- Repeatedly re-fit the model with simulated measurement errors added to the independent variables;
- Compute the average estimates as a function of the level of added measurement error;
- Extrapolate back to zero measurement error; Carroll et al. (1995, 1999) recommend quadratic extrapolation.

In our simulation models we have assumed the Poisson sampling model described above. We can therefore add a second "dose" of sampling error by simulating the process of sampling from a population of the observed size, and sample from that sample to add a third "dose", etc. Because the model ( .2) is linear in the coefficients, the average function estimates correspond to average parameter estimates. Let $\hat{\beta}_i$ denote the mean parameter vector with *i* doses of measurement error, *i*=1,2,3, with $\hat{\beta}_1$ being the estimate from the data. Quadratic extrapolation through these values back to *i*=0 yields the SIMEX estimator $\hat{\beta}_0 = 6\hat{\beta}_1 + 4\hat{\beta}_3 - 9\hat{\beta}_2$. Because of the large coefficients multiplying the $\hat{\beta}_i$ we used 500 refittings at each level of measurement noise. In the SIMEX example in the code that we provide, a univariate regression with 200 data points, 500 refittings takes about 40 seconds on an 800 MHz PC. The smoothing parameters were not re-estimated for the refittings, as the increase in measurement error variance had very small effects on the value of GCV2-optimal smoothing parameters. Our simulation results indicate that this shortcut is safe for the data analyzed here.

In applications, the capture or sampling probability *p* typically will have to be estimated. More generally, in order to use SIMEX one needs a parametric model of the measurement error process. This could be done theoretically, based on a model of the sampling process (as in Turchin and Ellner 2000) or empirically based on a small-scale set of replicate measurements at a range of population densities. It is tempting but sadly not feasible to parameterize a sampling error model by smoothing the time series and using the residuals as estimated errors, because the residuals are too dependent on the smoothing method. Two reasonable (and visually very similar) choices of a smooth gave maximum likelihood estimates of *p* that differed by about a factor of 5.

*More on smoothing parameter selection*

In addition to GCV2 we also considered a more conservative criterion based on the spatial autocorrelation of residuals from a fitted model (Ellner and Seifu, 2001). Residuals from an insufficiently complex model exhibit positive autocorrelation in the space of independent variables, due to the unfitted structure in the data. The selection criterion is to find the simplest model (as measured by the total effective degrees of freedom for both rate functions) such that the residual spatial autocorrelation (RSA), as measured by Moran's *I* statistic, is 0. Relative to GCV2 the RSA criterion almost invariably selects slightly simpler models, having slightly higher bias but fewer spurious wiggles on average (Ellner and Seifu, 2001). In the models considered here, fitted rate functions from the RSA and GCV2 criteria were visually almost indistinguishable. We therefore report results only for the GCV2 criterion, which has been tested more thoroughly (Nychka et al. 1992, Ellner and Turchin 1995) and is much quicker to compute.

**References**


Carroll, R.J., D. Ruppert, and L.W. Stefanski. 1995. Measurement Error in Nonlinear Models. Chapman and Hall, NY.

Carroll, R.J. J.D. Maca, and D. Ruppert. 1999. Nonparametric estimation in the presence of measurement errors. Biometrika 86: 541-554.

Cook, J.R. and L.A. Stefanski. 1994. Simulation-extrapolation estimation in parametric measurement error models. Journal of the American Statistical Association 89: 1314-1328.

Draper, N. R. and H. Smith. 1981. Applied Regression Analysis, 2nd edition. Wiley, NY.

Ellner, S. P., B.E. Kendall, S.N. Wood, E. McCauley, C.J. Briggs. 1997. Inferring mechanism from time-series data: delay-differential equations. Physica D 100:182-194.

Ellner, S. P. and Y. Seifu. 2001. Using spatial statistics to select model complexity. Journal of Computational and Graphical Statistics, *in press*.

Ellner, S. P. and P. Turchin, 1995. Chaos in a noisy world: new methods and evidence from time series analysis. American Naturalist 145: 343-375.

Hart, J. D. 1996. Some automated methods of smoothing time-dependent data. Journal of Nonparametric Statistics 6: 15-142.

Nychka, D.W., S. Ellner, A.R.Gallant, & D. McCaffrey. 1992. Finding chaos in noisy systems (with discussion). Journal of the Royal Statistical Society Series B 54, 399-426.

Ruppert, D. and R.J. Carroll. 1997. Penalized regression splines. Technical Report TR1249, Department of Operations Research and Industrial Engineering, Cornell University.

Ruppert, D. and R.J. Carroll. 2000. Spatially adaptive penalties for spline fitting. Australian and New Zealand Journal of Statistics 42: 205-223.

Stefanski, L.A. and J.R. Cook. 1995. Simulation-extrapolation: the measurement error jackknife. Journal of the American Statistical Association 90: 1247-1256.

Turchin, P. and S. P. Ellner. 2000. Living on the edge of chaos: population dynamics of Fennoscandian voles. Ecology 81: 3099-3116.

Wood, S.N. 1997. Inverse problems and structure-population dynamics. pp. 555-586 in: Tuljapurkar, S. and H. Caswell (eds.). 1997. Structured-population models in marine, terrestrial, and freshwater systems. Chapman and Hall, NY.