

Smoothing

Smoothing is an exploratory tool that can be applied on various forms of data to achieve various purposes. Tukey [22, Chapter 7] advocated smoothing in his classic work *Exploratory Data Analysis* (EDA) and illustrated its use on data sequenced by a single variable such as time. Like all exploratory tools, the emphasis is on insight and flexibility, in contrast to hypothesizing a specific functional form, estimating parameters, and testing for model adequacy. Although smoothing methodologies are still described most often in the context of one-dimensional data (i.e. a response as a function of one variable), with *time* as the explanatory variable, Tukey [22, Chapter 8 onwards] also showed by example that smoothing on variables other than time, as well as higher-dimensional smoothing (i.e. a response as a function of several variables), also provide insights into functional relationships between variables. In the context of environmental data, this higher-dimensional smoothing often takes the form of bivariate smoothing with a response (e.g. concentration of **radon** or volume of oil) as a function of two geographical variables (e.g. longitude and latitude).

The decomposition of data into $data = fit + residuals$ or $data = smooth + rough$ [22, p. 208] suggests a need to characterize the ‘smooth’ component of the data, which often leads to insight about the process that generated the data. Some of the goals of smoothing are:

1. To reveal the relationship between response and explanatory variables, which may suggest a functional model that describes their connections.
2. To magnify the underlying trend (*see* **Trend, detecting**).
3. To reduce attention to unusual values or **outliers**.
4. To examine patterns in the **residuals** that can be revealed once the smoothed trend has been removed.
5. To minimize the effect of aggregated values (sometimes called *binning* [12]), such as incidence measures that might apply to an entire region rather than to a specific point identified with the response.

Notice that these objectives differ from those for interpolation, where the fitted surface is constrained

to pass through the observed data values and the goal is to estimate the response when the explanatory variables take on values other than those at hand. Interpolation in particular generally applies when **measurement error** is thought to be negligible. Most environmental applications involve errors from various sources in the measuring device (e.g. upper and/or lower limits of detection, location, calibration), so smoothing, rather than interpolation, is often more appropriate.

Classes of Smoothers

Smoothers fall into one of two basic categories: *linear*, including local polynomial smoothing, *loess* (cf. [4] and [13]), splines [25] and kriging [5], and *nonlinear*, such as running medians and other median-based smoothers [14, 22, 24]. This entry introduces the concept and need for smoothing, and discusses briefly the smoothers in these two categories followed by examples in one and two dimensions.

A simple but illustrative example of both linear and nonlinear smoothing appears in Figure 1. These data are modified measurements of concentrations of a particular contaminant on $n = 20$ consecutive days. In the first row of plots, a shift has been added in the middle of the sequence; in the second row, an outlier has been added. The left column shows the effect of smoothing by running *means* of length 3: $\tilde{y}_i = (y_{i-1} + y_i + y_{i+1})/3$, $i = 2, \dots, 19$. The right column smooths by running *medians* of length 3: $\tilde{y}_i = \text{median}\{y_{i-1}, y_i, y_{i+1}\}$. (In both cases, $\tilde{y}_i = y_i$ at the endpoints $i = 1$ and 20.) Note that the eye is drawn to the trend line that is shown in each plot, which is one of the goals of smoothing (reduce attention to distracting detail). The smooths in the top two plots suggest immediately a change in level midway across the x -axis; the bottom two plots suggest almost constant values, apart from measurement error and possibly an outlier. These plots also show the advantage of smoothing by running medians in these two situations: median smoothing responds more quickly to abrupt features (vs. the more gradual shift when smoothing by means), and also is not influenced by single outliers unsupported by neighboring values (in contrast to the ‘tent’ constructed by the means smoother). Similar characteristics apply generally to other nonlinear smoothers. Outliers should be detectable readily in the *residuals* (original values

2 Smoothing

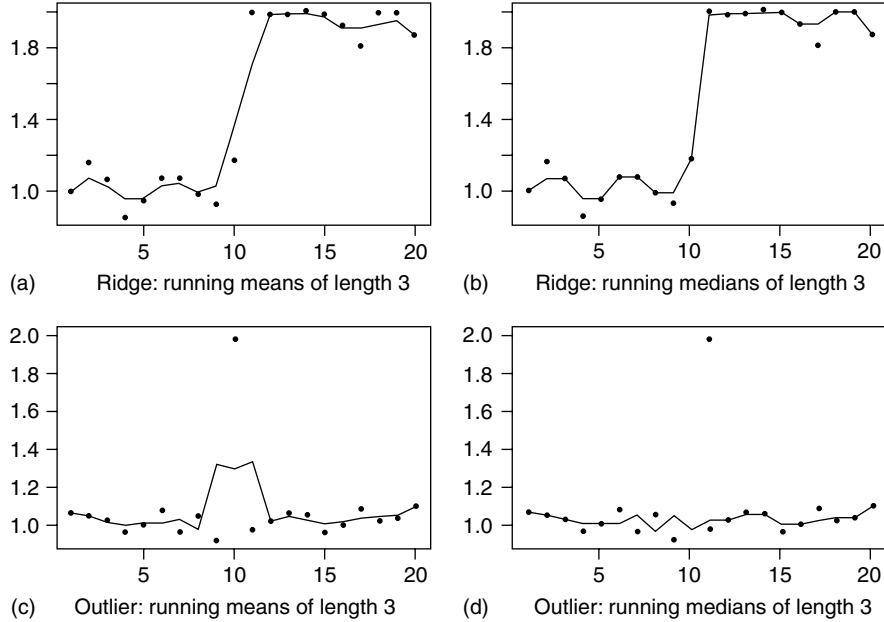


Figure 1 Four sets of smoothing a sequence of 20 data points. (a) Running means of length 3 on a sequence with an abrupt ridge. (b) Running medians of length 3 on a sequence with an abrupt ridge. (c) Running means of length 3 on a sequence with an outlier. (d) Running medians of length 3 on a sequence with an outlier

– smooth values); residuals from the median smooth are similar except for the one value at $i = 10$, whereas the mean smooth presents *three* potential outliers at $i = 9, 10$ and 11 .

Models and Characteristics of Smoothers

In the example above, the underlying model for data that motivates smoothing, $data = smooth + residual$, can be written more formally as

$$y_i = f(x_i) + r_i \quad (1)$$

We assume that f is reasonably smooth (apart from possible ridges or abrupt changes), and that r_i represents a mild departure (or occasionally a serious outlier) from the true $f(x_i)$. Note that x_i may involve more than one variable; for geographical smoothing, $\mathbf{x}_i = (x_{i1}, x_{i2}) = (\text{longitude}, \text{latitude})$. One way to understand the action of smoothing by running means is to consider the case where f is reasonably constant in a neighborhood of x_i and the residuals r_i are independent with constant variance σ^2 . Then $f(x_{i-1}) \approx f(x_i) \approx f(x_{i+1})$, so the mean of \tilde{y}_i is approximately equal to the desired $f(x_i)$, while its

variance is reduced by one-third. A wider *span* will reduce the variance even further, at the expense of introducing *bias* into the estimate [i.e. $f(x)$ may not look much like $f(x_i)$ if x is far from x_i], leading to the familiar *bias–variance* tradeoff. When the values of y_i and x_i are assumed to arise as realizations of some (possibly multivariate) random variables Y and X , and the mean of the residuals r_i is zero (for otherwise its mean could be incorporated into f), then $f(x_i)$ may be expressed as the conditional expectation $E(Y|X = x_i)$. This is the usual situation for *linear regression* when $f(x_i)$ is a line (e.g. $a + bx_i$ for some a and b – see **Linear models**) or, more generally, *parametric regression* when the function f is specified in terms of parameters (e.g. a and b above), and for *nonparametric regression* when the functional form of $f(\cdot)$ is left completely unspecified.

Characteristics of a good smoother are:

1. it should recreate f as accurately as possible;
2. it should recapture perfectly linear trends or surfaces;
3. it can handle unevenly spaced data;
4. its output should be ‘smooth’ (except at sharp break points);

5. extreme outliers should be properly ignored by the smooth and stand out clearly in the residuals.

Linear Smoothers

Linear smoothers can always be expressed as a linear function of the data, e.g. $\tilde{y}_i = \sum_{j=1}^n a_{ij} y_j / \sum_{j=1}^n a_{ij}$. The weights usually depend upon the target point being smoothed, e.g. in the running means example $(a_{3,1}, \dots, a_{3,20}) = (0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, \dots, 0)$, whereas $(a_{4,1}, \dots, a_{4,20}) = (0, 0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, \dots, 0)$. The *span* of the smoother is usually defined as the proportion of the weights that are nonzero; the larger the span, the smoother (and less variable) the surface, but the potentially greater bias in the estimate of f . Most nonparametric regression estimators take the form of linear smoothers, as can be seen in the equation for kernel-based estimators [(2) in **Nonparametric regression model**] and loess [(5) in **Nonparametric regression model**].

Spline smoothers (*see Splines in nonparametric regression*) may be viewed as extensions of the simple running mean/median smoothers, which implicitly assume that f is roughly constant in local neighborhoods, and of loess, which assumes that f is locally polynomial (usually linear, occasionally quadratic). Spline smoothers fit different polynomials in different segments of the data, where constraints are imposed to assure smoothness between segments at the knots that define segment endpoints. Various examples of spline smoothing appear in [2], [19] and [25].

When smoothing geographical data, two other linear smoothers are common: empirical Bayes smoothing [3], where the weights depend upon the **prior distribution** of the responses y_i (designed specifically for mapping disease rates – *see Disease mapping*), and **kriging** [15]. Kriging finds the best linear optimal predictor at $Y = y_i$ by estimating the parameter in a model for the covariance function of the response Y at various locations; usually this covariance function is defined in terms of the distance between these locations, and the ‘smoothness’ (i.e. local variability) of the surface is governed by this parameter [20, Chapter 4].

Another useful linear smoother is the **multivariate adaptive regression spline** (MARS) [10], wherein

the ‘smooth’ takes the form

$$\hat{f}(\mathbf{x}_i) = a_0 + \sum_m a_m \prod_j B_m(x_{ij}) \quad (2)$$

where $\mathbf{x}_i = (x_{i1}, \dots, x_{iJ})$ denotes the J components of \mathbf{x} [e.g. $J = 2$ for geographical smoothing and $\mathbf{x}_i =$ (latitude, longitude)], and the basis function $B_m(x_{ij})$ is a truncated power function of the single variable x_{ij} , $j = 1, \dots, J$ {i.e. $B_m(u) = [(u - c_0)^q]_+$, where $(y)_+ = \max[y, 0]$ }. MARS differs from typical spline smoothers in that the knots are not prespecified, but rather are found adaptively from the data themselves, and the basis functions are truncated linear functions, rather than polynomials. De Veaux et al. [8] developed a model for the spatial variations in seafloor topography that was suggested from applying the MARS smoother on data from sea ice concentrations in Antarctica. Similarly, **wavelet** smoothing proceeds by assuming that the smooth is a linear combination of some limited number of basis functions; various methods have been proposed for dictating which coefficients of these functions are nonzero and which can be omitted to yield a smooth approximation to the data [9]. MARS is particularly effective for exploratory purposes, but not when the process that generates the data is known to involve differential equations [7].

Nonlinear Smoothers

Nonlinear functions may actually be approximated fairly well by linear smoothers, as long as the function is *locally* linear and the smoother uses an appropriate span that does not smooth over nonlinear features. While linear smoothers are straightforward theoretically and are easy to implement, they can also often fail to capture the extent of interesting features: peaks are squashed, troughs are raised, and abrupt shifts (e.g. mountain ranges) appear as gradual changes; sometimes linear smoothers smooth over these features completely. (The use of truncated linear basis functions in MARS and flexible basis functions in wavelet smoothing moderates this tendency to some degree.) However, for such situations, nonlinear smoothers can be more satisfactory, as shown in the first example. The mathematical operator in such nonlinear smoothers is usually the median. Proposals for nonlinear smoothers using various combinations of spans of running medians have been

4 Smoothing

defined and illustrated for one-dimensional data by Tukey [22], Velleman [24] and Goodall [11], among others.

The electrical engineering literature has used median filters in two dimensions for image processing [17, 26]. Cressie [6] suggested a variant of this filter, called the *median polish smoother*, which does not require evenly spaced data such as pixels on a regular grid. Since the method is sensitive to orientation, Cressie recommends following the filter by kriging the residuals and adding the smoothed residuals to the result from the median polish filter. Another nonlinear smoother for two-dimensional data was proposed by Tukey and Tukey [23], but their ‘headbanging’ smoother can be shown to fail to capture linear surfaces. Both MARS and wavelet smoothing are likely to perform better when high-dimensional data involve ridges, edges or abrupt changes.

Regardless of which smoother is applied, two principles should be kept in mind. First, *some* smoothing of any form is almost always valuable. While the data are changed by smoothing (e.g. from y_i to \tilde{y}_i), one can argue that they are changed only slightly, and that the potential benefits (reduced noise, insight into functional relationships, etc.) far outweigh this concern. Second, process knowledge should always be incorporated first into any fitting strategy. The main strength of smoothers is their ability to suggest functional forms when such knowledge may not be available, or on the residuals from a previously

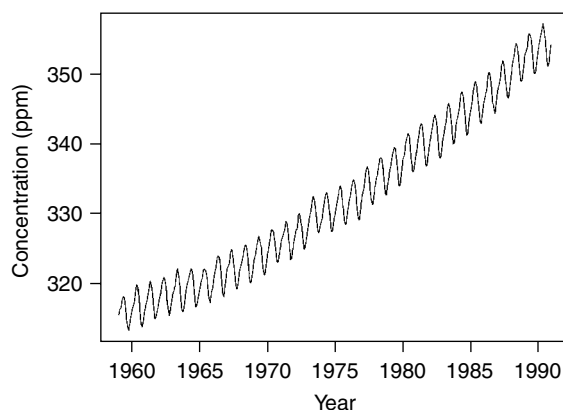


Figure 2 Monthly carbon dioxide concentrations in Mauna Loa volcano, from January 1959 to December 1990. Data from S-PLUS statistical software [16]

identified, possibly lower-order model (as illustrated by Example 1 below).

Example 1 Carbon dioxide in Mauna Loa.

Figure 2 shows monthly concentrations (ppm) of carbon dioxide in the Mauna Loa (Hawaii) volcano, from January 1959 to December 1990. The data were collected by the Scripps Institute of Oceanography in La Jolla, CA, and are described in the **S-PLUS** documentation [16]. The most obvious trends are the increase over time (mostly linear, but with a slight curvature) and regular yearly cycles. After first removing a quadratic trend and the month

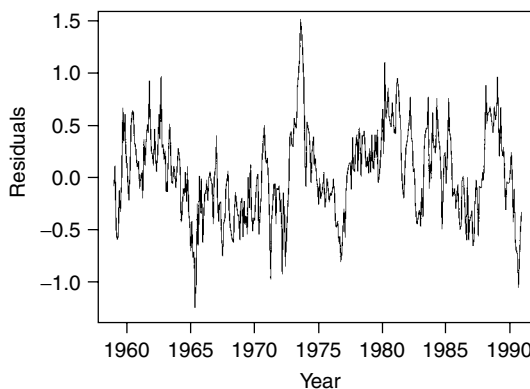


Figure 3 Residuals from fitting a quadratic trend and month effect to the data in Figure 2

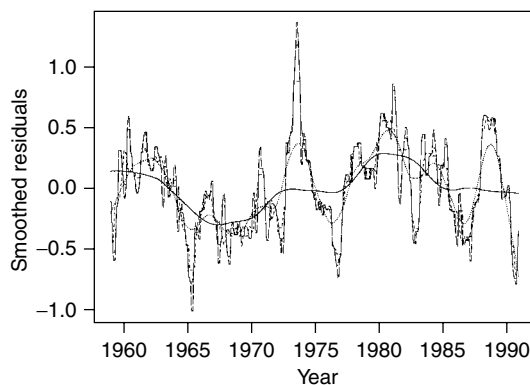


Figure 4 Five smoothers of the residuals shown in Figure 3. In order of smoothness: running medians of length 3 (least smooth), running medians of length 5, running medians of length 11, loess with span 0.10, loess with span 0.25 (most smooth)

effects [i.e. subtract a quadratic function of (year – 1975) and the January, February, etc., average from all January, February, etc., months], the resulting residuals are shown in Figure 3. Five smoothers of these residuals are shown in Figure 4; they all suggest

some elevated concentrations (above and beyond the overall increase and monthly effects) in 1960–1965, 1973–1974 and 1978–1985. Running medians of length 3 are the least smooth, but also capture best the large peak in August 1973 and January 1988 to

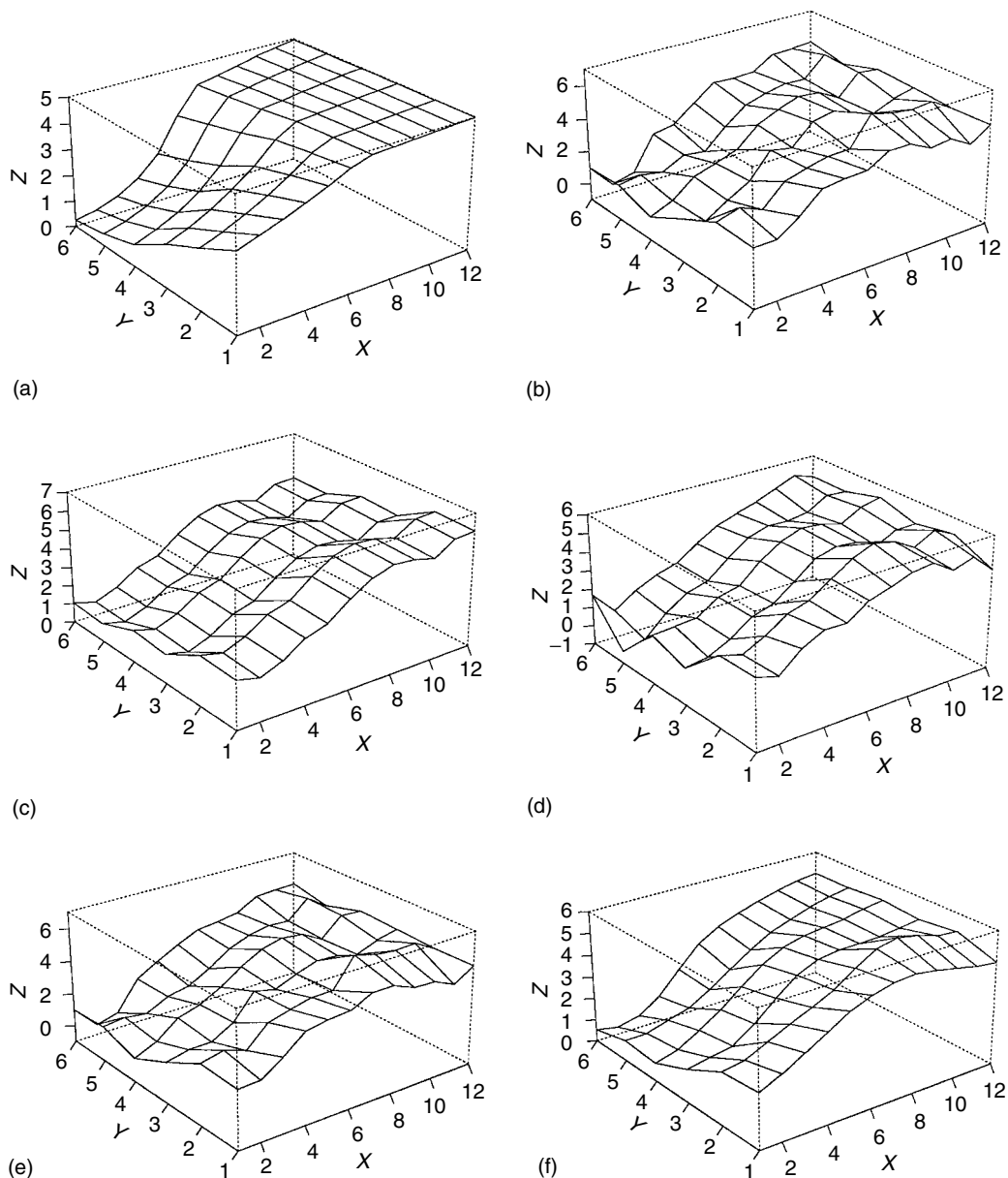


Figure 5 Nonlinear smoothers to two-dimensional data. (a) Smooth surface with no error. (b) Smooth surface with additive $N(0, 0.01)$ error. (c) Median polish fit to the data in panel (b). (d) Cressie's median polish smoother to the data in panel (b). (e) Loess smoother (span 0.10) on the data in panel (b). (f) Loess smoother (span 0.30) on the data in panel (b)

July 1989: as the span of this smoother increases to 5 and 11, the peaks become squashed. The two smoothest curves are loess with span 0.10 and loess with span 0.25; the former succeeds in capturing the general shape of the residuals over time but severely attenuates the more interesting features. The tradeoff between bias and variance in this example is well illustrated.

Example 2 Two-dimensional data. Nonlinear smoothers are not always so successful, especially in higher dimensions where measurement error is negligible. Figure 5(a) shows a smooth surface without error; normal (0, 0.01) error has been added to this surface in Figure 5(b). A median polish fit is shown in Figure 5(c); notice that it fails to capture well the nearly constant surface at values of $x > 8$. Cressie's median polish smoother [6] (which takes the result of Figure 5(c) and adds smoothed residuals) appears in Figure 5(d); the smoother recaptures the trend at the upper end of x but still suggests undulations that did not exist in the true underlying surface. In contrast, the loess smooths tend to reproduce the surface more faithfully, especially when the span lies in the range 0.10 (Figure 5e) to 0.30 (Figure 5f). Objective criteria for choosing the span have been proposed (e.g. **cross-validation**), but most people still prefer to choose the span 'by eye'.

References

- [1] Bowman, A.W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, London.
- [2] Chen, L. (1997). Multivariate regression splines, *Computational Statistics and Data Analysis* **26**, 71–82.
- [3] Clayton, D. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**, 671–682.
- [4] Cleveland, W.S. & Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting, *Journal of the American Statistical Association* **83**, 596–610.
- [5] Cressie, N. (1986). Kriging nonstationary data, *Journal of the American Statistical Association* **81**, 625–634.
- [6] Cressie, N. (1993). *Statistics for Spatial Data*, Revised Edition, Wiley, New York.
- [7] De Veaux, R.D., Bain, R. & Ungar, L.H. (1999). Hybrid neural network models for environmental process control, *Environmetrics* **10**, 225–236.
- [8] De Veaux, R.D., Gordon, A.L., Comiso, J.C. & Bacherer, N.E. (1993). Modeling of topographic effects on Antarctic sea ice using multivariate adaptive regression splines, *Journal of Geophysical Research* **98**, 20307–20319.
- [9] Donoho, D.L. & Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association* **90**, 1200–1224.
- [10] Friedman, J.H. (1991). Multivariate adaptive regression splines, *The Annals of Statistics* **19**, 1–141.
- [11] Goodall, C.R. (1991). A survey of smoothing techniques, in *Modern Methods of Data Analysis*, Chapter 3, J. Fox & J.S. Long, eds, Sage, Beverly Hills, pp. 126–176.
- [12] Hall, P. (1998). Binning, in *Encyclopedia of Statistical Sciences*, Update Vol. 2, S. Kotz, C.B. Read & D.C. Banks, eds, Wiley, New York, pp. 64–65.
- [13] Hastie, T. & Loader, C. (1993). Local regression: automatic kernel carpentry (with discussion), *Statistical Science* **8**, 120–143.
- [14] Mallows, C.L. (1980). Some theory of nonlinear smoothers, *The Annals of Statistics* **8**, 695–715.
- [15] Mathéron, G. (1963). Principles of geostatistics, *Economic Geology* **58**, 1246–1266.
- [16] Mathsoft (1993). *S-Plus*, Unix Version.
- [17] Narendra, P.M. (1981). A separable median filter for image noise smoothing, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **3**, 20–29.
- [18] O'Sullivan, F. (1988). Robust smoothing, in *Encyclopedia of Statistical Sciences*, Vol. 8, S. Kotz, N.L. Johnson & C. Read, eds, Wiley, New York, pp. 170–173.
- [19] Ramsey, J.O. (1988). Monotone regression splines in action (with discussion), *Statistical Science* **3**, 424–461.
- [20] Ripley, B. (1981). *Spatial Statistics*, Wiley, New York.
- [21] Simonoff, J.S. (1996). *Smoothing Methods in Statistics*, Springer-Verlag, New York.
- [22] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading.
- [23] Tukey, P.A. & Tukey, J.W. (1981). Graphic display of data sets in 3 or more dimensions, in *Interpreting Multivariate Data*, V. Barnett, ed., Wiley, Chichester, 1981, pp. 189–275. [Reprinted in Cleveland, W.S. (ed.) (1988), *The Collected Works of John W. Tukey*, Vol. V: *Graphics, 1965–1975*, Wadsworth, Belmont, pp. 188–288.]
- [24] Velleman, P.F. (1980). Definition and comparison of robust nonlinear data smoothing algorithms, *Journal of the American Statistical Association* **75**, 609–615.
- [25] Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.
- [26] Yang, C.J. & Huang, T.S. (1981). The effect of median filtering on edge location estimation, *Computer Graphics and Image Processing* **15**, 224–245.

Further Reading

Two particularly useful books on smoothing are Simonoff [21] and Bowman and Azzalini [1]; the former emphasizes applications of smoothing for density estimates, while the latter

emphasizes more general smoothing applications as described in this article. Both concentrate on one-dimensional smoothing. These books, as well as the entry on **Nonparametric regression model**, also discuss the issue of selecting a bandwidth. Ripley [20] and Cressie [5] offer more methods and examples for higher-dimensional data. O'Sullivan [18] presents a discussion of robust smoothing.

(*See also* **Generalized additive models; Kernel estimator by Fourier transform; Robust regression; Smoothing in environmental epidemiology; Time series**)

KAREN KAFADAR & PAUL S. HORN