Journal of Epidemiology and Biostatistics (2001) Vol. 6, No. 6, 433–443

Data visualisation for time series in environmental epidemiology

B ERBAS[1] and RJ HYNDMAN[2]
[1] *Department of Public Health, The University of Melbourne, Australia*
[2] *Department of Econometrics and Business Statistics, Monash University, Australia*

Background Data visualisation has become an integral part of statistical modelling.

Methods We present visualisation methods for preliminary exploration of time-series data, and graphical diagnostic methods for modelling relationships between time-series data in medicine. We use exploratory graphical methods to better understand the relationship between a time-series reponse and a number of potential covariates. Graphical methods are also used to examine any remaining information in the residuals from these models.

Results We applied exploratory graphical methods to a time-series data set consisting of daily counts of hospital admissions for asthma, and pollution and climatic variables. We provide an overview of the most recent and widely applicable data-visualisation methods for portraying and analysing epidemiological time series.

Discussion Exploratory graphical analysis allows insight into the underlying structure of observations in a data set, and graphical methods for diagnostic purposes after model-fitting provide insight into the fitted model and its inadequacies.

Keywords additive models, environmental epidemiology graphical methods, nonparametric smoothing seasonality.

Introduction

Graphics can be used to uncover and present complex trends and relationships in an informal and simple way[1], enabling the researcher to better understand the underlying structure of the data[2]. Time-series data consist of observations equally spaced through time e.g. daily, monthly, or quarterly observations. In this paper we explore the use of graphical methods to improve understanding of the underlying variation within and between time-series data.

To examine the underlying structure of the pattern within a time series, we will discuss time-series plots with a nonparametric smooth-curve superimposed, and sub-series plots for studying seasonal patterns. In addition, we will present decomposition plots, to detect the magnitude and strength of the seasonal and trend components within each series, and scatterplot matrices, to examine the relationships between various time series. We will consider coplots as a method suitable for identifying potential interactions between two time series.

We will also discuss partial residual plots, as a diagnostic graphical tool useful for identifying nonlinearities in the covariates for generalised linear models, and for examining the nonlinear relationship between the response and the covariates for generalised additive models. A natural feature of time-series data is the serial correlation within the series. We will discuss plots of the autocorrelation and partial autocorrelation functions, to examine the underlying correlation structure of the residuals from a particular model.

Correct specification of the relationship between the response and the covariates, and identifying remaining autocorrelation in the residuals, are major methodological issues in daily mortality/morbidity research[3,4]. The data-visualisation methods we present will assist the investigator in resolving these issues.

None of the methods we discuss are original. In recent years, they have come to be used in diverse fields, especially in the analysis of business and economic time-series data. We provide a link with this literature, and demonstrate how the techniques are transferable to epidemiological time-series data.

As an illustration, we will use time-series data from a study of asthma hospital admissions in Melbourne, Australia. Asthma hospital admissions from all short-stay acute public hospitals in Melbourne, registered on a daily basis by the Health Department of Victoria, was used as the response variable for the period 1 July 1989 to 31 December 1992. International Classification of Disease (ICD) code (493) was used to define asthma.

Air pollution data were obtained from the Environmental Protection Authority (EPA). Maximum hourly values were averaged each day across nine monitoring stations in Melbourne, for nitrogen dioxide (NO_2),

Correspondence to: B. Erbas, Department of Public Health, The University of Melbourne, VIC 3010 Australia.

Received 30 August 2000
Revised 14 February 2001
Accepted 29 August 2001

sulphur dioxide (SO_2), and ozone (O_3), all measured in parts-per-hundred-million (pphm). Particular matter was measured by a device that detects back-scatterng of light by visibility-reducing particulates 0.1–1 μm in diameter. Particulates were derived from $B_{scat} \times 10^{-4}$ and the resulting variable is denoted by API.

Meteorological data were obtained from the Commonwealth Bureau of Meteorology. Three hourly maximum daily levels of relative humidity (hu), wind speed and dry bulbs temperature (db) were averaged across four monitoring stations in the Melbourne area.

To simplify the analysis of seasonality, we excluded the leap day of 29 February 1992 in each series.

Exploratory graphical methods

Time-series plots and smooth curves

A time-series plot is a commonly used graphical tool useful for examining the underlying long-term pattern of a time series. Time (days, months, years, etc.) is plotted on the horizontal axis, and the series of interest is plotted on the vertical axis. Although these plots provide a visual insight into the long-term trend, they do not always allow a clear visual understanding of the fluctuations of each observation from one time-period to the next. The trend can be reinforced, and we can visually examine specific time-periods by supplementing the plot with a nonparametric smooth trend-curve[5–9]. The smooth curve can be estimated in a number of different ways, the most commonly used being loess smoothers[6], kernel smoothers[7] and cubic smoothing splines[8]. A brief description of these smoothers is provided in the Appendix.

A time-series plot of daily counts of asthma hospital admissions is displayed in Figure 1. Although we observe various fluctuations in the trend throughout the series, it is difficult to visually determine the pattern of variation between time periods. We fitted a loess smooth[6], with span = 10%, to the daily counts of asthma admissions and superimposed the smooth curve onto the time-series plot in Figure 1. The result is shown in Figure 2 and enables a clearer visual interpretation of the long-term pattern in asthma admissions. For example, there is evidence of a decrease in admissions in December–January each year; this decrease is less pronounced in 1990–91 and almost nonexistence at the end of 1992. There is a sharp increase in admissions in autumn for 1990 and 1992, but no evidence of this increase in 1991.

Seasonal subseries plots

Seasonality is a strong confounder in the analysis of daily hospital admissions/mortality data[3,10]. It is important, therefore, to identify the strength and magnitude of the seasonal component in a time series. We can construct a seasonal subseries plot to assess the

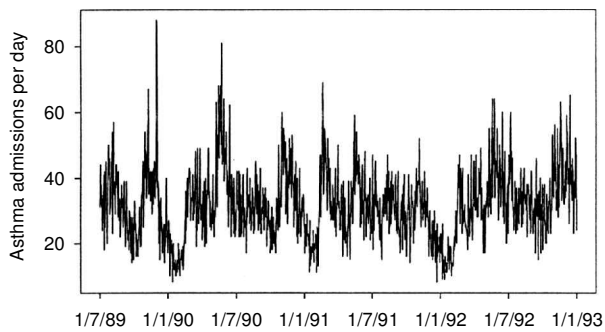

Fig. 1 A time-series plot of daily counts of asthma hospital admissions in Melbourne, Australia from 1 July 1989 to 31 December 1992.

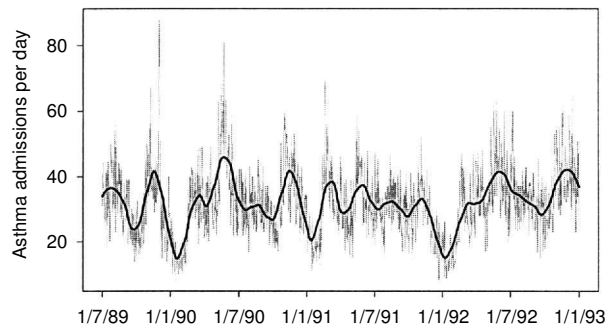

Fig. 2 A loess smooth of daily counts of hospital admissions for asthma in Melbourne, Australia, from 1 July 1989 to 31 December 1992. (Smoothing span set to 10%.)

behaviour of each monthly subseries. A subseries plot is constructed by plotting January values for successive years, then February values, and so on up to values for December. For each month, the mean of the values is drawn by a horizontal line, and vertical lines emanating from these horizontal lines are values for each subseries. We can assess the overall seasonal pattern of the series by the means of each subseries and we can assess the variation within each subseries by the vertical lines[11,12] (Figure 3).

Figure 3 shows this style of plot (a) hu and (b) db in Melbourne, Australia from 1 July 1989 to 31 December 1992. Note there are only three vertical lines emanating from each mean from January to June for both series. This is because both series began in July, rather than January, for the study period. The subseries plot for hu in (a) shows a large variation in the means for each month, and the vertical lines emanating from each mean is indicative of a large variation within each month. The overall pattern in (a) is one in which June is the highest month for hu. The monthly variation for db in (b) is far more stable than the monthly variation in (a). The overall pattern for db shows a clear minimum in July and

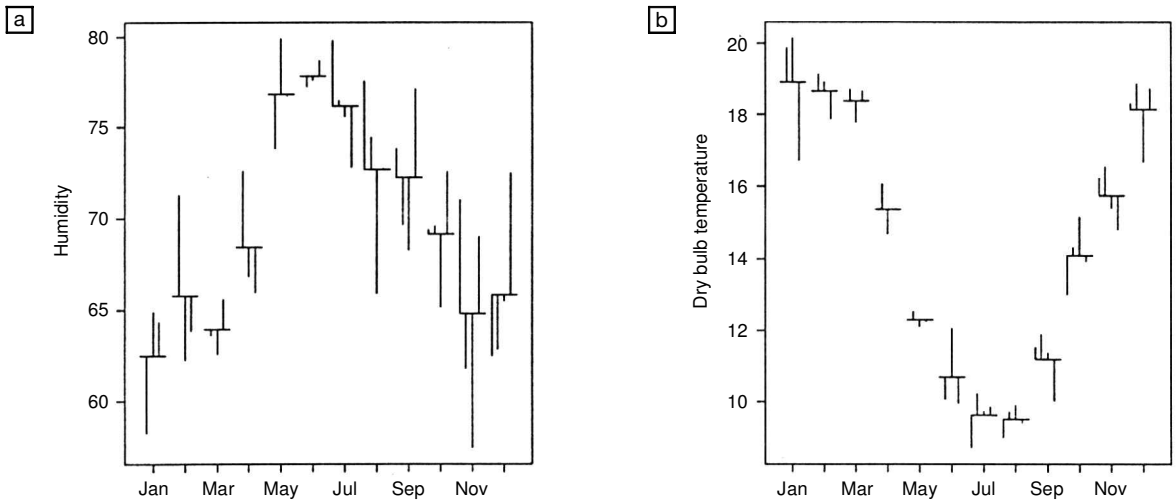

Fig. 3: (a) *A subseries plot for relative humidity. The variation in the subseries is large, as are the variations in the means for each month.* ***(b)*** *A subseries plot for dry bulb temperature. The monthly variation is smaller than the monthly variation in (a).*

August, and a relatively level period from December through March. Although there is some variation within each month, particularly January, the variation in each subseries is relatively small compared with the overall pattern in the monthly series.

Decomposition plots

A common approach to time-series in business and economics is to consider them as a mixture of several components[13]: trend, seasonality and an irregular random effect. It is important to identify the trend and seasonal components, and remove them from the time series when modelling relations between time-series. When this is not done, highly seasonal series can appear to be related, purely because of their seasonality rather than because of any real relationship. Similarly, trended series can exhibit spurious collinearity. Consequently, the estimation of trend and seasonal terms, and the calculation of adjusted series removing their effect, are important issues in any time-series analysis[12].

Time-series decomposition methods allow an assessment of the strength of the seasonal component in each of the pollutants and climatic variables. After identification, the seasonal component is removed and the resultant seasonally adjusted series is used in subsequent analysis. Extracting the seasonal component thus allows a clearer picture of other characteristics of the data.

A number of time-series decomposition methods are available. Classical decomposition[12] is a relatively simple method but has several drawbacks, including bias problems near the ends of the series and an inability to allow a smoothly varying seasonal component. To overcome these difficulties we adopt the seasonal trend decomposition procedure based on loess (STL) method[14].

We assume an additive decomposition:

$$Y_t = T_t + S_t + E_t \qquad (1)$$

where Y_t denotes the time-series of interest, T_t denotes the trend component, S_t denotes the seasonal component and E_t denotes the remainder (or irregular) component. The seasonally adjusted series, Y_t^* is computed simply by subtracting the estimated seasonal component from the original series, $Y_t^* = Y_t - \hat{S}_t$.

STL consists of a sequence of applications of the loess smoother, to provide robust estimates of the components T_t, S_t and E_t from Y_t[12]. The STL method involves an iterative algorithm, to progressively refine and improve estimates of trend and seasonal components. After appropriate seasonal adjustment of the pollutants and climatic variables using STL, we construct decomposition plots.

A decomposition plot normally consists of four panels: the original series, the trend component, the seasonal component and the irregular (or random) component. The panels are arranged vertically, so that time is a common horizontal axis for all panels[11]. To illustrate the usefulness of this method we display the decomposition of the asthma admissions data in Figure 4. Where we have applied the STL method[14] to compute the trend and seasonal components of this series. These are shown in the second and third panels respectively. In effect, we are separating the smooth curve estimated in Figure 2 into two components, trend and seasonality. Note that the seasonal component captures the drop in January and the increase in autumn each year. The trend component in Panel 2 makes it clear that the trend is increasing near the end of the series, thus explaining the smaller

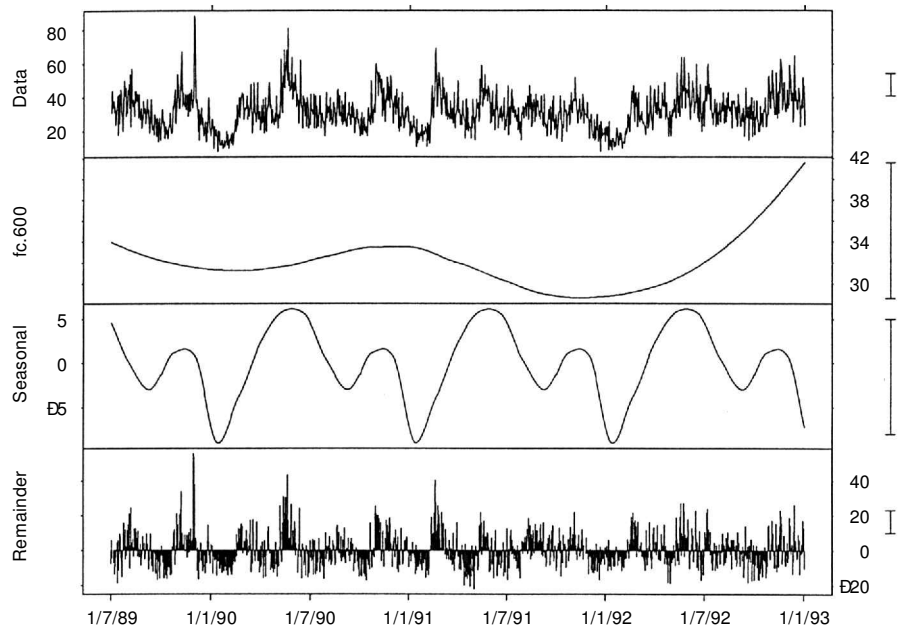

Fig. 4 *The original daily asthma admissions are displayed in the top panel. The three components are displayed on the other panels. In this case, the trend and seasonal components do not account for much of the variation.*

decrease than was expected in admissions in December 1989.

The bottom panel displays what remains when the trend and seasonal components are removed from the data. This allows unusual days or periods to be determined without the confounding that occurs because of seasonality or trend. The bars on the right of each panel depict the amount of variation in the data and components. These are all the same length, but plotted on different scales. Thus, a decomposition plot can be used to assess the strength and magnitude of the trend, and seasonal components within a series. In this example, the length of the bar to the right of the third panel indicates that a small amount of variation in the original series is accounted for by the seasonal component.

Scatterplot matrices

A scatterplot is a useful graphical tool that allows a visual examination of the possible relationship between two series. However, its usefulness is limited to two dimensional space[15]. The scatterplot matrix, first described in Chambers, et al.[16] (1983), is a simple graphical tool for displaying pairwise scatterplots for three or more variables. Visually, it represents multi-dimensional space, thus allowing insight into the complex relationship between multiple series. The graphs are arranged in a matrix with a shared scale, so that each panel of the matrix is a scatterplot of one variable against another variable. If there are k variables, there are $k(k-1)$ panels in total, and each pair of variables is plotted twice. The scatterplot matrix is symmetric, in that the upper left triangle consists of all $k(k-1)/2$ pairs, as does the lower right triangle[15].

Scatterplot matrices can be especially useful in selecting covariates for modelling and detection of multi-collinearity[5]. They can also be useful in detecting a nonlinear bivariate relationship between a response variable and a potential covariate in modelling. Scatterplot matrices provide a clear sense of the underlying patterns in the relationships, and allow multiple comparisons on a single plot[5].

A pairwise scatterplot matrix for daily counts of asthma hospital admissions, pollutants and climatic variables is shown in Figure 5. This clearly exhibits a nonlinear bivariate relationship between asthma admissions and the climatic variables dry bulbs temperature, and relative humidity. This is consistent with results reported in previous studies[17–19]. Collinearity is an important methodological issue in the analysis of daily counts of morbidity/mortality[20]. Figure 5 allows a visual insight of the inherent collinearity between db and hu, NO_2 and wind-speed. Although a scatterplot matrix allows us to visually detect collinearity, the extent of collinearity should be determined by a formal test[21].

Coplots

When modelling relationships between a response variable and two or more explanatory variables, we are often interested in the interaction between the

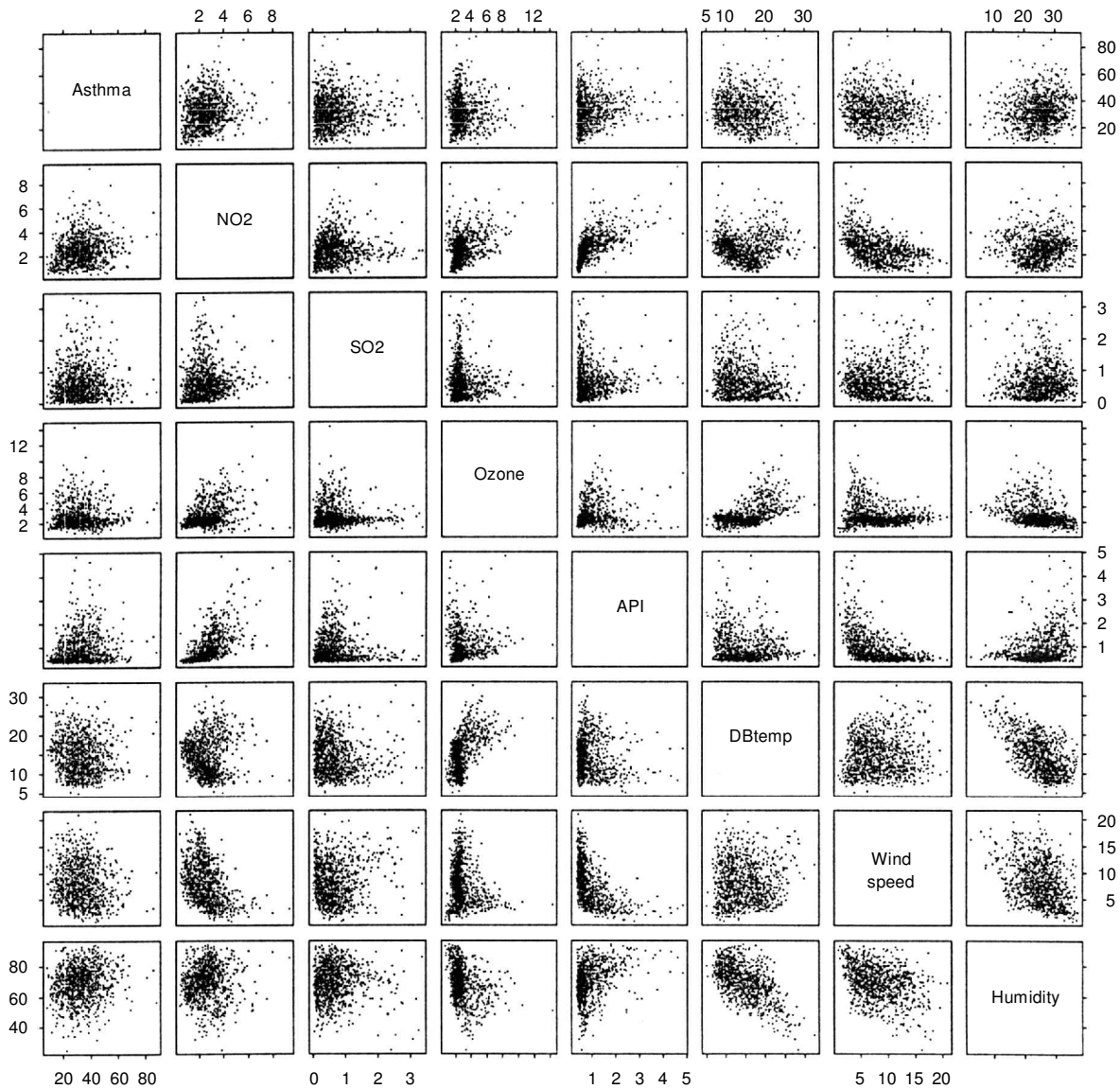

Fig. 5 *Pairwise scatterplots for asthma hospital admissions, pollutants and climatic data.*

explanatory variables. Interaction is a condition where the levels of the response variable are different at various levels of the explanatory variables[21]. To study how a response variable depends on two explanatory variables we can use a coplot (conditioning plot), as introduced by Cleveland[22] (1993) (Figure 6).

One of the explanatory variables act as the 'given series'. A coplot consists of a top panel, the 'given' panel (each point represents a value for the given series) and a bottom panel, the 'dependence' panel. The latter comprises a series of scatterplots of the response variable with the other explanatory variable for those observations, whose values of the given series are equal to each point in the given panel. (For a given series with many unique values, the plot is constructed based on intervals of the given series, rather than at unique values.)

Smooth curves and confidence intervals (CI) can be added to each dependence panel to detect patterns in the bivariate relationship for each panel. A coplot is useful in detecting the presence of interaction. If the panels exhibit relationships of different shapes, then this is indicative of an interactive effect between the explanatory variables on the response.

Figure 6 shows a coplot where each panel demonstrates the relationship between asthma admissions and db, where the data are conditioned on the month of observation. Smoothing spline curves and 95% point-

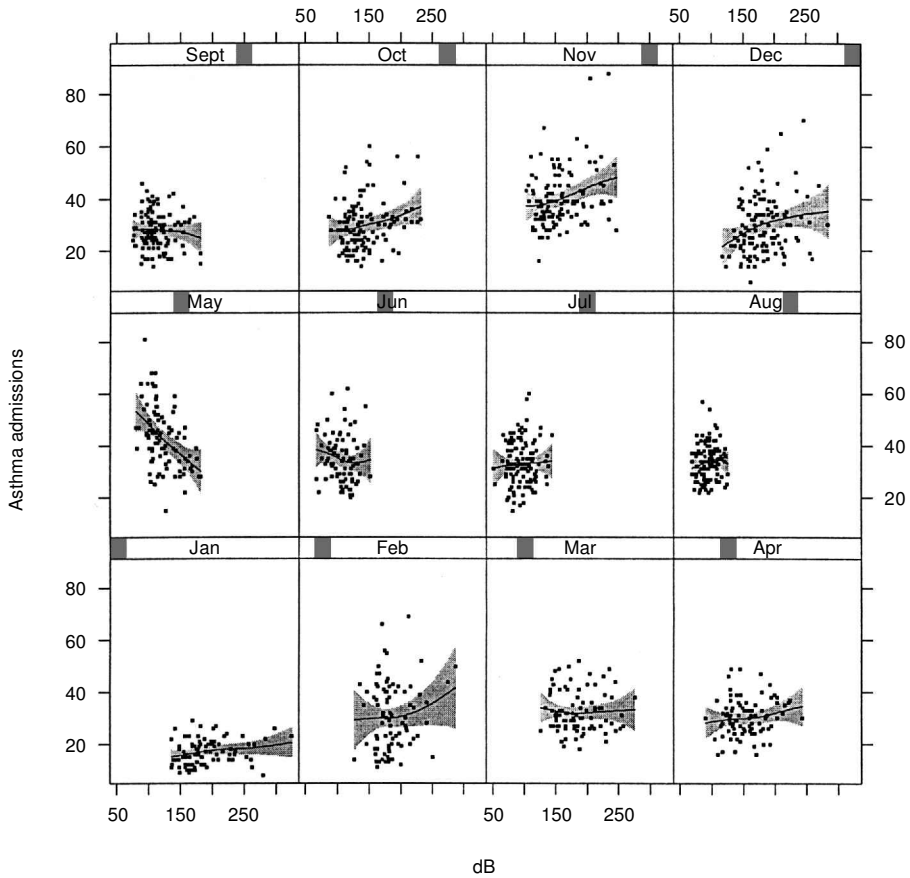

Fig. 6 *A coplot of asthma admissions against db given the month of observation.*

wise CI are added to enable easier visualisation of the changing relationship throughout the year. Note that the relationship is negative for May, but positive in the hotter months of November through February. In the winter months of June–September, there is little relationship at all.

Diagnostic graphical methods

Graphical analysis of residuals are central to the process of statistical model building[10]. These plots can be used to identify any undetected tendencies in the data, outliers, and homogeneity in the variance of the response variable[1]. In addition, a variety of exploratory graphical methods are available to assess the nonlinearity between a response variable and an explanatory variable in statistical modelling[23]. This paper discusses partial residual plots as an exploratory diagnostic graphical method and demonstrates diagnostic plots after fitting a generalised linear model (GLM) and generalised additive model (GAM) to asthma hospital admissions in Melbourne, Australia from 1 July 1989 to 31 December 1992.

GLM and GAM

Preliminary exploratory analysis of asthma hospital admissions exhibited strong long-wave length patterns. To adequately model these patters we use Fourier series terms of $\sin(2\pi jt/365)$ and $\cos(2\pi jt/365)$ for $j = 1, \ldots, 10$. Fourier terms of sine and cosine pairs for $j > 10$ were not included in the analysis because neither the sine nor cosine pair was statistically significant ($p < 0.05$) when included in the model. Also, the inclusion of sine and cosine pairs for $j > 10$ did not contribute to a reduction in the AIC[24] goodness-of-fit statistic.

Dummies are used to model day-of-week patterns using a parameterisation based on Helmert contrasts[25]. Five-day lagged pollutants and db and hu are also included as potential covariates. Lag combinations that contribute to the largest decrease in AIC are included in the final model. Step-wise procedures are used to evaluate the contribution of each covariate and determine the GLM and GAM that best describes asthma hospital admissions. For this analysis, seasonally adjusted NO_2, O_3 and seasonally adjusted hu and db are included in the analysis.

GLM

For a GLM we estimate the expected value of the response variable, as a function of a linear combination of covariates:

$$E(Y_t|X_t)=\exp\{\beta_0+\beta_1 t+\beta_2 t^2+\beta_3 NO_{2,t}+\beta_4 O_{3,t-1}+\beta_5 db_t+\beta_6 db_{t-1}+\beta_7 db_{t-2}+\gamma_t+\sum_{j=1}^{10}[a_j \sin(2\pi jt/365)+\varphi_j \cos(2\pi jt/365)]\} \quad (2)$$

where X_t denotes the vector of covariates, t denotes the day of observation (1, . . ., 1279), $NO_{2,t}$ denotes the seasonally adjusted NO_2 measurement, $O_{3,t}$ denotes the seasonally adjusted O_3 measurement and db_t denotes the seasonally adjusted db measurement, all on day t. The day-of-week effect is captured using the variable γ_t, which takes on a different value for each day of the week. We assume Y_t is pseudo-Poisson (i.e. Poisson with overdispersion), conditional on the values of the covariates.

GAM

For a GAM, we adopt a similar modelling framework, except the linear terms in a GLM may be replaced by smooth nonlinear functions that are estimated nonparametrically. The following model was fitted:

$$E(Y_t|X_t)=\exp\{\beta_0+g_1(t)+\beta_2 NO_{2,t}+\beta_3 NO_{2,t-1}+\beta_4 O_{3,t}+g_5(O_{3,t-2})+\beta_6 db_t+g_7(db_{t-1})+\beta_8 db_{t-2}+\gamma_t+[\sum_{j=1}^{10} a_j \sin(2\pi jt/365)+\varphi_j \cos(2\pi jt/365)]\} \quad (3)$$

where t denotes the day of observation (1, . . ., 1279), $NO_{2,t}$ denotes the NO_2 measurement, $O_{3,t}$ denotes the O_3 measurement and db_t denotes the db measurement, all on day t. The summation term involving sins and cos functions models the seasonal pattern throughout the year using a Fourier series approximation. The day of week effect is captured using the variable γ_t, which takes on a different value for each day of the week. Note that $O_{3,t-2}$, db_{t-1} and t are modelled using smooth nonparametric functions; the other variables are all modelled linearly. The choice of a smooth or linear term for each covariate was made using the AIC. We assume Y_t is pseudo-Poisson (i.e. Poisson with overdispersion) conditional on the values of the covariates. The choice of smoothing constant for the smooth terms (g_1, g_5 and g_7) was determined by setting the equivalent degrees of freedom for these terms to 4. While this choice is relatively arbitrary, we felt it provided sufficient flexibility to model any non-linearity in the series without allowing the smooth term to be unnecessarily affected by individual observations.

Partial residual plots

Partial residual plots are useful in detecting nonlinearities and for identifying the possible cause of unduly large observations[26] in a generalised linear model and a generalised additive model. The partial residuals for a variable X_j in an ordinary linear regression model are the same as the ordinary residuals except the jth term is added to each value:

$$d_t^j = \hat{\beta}_j X_{t,j} + (Y_t - \mu_t) \quad (4)$$

where μ_t denotes the fitted value for observation t. This is equivalent to adjusting the observation Y_t for every variable in the model except for the jth term. For GLMs or GAMs, we use partial deviance residuals instead, which are defined analogously[8]. Other possibilities are to use partial Pearson or partial quantile residuals[27]. To detect nonlinearities more easily from a partial residual plot, one may superimpose a smooth curve to enhance the display and allow easier detection of a nonlinear relationship between the response variable and a covariate.

We use partial residual plots as a diagnostic graphical method to detect nonlinearities in generalised linear models, and to better understand the relationship between the response variable (asthma hospital admissions) and the explanatory variables (pollution and climate) in generalised additive models.

To demonstrate the diagnostic usefulness of a partial residual plot we present results from the GLM analysis of the time trend on asthma hospital admissions. Figure 7 shows the partial residual of the time

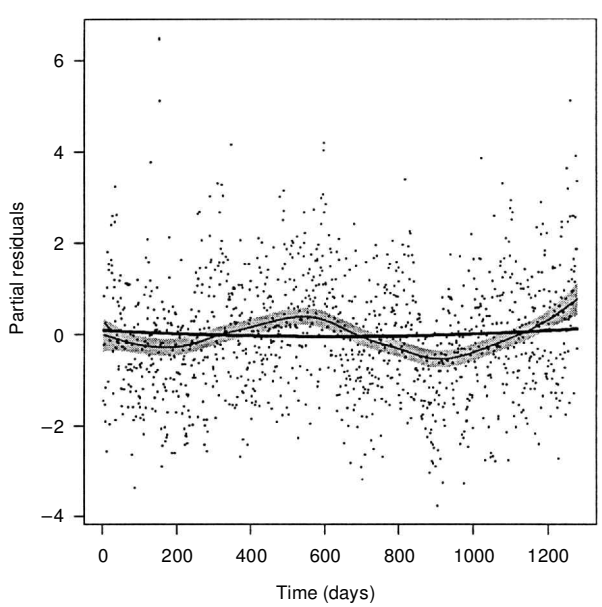

Fig. 7 *Partial deviance residuals for the time trend from the GLM. The fitted quadratic time trend is shown as a thick line. A smoothing spline has been fitted to the partial deviance residuals and is shown as a fine line. The shaded area represents 95% pointwise CI around the spline curve.*

covariate from a GLM analysis with a quadratic term fitted for the time covariate. We fit smoothing spline to these residuals to enable a clearer examination of the nonlinear relationship between time and asthma admissions. The shaded area represents 95% pointwise CI around the fit. The plot displays a striking nonlinear relationship between time and asthma admissions that has not been adequately captured by the quadratic fit in the GLM.

Figure 8 shows the smoothing spline estimates for the time trend, ozone term and db term in the GAM model. The seasonal term constructed from the Fourier approximation is also shown. The time trend is clearly more complicated than a simple quadratic. The seasonal structure is also quite complex, as would be expected with 10 Fourier terms. The ozone and db effects are significantly non-linear (hence the use of smooth functions in the model), but only near the centre of the range of observations is the function estimated with any degree of precision (shown by the wide confidence bands elsewhere).

Autocorrelation and partial autocorrelation plots

A characteristic feature of time series is that the observations are ordered through time; a consequence of this is autocorrelation, that is an underlying pattern between observations from one time period to the next within a time series. We can define autocorrelation at lag k as the correlation between observations k time-periods apart. This can be estimated for lags $k = 0, 1, 2, \ldots$, as:

$$r_k = \frac{\sum_{t=k+1}^{n} (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n} (x_t - \bar{x})^2} \tag{5}$$

where x_t denotes the observation at time t and $\bar{x}$ denotes the sample mean of the series $\{x_t\}$.

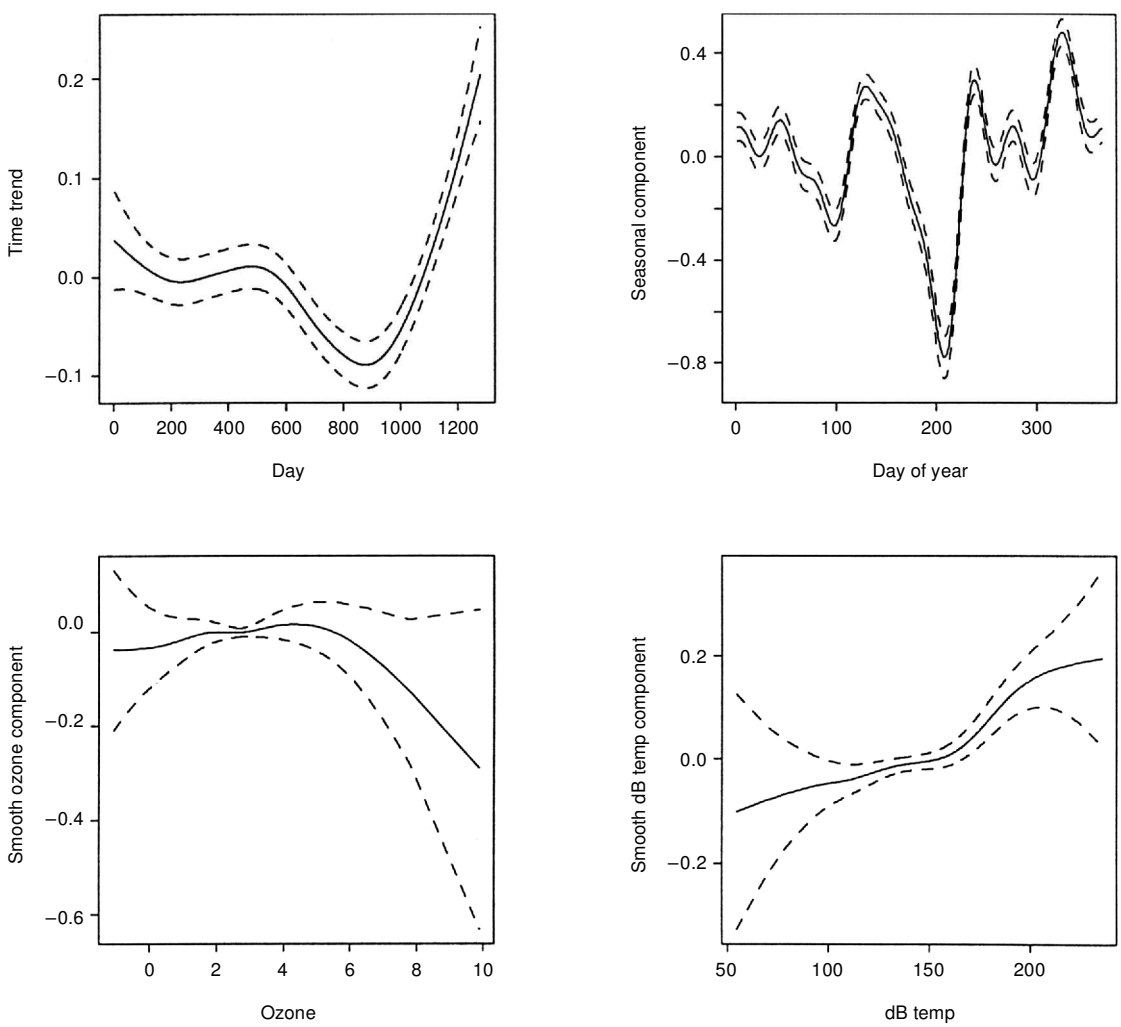

Fig. 8 *Smooth terms from the GAM. Dashed lines show 95% pointwise CI.*

The partial autocorrelation at lag k is defined as the partial correlation between observations k time-periods apart. That is, it is the correlation between x_t and x_{t-k} after removing the effect of the intervening observations $x_{t-1}, \ldots, x_{t-k+1}$. It can be estimated using a fast computational algorithm[28].

The collection of autocorrelations at lags $k = 1, 2, \ldots$ is known as the autocorrelation function (ACF). Similarly, the collection of partial autocorrelations is known as the partial autocorrelation function (PACF). To help visualise the temporal pattern in a time series, we plot the ACF and PACF of the series.

Figure 9 displays ACF and PACF plots of the residuals from the GAM model fitted to the asthma admissions data. The dashed lines denote 95% critical values for each autocorrelation. Any spike lying outside these bounds shows a correlation that is significantly different from zero. The graph shows there is strong autocorrelation remaining in the residuals, for which allowance needs to be made in the model. It is possible that a low order autoregressive term will adequately capture this pattern[3]. GAM with autocorrelation are currently being developed by the authors.

A related visual diagnostic tool is the cross-correlation function[28], which is useful for examining the lag-effect of a change in pollution on a change in the response variable[29].

Discussion

We have presented some simple data-visualisation methods that allow an effective graphical examination of quantitative information. Most exploratory graphical methods fall into two categories: displaying the original data, or displaying quantities associated with fitted models. Methods in the first category method are used to explore the data and those in the second are used to enhance statistical analysis that are based on assumptions about relations in the data[26].

We have reviewed exploratory graphical methods for time-series analysis in environmental epidemiology, and have applied time-series plots, subseries plots, decomposition plots, scatterplot matrices and coplots, to enable a visual assessment of long-term trend, seasonality and nonlinearity in time-series data. We have also applied partial residual plots and ACF and PACF residual plots as diagnostic methods to assess nonlinearities in explanatory variables, and to assess autocorrelation patterns in the residuals from a fitted model.

We found that a smooth plot superimposed on a time-series plot provided a more enhanced visual representation of the long-term pattern and short-term variation between time periods in a time series. A seasonal subseries plot was used to assess the variation within each month, and between successive months in a time series. This allows visual insight into the seasonal behaviour of a time series. Decomposition plots provide understanding of the underlying trend-cycle, seasonal and random components of a time series, and their relative strengths.

Scatterplot matrices are particularly useful for detecting collinearity betwen potential explanatory variables, and nonlinearity between the response and potential explanatory variables. They provide an immediate visual insight into the complex relations between each series.

Coplots were used to assess the interactive effects of pollution and climate on asthma admissions. These plots are particularly useful for detecting interactive effects between explanatory variables.

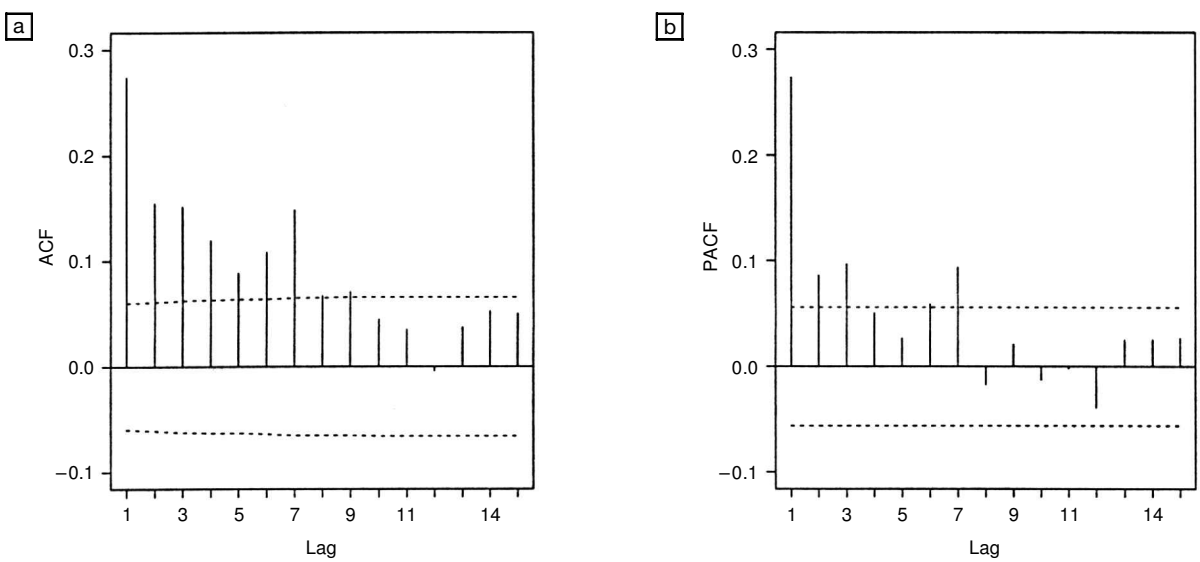

Fig. 9 *The autocorrelation* ***(a)*** *and partial autocorrelation* ***(b)*** *function for the residuals from a GAM.*

Partial residual plots are useful in detecting nonlinear relationships between the response and covariates and in understanding the effects of each covariate on the response variable. Plots of the autocorrelation and partial autocorrelation function allowed a visual insight into any remaining correlation structure in the residuals.

We have demonstrated that data visualisation is an integral part of statistical modelling. Many of the diagnostic graphical methods presented here are not specific to time series, they can be used in most applications of statistical methodology. Methodological issues in the analysis of morbidity/mortality have been an important issue for many decades, the inclusion of exploratory graphical tools for exploratory and diagnostic purposes is long overdue.

References

1 Toit SHC, Steyn AGW, Stumpf RH. *Graphical exploratory data analysis* New York: Springer-Verlag, 1986.

2 Tukey JW. *Exploratory data analysis*. Massachusetts: Addison-Wesley, 1977.

3 Schwartz J, Spix C, Touloumi G *et al*. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *J Epidemiol Commun Health* 1996; 50(Suppl 1): s3–s11.

4 Thurston GD, Kinney PL. Air pollution epidemiology: considerations in time series modeling. *Inhal Toxicol* 1995; 7: 71–83.

5 Henry TG. *Graphing data: techniques for display and analysis*. California: SAGE Publications, 1995.

6 Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Statist Assoc* 1979; 74: 829–36.

7 Bowman AW, Azzalini A. *Applied smoothing techniques for data analysis*. New York: Oxford Press, 1997.

8 Hastie T, Tibshirani RJ. *Generalized additive models*. London: Chapman and Hall, 1990.

9 Simonoff JS. *Smoothing methods in statistics*. New York: Springer-Verlag, 1996.

10 Erbas B, Hyndman RJ. The effect of air pollution and climate on hospital admissions for chronic obstructive airways disease: a nonparametric alternative. In press.

11 Cleveland WS, Terpenning IJ. Graphical methods for seasonal adjustment. *J Am Statist Assoc: Theory Methods* 1982; 77: 52–62.

12 Makridakis S, Wheelwright SC, Hyndman RJ. *Forecasting: methods and applications*. 3rd edition. New York: Wiley & Sons, 1998.

13 Kendall M, Ord KJ. *Time series*. 3rd edition. Kent: Hodder and Stoughton Ltd, 1990.

14 Cleveland RB, Cleveland WS, McRae JE, Terpenning I. *STL: a seasonal-trend decomposition procedure based on loess* [with discussion]. *J Oficial Stat* 1990; 6: 3–73.

15 Cleveland WS. *The elements of graphing data*. Revised edn. New Jersey: Hobart Press, 1994.

16 Chambers JM, Cleveland WS, Kleiner B, Tukey PA. *Graphical methods for data analysis*. New York: Chapman and Hall, 1983.

17 Saez M, Sunyer J, Castellsague J *et al*. Relationship between weather temperature and mortality: a time series analysis approach in Barcelona. *Int J Epidemiol* 1995; 24: 576–581.

18 Ballester F, Corella D, Perez-Hoyos S *et al*. Mortality as a function of temperature. A study in Valencia, Spain, 1991–1993. *Int J Epidemiol* 1997; 26: 551–61.

19 Morgan G, Corbett S, Wlodarcyzk J. Air pollution and hospital admissions in Sydney, Australia, 1990–1994. *Am J Pub Health* 1998; 88: 1761–6.

20 Pitard A, Viel J. Some methods to address collinearity among pollutants in epidemiological time series. *Stat Med* 1997; 16: 527–44.

21 Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied regression analysis and other multivariable methods*. 3rd edition. Pacific Grove: Duxbury Press, 1998.

22 Cleveland WS. *Visualizing data*. New Jersey: Hobart Press, 1993.

23 Cook RD, Weisberg *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

24 Akaike H. A new look at statistical model identification. *IEEE Transac Auto Control*. 1974; AC-19: 716–23.

25 Venables WN, Ripley BD. *Modern applied statistics with S-PLUS*, 2nd edition. New York: Springer-Verlag, 1997.

26 Chambers J, Hastie T. *Statistical models in S*. California: Wadsworth and Brooks/Cole Advanced Books and Software, 1992.

27 Dunn P, Smyth G. Randomized quantile residuals. *J Comput Graphical Stat* 1996; 5: 236–44.

28 Diggle P. *Time series: a biostatistical introduction*. Oxford University Press, 1990.

29 Campbell MJ, Tobias A. Causality and temporality in the study of short-term effects of air pollution on health. *Int J Epidemiol* 2000; 29: 271–3.

Appendix

Nonparametric smoothing

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote pairs of observations. Then a nonparametric smooth curve is a function $s(x)$ with the same domain as the values in $\{X_t\}$:

$$s(x) = \mathrm{E}(Y_t | X_t = x)$$

The function $s(x)$ is intended to show the mean value of Y_t for given values of $X_t = x$.

A kernel smoother is a weighted average of the form:

$$\hat{s}(x) = \sum_{j=1}^{n} w_j(x) Y_j$$

where the weights are given by:

$$w_j(x) = \frac{K\left(\frac{x-j-x}{b}\right)}{\sum_{i=1}^{n} K\left(\frac{x-i-x}{b}\right)}$$

The function K is a 'kernel function', which is usually non-negative, symmetric about zero and integrates to one. A common choice for K is the standard normal density function. The value of b (the 'bandwidth') determines how smooth the resulting curve will be. Large bandwidth values will produce smoother curves and small bandwidth values will produce more wiggly curves.

A cubic smoothing spline is defined as a function $\hat{s}(x) = g(x)$ that minimises the penalised residual sum of squares:

$$S_{\lambda}(g) = \sum_{t=1}^{n} [Y_t - g(X_t)]^2 + \lambda \int (g''(x))^2 \, dx$$

Thus, S is a compromise between goodness-of-fit and the degree of smoothness. The smoothness of the fitted curve is controlled by λ (the 'smoothing parameter'). Large values of λ will produce smoother curves, and smaller values will produce more wiggly curves.

In local linear regression, instead of fitting a straight line to the entire data set, we fit a series of straight lines to sections of the data. The resulting curve is the smoother $\hat{s}(x)$. The estimated curve at the point x is $y = a_x$ where a_x and b_x are chosen to minimise the weighted sum of squares:

$$\sum_{j=1}^{n} w_j(x)[Y_j - b{-}x - a{-}x(X_j - x)]^2$$

Here, $w_j(x)$ represents the weights. The weight function $w_j(x)$ can be the same as in kernel smoothing.

Loess smoothing is a form of local linear regression with some additional steps to make it more robust to outliers. The 'span' of a loess smoother describes the percentage of points receiving non-zero weight in the calculation of the smoother at each value of x.