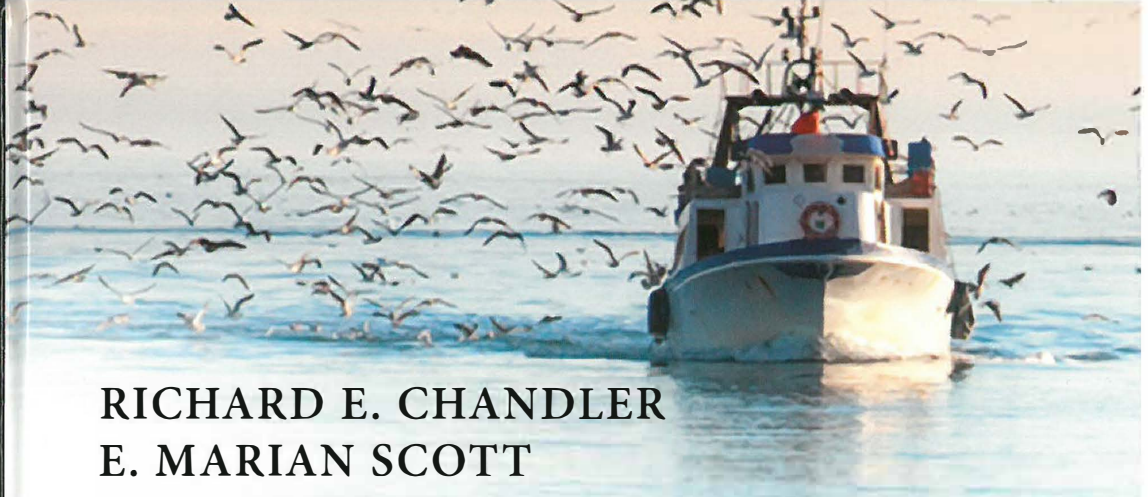

RICHARD E. CHANDLER
E. MARIAN SCOTT

Statistical Methods for Trend Detection and Analysis

in the Environmental Sciences

WILEY

STATISTICS IN PRACTICE

This edition first published 2011

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Chandler, R. E. (Richard E.)
Statistical methods for trend detection and analysis in the environmental sciences / Richard Chandler, E. Marian Scott.
p. cm. – (Statistics in practice ; 90)
Includes bibliographical references and index.
ISBN 978-0-470-01543-8 (hardback)
1. Environmental sciences – Statistical methods. I. Scott, E. Marian. II. Title.
GE45.S73C43 2011
577.01′1 – dc22

2010051076

A catalogue record for this book is available from the British Library.

Print ISBN: 978-0-470-01543-8
ePDF ISBN: 978-1-119-99156-4
oBook ISBN: 978-1-119-99157-1
ePub ISBN: 978-1-119-99196-0

Typeset in 9/11 Times by Laserwords Private Limited, Chennai, India
Printed in Great Britian by TJ International Ltd, Padstow, Cornwall

prior distributions for which such calculations are possible. Thus, Bayesian estimation and inference is often carried out using *Markov chain Monte Carlo* (MCMC) computer algorithms that are designed to provide a simulated sample from the posterior distribution, from which any quantities of interest can be estimated (see Gelman *et al.*, 1995, Part III for an overview). MCMC methods are also extremely useful for the fitting of complex models in situations where direct calculation of the likelihood is infeasible. In R, libraries `R2WinBUGS` (Sturtz, Ligges and Gelman, 2005) and `BRugs` (Thomas *et al.*, 2006) provide interfaces to packages such as WinBUGS (Lunn *et al.*, 2000) and OpenBUGS (Thomas *et al.*, 2006), which can be used to carry out MCMC computations for a wide range of Bayesian analyses. BayesX (Belitz *et al.*, 2009) is another freely available package for Bayesian computation; this is used in Chapter 10 of the present volume.

In many environmental situations, a Bayesian approach provides an appealing means of incorporating genuine subject-matter knowledge into an analysis via the prior distribution. For example, Leith and Chandler (2010) carry out a Bayesian analysis of simulated climate data for the end of the twenty-first century, in which physical limits on the climate system are incorporated by using the ranges of historical climate observations to set prior distributions for regression coefficients representing future means and linear trends. If, on the other hand, little subject-matter knowledge is available then this can be represented by using prior distributions with very large variances, although such 'noninformative' or 'diffuse' priors can have unexpected implications in some circumstances (Davison, 2003, Section 11.1.3).

The need to specify a prior distribution is sometimes seen as a disadvantage of Bayesian methods because it introduces an element of subjectivity: two analysts, fitting the same model to the same data but with different priors, could reach different conclusions. However, it can be shown that in large samples irrespective (within reason) of the choice of prior, the results from Bayesian and likelihood based analyses are almost the same: for example, the Bayes estimator and credible intervals are very similar to the MLE and corresponding confidence intervals (Davison, 2003, Section 11.2). Therefore, given enough data, our two analysts should be able to resolve their differences. If they cannot do this and their priors are both justifiable on subject-matter grounds, the implication is that the available data do not contain enough information to discriminate between alternative plausible scenarios.

3.2 Multiple regression techniques

An obvious extension to the linear trend model (3.2) is to supplement, or replace, the time index t with the values of other quantities that may be responsible for changes in the variable of interest. Such quantities are referred to as *covariates*, and the variable of interest is often called the *response variable*. The simplest way to model the effect of several covariates is via a multiple regression model of the form

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i x_{it} + \varepsilon_t = \mu_t + \varepsilon_t \text{, say} \qquad (t = 1, \ldots, T). \tag{3.34}$$

Here, x_{it} denotes the value of the ith covariate at time t.

The multiple regression model can once again be written in the matrix form (3.10), the only differences being that $\boldsymbol{\beta} = (\beta_0 \quad \beta_1 \quad \cdots \quad \beta_p)'$ now has length $k = p + 1$ and that $\boldsymbol{X}$ has dimensions $T \times k$. As before, the intercept β_0 is accommodated by filling the

first column of $\boldsymbol{X}$ with ones. A consequence of the matrix representation is that all of the results from the previous section are applicable here as well. Techniques for identifying influential observations can be used without modification; the diagnostics described in Section 3.1.1 can also be used to check the model. Additionally, it can be useful to plot the residuals against each of the covariates individually to check for unmodelled structure. Some effort can be saved by plotting the residuals against the fitted values, rather than against each covariate individually. The `plot.lm` function in R produces a variety of diagnostics for multiple regression models; straightforward summaries of the main ideas can be found in Davison (2003, Section 8.6) and Faraway (2005, Chapter 4). For more extensive discussion and details of more sophisticated diagnostics, see Cook and Weisberg (1999, Chapter 14) and Fox (2002, Chapter 6).

When the underlying assumptions are satisfied, multiple regression provides the ability to represent all of the processes affecting the quantity of interest within a single model. This contrasts with the common approach of standardising time series data prior to analysis so as to remove structure that is not of direct interest. A disadvantage of the latter approach is that any form of adjustment, such as the removal of seasonality, is a form of preprocessing and therefore needs to be accounted for subsequently. By way of illustration, consider a hypothetical example involving the association between air pollution and human mortality. Mortality time series typically show seasonal fluctuations, some but not all of which may be attributable to seasonal variation in pollution (Schwartz, 1994). If seasonality is removed from a mortality time series prior to analysis, for example by subtracting monthly means, it is likely that some of the pollution effect will be removed inadvertently at the same time. Moreover, if the resulting *anomalies* are regressed on raw or deseasonalised pollution levels, the standard errors of regression coefficients will tend to be underestimated because the analysis does not allow for the possibility that pollution is responsible for some of the discarded seasonal structure. The problem can be avoided entirely by fitting a multiple regression model to the raw data, containing covariates that represent seasonality explicitly as described below.

In principle, multiple regression models can also be used for extrapolation, using the methodology described in Section 3.1.4. However, in practice this is only possible if future values of the covariates are available. One way to achieve this is by using lagged values of the covariates in the model, if such lagged values have any explanatory power at the time horizon of interest. An alternative is to base extrapolations on 'scenarios', whereby the effect of a prespecified sequence of covariates is investigated. Scenario based extrapolation is useful in situations where the future values of the covariates are, at least nominally, under the control of policymakers – examples of such covariates might include levels of industrial sulfur emissions and fisheries quotas.

3.2.1 Representing seasonality in regression models

Seasonality is often one of the most important factors controlling environmental processes at sub-annual timescales. In some cases, it arises mainly due to dependence on one or more seasonally varying covariates, and can be accounted for by including these covariates in a multiple regression model. However, if there are no plausible covariates to which seasonality can be attributed, or if data on such covariates are not available, a different approach is required.

A crude way to handle seasonality is to fit separate models for different times of year. However, the multiple regression framework offers the possibility of representing seasonal controls explicitly via the use of 'dummy' covariates. The simplest option is perhaps

to define binary *indicator variables* for each time period. For example, for quarterly data one could define variables x_{1t}, x_{2t}, x_{3t} and x_{4t}, such that x_{jt} takes the value 1 for observations in quarter j and zero otherwise. A multiple regression model involving just these covariates takes the form

$$Y_t = \beta_0 + \sum_{j=1}^{4} \beta_j x_{jt} + \varepsilon_t. \tag{3.35}$$

In this model, the fitted value for quarter j will be $\beta_0 + \beta_j$, since $x_{jt} = 1$ during this quarter and the other covariates are all zero. A least squares fit will, in principle, equate $\beta_0 + \beta_j$ with the mean of the observations for quarter j ($j = 1, \ldots, 4$). However, this reveals a problem: since there are only four quarterly means, it is not possible to estimate the five coefficients β_0 to β_4. The model is said to be *overparameterised*. The difficulty is usually resolved by imposing a constraint on the coefficients, for example by setting one of them to zero. The precise choice of constraint does not affect the fitted values from the model, but it does affect the interpretation of the coefficients. Consider, for example, setting $\beta_0 = 0$ in (3.35). In this case, the fitted value for quarter j is just β_j, which can therefore be interpreted as the mean level for that quarter. If instead we set $\beta_1 = 0$, then the fitted value for quarter 1 is β_0 and the fitted value for any other quarter j is $\beta_0 + \beta_j$. In this case therefore, β_0 is the mean for quarter 1 and, for quarter $j > 1$, β_j is the difference between the means for quarters 1 and j.

The use of indicator variables can also be regarded as a means of adjusting for seasonality. The residuals from model (3.35) are precisely the anomalies that would be obtained by subtracting the quarterly means prior to analysis. If the purpose of the analysis is to assess the effect of some other covariate on the response, this can be quantified by fitting an extended model incorporating the extra covariate in addition to the seasonal indicators. The fitted values from such a model will be the same as those from a separate analysis of the anomalies, but the regression coefficient corresponding to the covariate of interest, and its standard error, may be rather different since they take into account all of the available information.

In the discussion above, the dummy covariates x_{1t} to x_{4t} effectively code for a single variable 'quarter', which defines four separate groups or categories. Such grouping variables are called *factors*; the separate groups are referred to as *levels*. Regression software will usually handle factors automatically, providing they are defined as such (correct behaviour can be guaranteed in R by using characters, rather than numbers, to represent the different groups). The issue of overparameterisation is, however, always present and, to interpret software output, it is necessary to know what constraints have been imposed. In R, the default behaviour for unordered factors is to use 'corner-point' constraints, in which the coefficient associated with the first level (β_1 in the discussion above) is set to zero. Another option is to constrain all of the coefficients associated with a factor to sum to zero. In model (3.35), if there were equal numbers of observations in each quarter then, under a sum-to-zero constraint, the estimate of β_0 would be the overall mean of the series and β_j would be the average deviation from this overall mean in quarter j. The interpretation is less straightforward with differing numbers of observations per quarter. For further details of factor coding in general, see Dobson (2001, Section 2.4). Fox (2002, Chapter 4) and Venables and Ripley (1999, Section 6.2) give a comprehensive account of the facilities available in R.

The factor based approach to modelling seasonality is similar in spirit to the practice of fitting separate models to different subgroups of observations. In both cases the need to