

## MODELLING TRENDS IN GROUNDWATER LEVELS BY SEGMENTED REGRESSION WITH CONSTRAINTS

QUANXI SHAO<sup>1\*</sup> AND N.A. CAMPBELL<sup>1</sup>

*CSIRO Mathematical and Information Sciences*

### Summary

This paper provides a statistically unified method for modelling trends in groundwater levels for a national project that aims to predict areas at risk from salinity in 2020. It was necessary to characterize the trends in groundwater levels in thousands of boreholes that have been monitored by Agriculture Western Australia throughout the south-west of Western Australia over the last 10 years. The approach investigated in the present paper uses segmented regression with constraints when the number of change points is unknown. For each segment defined by change points, the trend can be described by a linear trend possibly superimposed on a periodic response. Four different types of change point are defined by constraints on the model parameters to cope with different patterns of change in groundwater levels. For a set of candidate change points provided by the user, a modified Akaike information criterion is used for model selection. Model parameters can be estimated by multiple linear regression. Some typical examples are presented to demonstrate the performance of the approach.

*Key words:* Akaike information criterion; change point; groundwater levels; model selection; segmented regression with constraints.

### 1. Introduction

Dryland salinity is a major problem for agriculture in Australia. More than 1.2 million hectares of productive land are currently affected by salinity, while some estimates put a further 1.6 million hectares at risk. The loss of agricultural production is more than \$A240 million per year. A current project known locally as Salt Scenario 2020 (SS2020), which involves Agriculture Western Australia (AgWest) and the Remote Sensing and Monitoring Project in CSIRO Mathematical and Information Sciences, aims to incorporate information on trends in groundwater levels, based on borehole data, into salinity risk prediction. A groundwater level is a measurement of the vertical distance from land surface to watertable. Data have been collected irregularly in thousands of boreholes throughout Western Australia. The SS2020 has two components — describing the trends within a borehole over time, and relating the trends to characteristics of the landscape. This paper gives details of an approach developed for modelling the trends within a borehole over time.

---

Received October 1999; revised January 2001; accepted September 2001.

\* Author to whom correspondence should be addressed.

<sup>1</sup> CSIRO Mathematical and Information Sciences, Leeuwin Centre, 65 Brockway Road, Floreat Park, WA 6014, Australia. e-mail: Quanxi.Shao@cmis.csiro.au

*Acknowledgments.* This work was funded by the National Land and Water Resources Audit Project. The authors thank staff of Agriculture Western Australia, especially Ruhi Ferdowsian, Arjen Ryder, Richard George and Don Bennett, for providing the borehole data and for many useful discussions over the nature of the data. They also thank Dr Eddy Campbell for helpful discussion at an early stage of the project, and the editor and two anonymous referees for their careful reading and valuable comments.

The method used currently by AgWest and in other states in Australia is a simple linear regression (AgBores version 2.1a, Heinrich & Bennett, 1997–2000), which often provides questionable results as the groundwater trends tend to follow different patterns over different periods. As an alternative, the simple linear regression model is sometimes applied to annual maxima of groundwater levels only. (An annual maximum is the highest watertable observed in each calendar year.) Four issues arise from this approach. First, annual maxima use only part of the collected data and do not capture the entire change in trend. Second, the maxima derived from data may not be the true values due to low and irregular sampling frequencies. Third, the magnitude of the seasonal variation is completely ignored and the results cannot represent the trends in groundwater levels. Finally, only small sample sizes for annual maxima are available because of the short sampling periods, and therefore it is often difficult to obtain reliable statistical results.

AgWest required a consistent approach to be applied to all borehole data; such an approach is developed and evaluated in this paper, based on a segmented regression with constraints. Details of the problem description and definitions of change points are given in Section 2.

A major part of the modelling procedure is the determination of the number of change points, which is an active area of statistical research; see Yin (1988), Yao & Au (1989), Barry & Hartigan (1992) and Lee (1997). Many results in the existing literature deal with a known number of change points, and seek to determine their locations. However, for the borehole data, the number of change points is unknown and ideally needs to be determined during the modelling procedure. In this paper, we establish a model that identifies the number and type of change points. At the same time, the linear trends and seasonal variations are estimated. A modified Akaike information criterion ( $AIC_C$ ) is applied to the case where the number of change points is unknown. Section 3 discusses parameter estimation and model selection. Section 4 demonstrates the approach by analysing some typical boreholes. Section 5 gives conclusions and discussion.

## 2. The problem

Thousands of boreholes have been drilled and monitored throughout Western Australia. For these boreholes, the sampling intervals are unequal and can vary from one month to three months, and sometimes as much as six months. The total monitoring period for a borehole can vary from less than a year to over 10 years. It is difficult to handle data collected in short time periods and/or in low frequencies (e.g. six months). For the model used in this paper, reasonable sampling frequencies are required (e.g. at least one sample for each season) for modelling seasonal variations. At least four samples are required between candidate change points to ensure that the fitting procedure works. However, we do not intend to set a clear exclusion criterion. Instead, we allow the user to decide whether or not a borehole should be analysed.

Different boreholes in the available database show varying trends in groundwater level. Generally speaking, a deep watertable shows a linear trend, whereas a shallow watertable can be affected dramatically by seasonal variations. Typical examples of these trends are shown in Figure 1. There are many boreholes where the trends change abruptly; these changes may be attributable to external factors such as weather and land management practices.

Examination of the database identifies three distinct types of change point. A 'break point' is defined here as a time at which both the linear trend and periodic response change and a discontinuity may occur. A 'join point' is a time at which the fitted linear trend changes

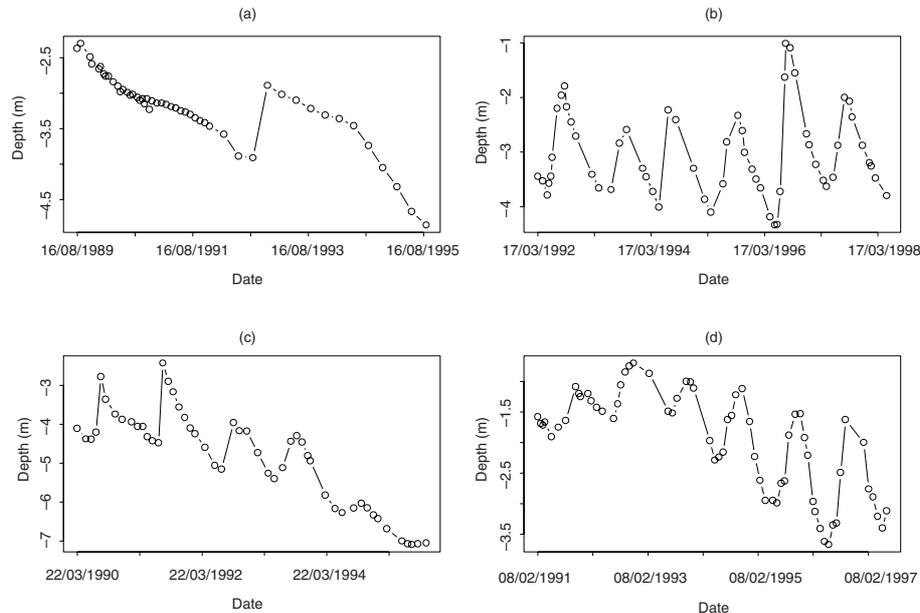


Figure 1. Some typical groundwater data: (a) Bore English #8; (b) Bore DJ07I; (c) Bore AC3D90; (d) Bore WW07D. Depth represents the groundwater level, which is the vertical distance from land surface to watertable.

but is continuous, while the periodic response is the same for the segments at both sides. A ‘knot point’ is a time at which the periodic response changes but is continuous, while the linear trend is the same for the segments at both sides. A knot and a join point can occur together and form another type of change point. That is, the linear trend and periodic response can change at the same time, but in a continuous manner. These four types of change points are sometimes observed for a single borehole during the sampling period. Technically, break points should not be seen in any borehole because of the continuous nature of the change in the groundwater levels. However, a marked change in groundwater levels due to an external event such as heavy rainfall may be recorded as an abrupt change because of the low sampling frequency. The use of break points accommodates these observed abrupt changes in the modelling procedure. The model given in the next section identifies the number and type of change points.

### 3. Statistical approaches to modelling borehole data

This section develops, in stages, a statistical model that is capable of modelling the trends identified. A method is also required for selecting an optimal model. An extended version of the Akaike information criterion (AIC) is proposed.

#### 3.1. A simple segmented regression

Assume that  $x_i$  denotes the measured groundwater level at time  $t_i$  ( $i = 1, 2, \dots, N$ ). A simple segmented regression model is written as

$$x(t) = \mu_\ell(t) + \varepsilon(t), \quad r_{\ell-1} < t \leq r_\ell \quad (\ell = 1, 2, \dots, L + 1). \quad (1)$$

The deterministic part of the model,  $\mu_\ell(t)$  ( $\ell = 1, \dots, L + 1$ ), is given by

$$\mu_\ell(t) = a_\ell + b_\ell t + \rho_\ell \sin(2\pi\omega_0 t + \theta_\ell) \quad (\ell = 1, \dots, L + 1); \quad (2)$$

the errors  $\{\varepsilon(t_i); i = 1, 2, \dots, N\}$  are assumed to be independent and identically distributed as  $N(0, \sigma^2)$ . Here  $r_1 < r_2 < \dots < r_L$  are real numbers representing the break points, and we take  $r_0 = 0$  and  $r_{L+1} = \infty$ . Denote this model by SRST[ $L, r_1, \dots, r_L$ ] (segmented regression system over time). The model is characterized by  $r_1 < r_2 < \dots < r_L$ , which divide the time dimension into segments, each of which exhibits a different pattern. The main requirement in fitting SRST[ $L, r_1, \dots, r_L$ ] is the determination of the number of change points,  $L$ , and their locations  $r_1, r_2, \dots, r_L$ . By convention,  $L = 0$  corresponds to a simple linear regression. For the  $\ell$ th segment, the linear trend is  $a_\ell + b_\ell t$  and the periodic response is  $\rho_\ell \sin(2\pi\omega_0 t + \theta_\ell)$ , where  $\rho_\ell$  is the amplitude,  $\omega_0$  is the period and  $\theta_\ell$  is the phase. The linear trend is superimposed on the periodic response.

In our applications, the period  $\omega_0$  is assumed known and equal to one year ( $\omega_0 = 1$ ), representing the annual cycle. Rewrite

$$\rho_\ell \sin(2\pi t + \theta_\ell) = \rho_\ell \cos(\theta_\ell) \sin(2\pi t) + \rho_\ell \sin(\theta_\ell) \cos(2\pi t),$$

and let  $c_\ell = \rho_\ell \cos(\theta_\ell)$  and  $d_\ell = \rho_\ell \sin(\theta_\ell)$ . Each segment of the model is simply a multiple linear regression with parameters  $\{a_\ell, b_\ell, c_\ell, d_\ell\}$  ( $\ell = 1, 2, \dots, L + 1$ ). The amplitude and the phase are estimated by  $\hat{\rho}_\ell = \sqrt{\hat{c}_\ell^2 + \hat{d}_\ell^2}$  and  $\hat{\theta}_\ell = \arctan(\hat{d}_\ell/\hat{c}_\ell)$ , respectively.

### 3.2. Model selection

Ideally, all possible combinations of sampling points would be examined and the optimal model would then be determined using a suitable criterion. However, this is not likely to be practical because of the potentially very large number of combinations involved. Therefore, before selecting the optimal model, it is necessary to pre-determine a set of candidate break points  $r_1 < r_2 < \dots < r_{L_0}$ . From a modelling perspective, the candidate break points do not have to be the sampling times. Steps in the process of model selection are given below.

For a given  $L \leq L_0$ , there are  $\binom{L_0}{L}$  different sets having  $L$  break points. A major issue in the statistical modelling is the selection of an optimal model for change point detection. There are many results published. Some references related to multiple change points include Yin (1988), Seber & Wild (1989 Chapter 9), Yao & Au (1989), Barry & Hartigan (1992) and Lee (1997). In regression analysis, a model fit can be measured by its residual sum of squares (RSS), which decreases as the number of parameters increases. A criterion is needed so that not too many parameters are used in selecting competing models. The so-called Akaike information criterion (AIC) (see Akaike, 1974) is widely used in regression models. Tong (1980) applied AIC for model selection in a threshold autoregression model. By extending the method proposed by Sugiura (1978) for linear regression, Hurvich & Tsai (1989) derived a modified Akaike information criterion ( $AIC_C$ ) in small samples, and in a later paper (Hurvich & Tsai, 1991) they claimed that the  $AIC_C$  dramatically reduced the bias and improved model selections. The  $AIC_C$  penalizes the RSS by a function of the number of free parameters and is given by

$$AIC_C(p) = \ln \frac{RSS}{N} + \frac{N + p}{N - p - 2}, \quad (3)$$

where  $N$  is the number of observations and  $p$  is the number of free parameters in the model.

For the model specified by (2),  $p = 4(L + 1)$ . We retain  $p$  in the notation of  $AIC_C(p)$  so that it can be applied to extended models.

For a given set of break points, the observations can be grouped by the break points in the set. Each segment of the model can be fitted separately and the  $AIC_C$  can be calculated accordingly. The optimal set of break points for a given  $L$  is the one which minimizes  $AIC_C$ . The optimal model is then the one with the overall minimum  $AIC_C$ .

### 3.3. Segmented regression with constraints

Break points are unconstrained change points in the sense that model parameters can change freely at a break point, whereas join points and knot points are constrained change points. Some common parameters are used by consecutive time periods separated by a join point and/or a knot point, while there are no common parameters between consecutive time periods separated by a break point. Therefore, in the modelling procedure, regressions can be implemented separately for segments defined by break points, but need to be fitted simultaneously for consecutive segments defined by join and/or knot points.

We now discuss parameter estimation for each segment defined by break points. For ease of notation, it is assumed that there is no break point in the model (because the model can be fitted separately for segments defined by break points). To ensure the continuity of the regression curve at the join points,  $\theta_i$  ( $i = 0, \dots, k$ ) are assumed to be the same, i.e.  $\theta_0 = \dots = \theta_k = \theta$ . Ideally, the common phase  $\theta$  should be determined by model fitting. However, in this application the seasonal variations are strongly associated with weather and temperatures, which have clear seasonal patterns in Western Australia. Therefore, we can assume that  $\theta$  is known. Without loss of generality, we can assume that  $\theta = 0$  by selecting the start time point. The determination of the starting time point is based on expert judgement rather than statistical optimum. In this application, February is a preferred starting point and works well. For statistical completeness, an iteration algorithm on parameter estimation with unknown  $\theta$  is described in the Appendix, although it was not used in the salinity project.

Assume that there are  $j$  join points  $J_i$ ,  $i = 1, \dots, j$ , and  $k$  knot points  $K_i$ ,  $i = 1, \dots, k$ . (The  $J_i$  and  $K_i$  do not have to correspond to observation times.) Assume further that  $K_i \subset \{1, 2, \dots\}$ , which means that the amplitudes can only change after the completion of a full cycle and ensures the continuity of the regression curve at knot points. The candidate join points should be pre-determined, to avoid having a very large number of combinations. The full model for a segment defined by break points can then be written as

$$\mu_t = a + \sum_{i=0}^j b_i(t - J_i)_+ + \sum_{i=0}^k \rho_i \delta(t - K_i) \sin(2\pi t + \theta), \quad (4)$$

where  $u_+ = u$  if  $u > 0$  and 0 if  $u \leq 0$ , and  $\delta(u) = 1$  if  $u > 0$  and 0 if  $u \leq 0$ . By convention,  $J_0 < \min(\text{sampling dates})$  and  $K_0 < \min(\text{sampling dates})$ .

A set-up for parameter estimation for each segment defined by break points is outlined in the Appendix. Assume that there are  $j_i$  join points and  $k_i$  knot points for segment  $i$  defined by break points. The number of free parameters in segment  $i$  is  $p_i = 3 + j_i + k_i$ . Therefore the total number of independent parameters is  $p = \sum_i (3 + j_i + k_i)$ . The optimal model can then be obtained by minimizing the overall  $AIC_C$ .

In this paper,  $AIC_C$  is employed for model selection. The number of candidate change points is usually small in this application. Exhaustive combinations are examined and the one with smallest  $AIC_C$  is selected as the final model. However, in general, exhaustive exami-

nation can be very time-consuming if  $L_0$  is large. In that case, a stepwise model selection can be employed. A step-forward-and-then-backward procedure was suggested by An (pers. comm.); some relevant results can be found in An & Gu (1984), Wang & An (1984) and An & Gu (1986 Chapters 2 and 6).

### 3.4. Statistical hypothesis testing

Although hypothesis testing is not used in this application, it is useful to briefly discuss hypothesis testing, as distinct from model selection. Discussions on the differences between hypothesis testing and model selection can be found in Geisser (1993 Chapter 1). Gallant & Fuller (1973) derived hypothesis testing for segmented polynomial regression models. The techniques of hypothesis testing can be used in model selection for our model. Note that deleting a change point from a model (denoted by  $M_0$ ) results in a degenerate form (denoted by  $M_1$ ). That is,  $M_1$  is a special case of  $M_0$ . The model selection between  $M_0$  and  $M_1$  can be written as a test of

$$H_0: M_1 \text{ is true} \quad \text{against} \quad H_1: M_0 \text{ is true.}$$

Following the arguments of Gallant & Fuller (1973), we can form a test statistic as

$$t(x) = \frac{\frac{1}{2} \sum_{t=1}^N \left( (x_i - \hat{x}_0(t_i))^2 - (x_i - \hat{x}_1(t_i))^2 \right)}{\frac{1}{N-p} \sum_{t=1}^N (x_i - \hat{x}_1(t_i))^2},$$

where  $\hat{x}_0(t_i)$  and  $\hat{x}_1(t_i)$  are the estimated groundwater levels at sampling time  $t_i$  under models  $M_0$  and  $M_1$ , respectively, and  $p$  is the number of parameters (rather than the number of free parameters as used in  $AIC_C$ ) in model  $M_0$ . The approximate test rejects  $H_0$  if  $t(x) > c_{1-\alpha}(F_{2,N-p})$  where  $\alpha$  is the significance level.

The above significance testing relies on the choice of significance level. For the salinity project a tidy method is preferable so that there can be a unique result for each borehole. Therefore, we do not adopt this method here. Further study is needed to compare the different model selection techniques.

## 4. Examples

Our approach is being applied to all the boreholes in the database of AgWest. In this section, we demonstrate the approach by applying it to a range of typical boreholes. The measurement units of time and depth are years and metres, respectively. The starting point is taken as 1 February 1960 because it appears that the periodic variation starts in February for the examples given here, and so that all sampling times have positive values.

There are 54 observations for Bore English #8 plotted in Figure 1(a). The candidate join points are (15/11/1990, 01/06/1994). The candidate break point is 01/09/1992. Seasonal variation is not fitted. The final model is given by

$$\mu(t) = \begin{cases} 14.0602 - 0.5580t + 0.1935(t - t_1)_+ & \text{if } t \leq 32.61, \\ 9.2559 - 0.3705t - 0.7708(t - t_2)_+ & \text{if } t > 32.61, \end{cases}$$

where  $(t_1, t_2) = (30.81, 34.35)$  corresponds to (15/11/1990, 01/06/1994). The break point value 32.61 corresponds to 01/09/1992. The criterion value  $AIC_C = -4.3156$ . The model fit and comparison with a simple linear regression are shown in Figure 2. The estimated

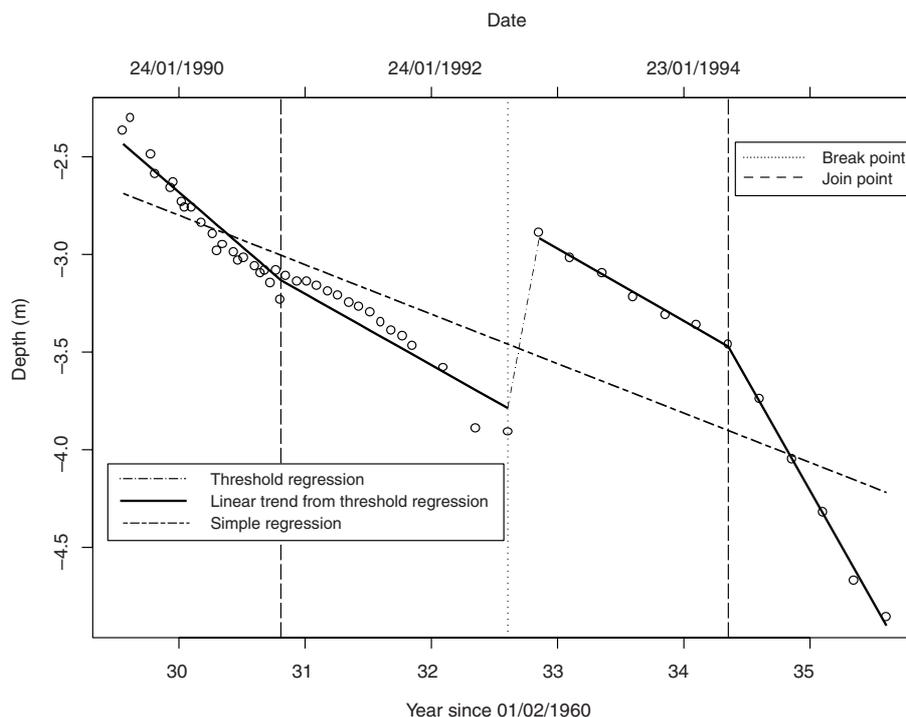


Figure 2. Observed and fitted depths for Bore English #8. Depth represents the groundwater level. Circles represent observed values.

rates of decrease are 0.5580 m/year, 0.3645 m/year, 0.3705 m/year and 1.1413 m/year, respectively. The simple linear regression is  $\mu_t = 4.7972 - 0.2532t$  with  $AIC_C = -1.3814$ . It is quite possible to obtain a smaller  $AIC_C$  by selecting more join points before the break point 01/09/1992. However, we chose not to do so.

There are 57 observations in Bore DJ07I plotted in Figure 1(b). (01/02/1994, 01/02/1995) are selected as both the candidate join points and the candidate knot points. The candidate break point is 27/06/1996. The final model is given by

$$\mu(t) = \begin{cases} 2.9822 - 0.1782t - 0.7489 \sin(2\pi t) & \text{if } t \leq 32.43, \\ 25.9824 - 0.7673t - 0.7482 \sin(2\pi t) & \text{if } t > 32.43, \end{cases}$$

with  $AIC_C = -0.8624$ , where the break point value 32.43 corresponds to 27/06/1996. The model fit and comparison with a simple linear regression are shown in Figure 3. In the segmented model, the estimated rates of decrease are 0.1782 m/year and 0.7673 m/year, respectively. The simple linear regression is  $\mu(t) = -3.4557 + 0.0121t$ , with  $AIC_C = 0.6488$ , giving an estimated rate of increase of 0.0121 m/year.

There are 58 observations for Bore AC3D90. The observations collected after 19/02/1996 (inclusive) are omitted from the analysis because the bore dried up. Only the first 50 observations are used in the analysis. The plot is shown in Figure 1(c). There was a change in land use in 1991. After discussion with experts in hydrology, the data before 11/07/1991 are considered as a single segment. 01/02/1994 is selected as both the candidate join point and the candidate knot point. The candidate break points are (11/07/1991, 14/07/1992). The final

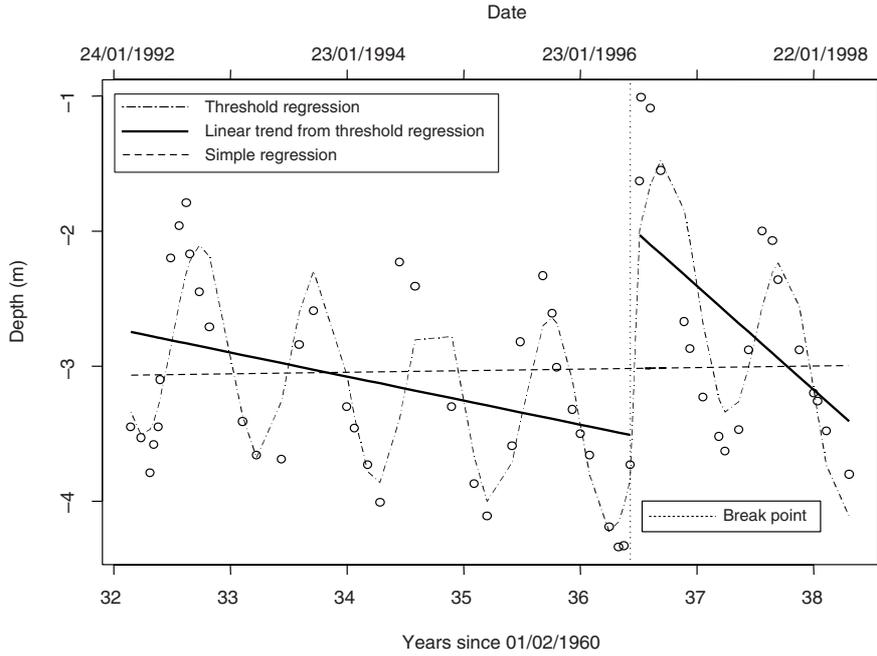


Figure 3. Observed and fitted depths for Bore DJ07I. Depth represents the groundwater level. Circles represent observed values.

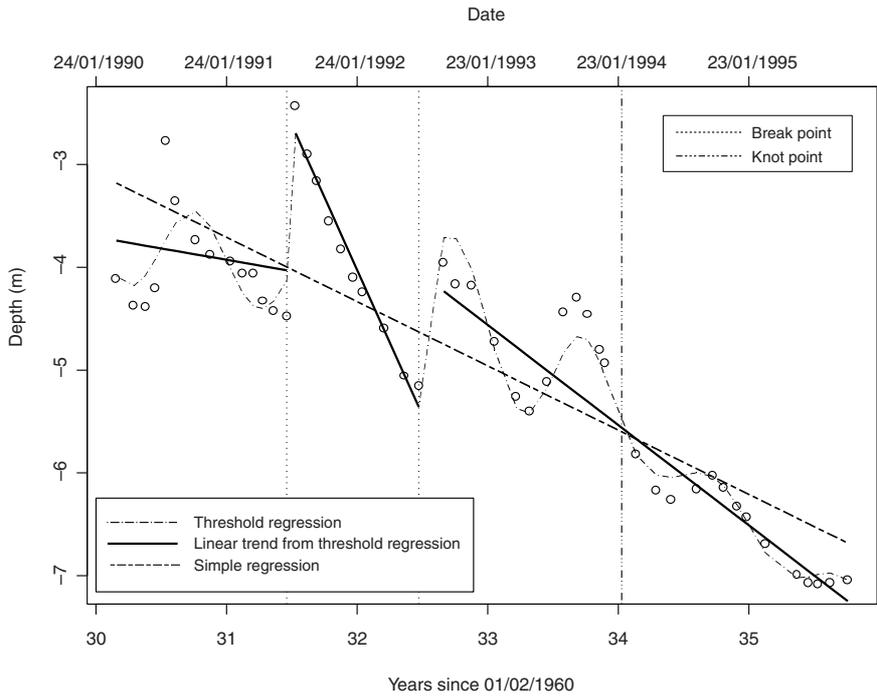


Figure 4. Observed and fitted depths for Bore AC3D90. Depth represents the groundwater level. Circles represent observed values.

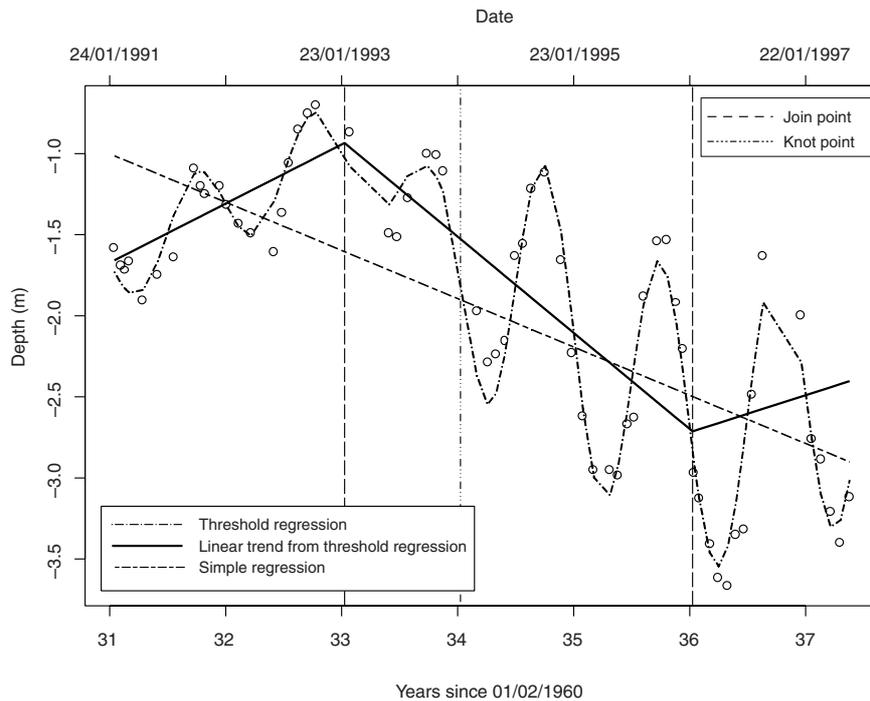


Figure 5. Observed and fitted depths for Bore WW07D. Depth represents the groundwater level. Circles represent observed values.

model adopted is

$$\mu(t) = \begin{cases} 2.9118 - 0.2206t - 0.4209 \sin(2\pi t) & \text{if } t \leq 31.46, \\ 86.3418 - 2.8241t & \text{if } 31.46 < t \leq 32.47, \\ 27.5989 - 0.9746t - 0.6047 \sin(2\pi t) \\ \quad + 0.3987 \sin(2\pi t)\delta(t - t_0) & \text{if } t > 32.47, \end{cases}$$

where the knot point  $t_0 = 34.02$  corresponds to 01/02/1994. The break point values (31.46, 32.47) correspond to (11/07/1991, 14/07/1992). The criterion value  $AIC_C$  is  $-1.9013$ . The model fit and comparison with a simple linear regression are shown in Figure 4. In the segmented model, the estimated rates of decrease are 0.2206 m/year, 2.8241 m/year and 0.9746 m/year. The rapid change in 1991–1992 may be caused by the disturbance due to the change in land use. The simple linear regression is  $\mu(t) = 15.6739 - 0.6252t$ , with  $AIC_C = 0.1750$ , giving an estimated rate of decrease of 0.6252 m/year.

There are 63 observations for Bore WW07D plotted in Figure 1(d). It can be seen that the magnitude of the seasonal variation is related to the groundwater level. (01/02/1992, 01/02/1993, 01/02/1994, 01/02/1995, 01/02/1996) are selected as both the candidate join points and the candidate knot points. The final model is given by

$$\mu_t = -12.9793 + 0.3647t - 0.9574(t - t_1)_+ + 0.8217(t - t_2)_+ \\ - 0.2828 \sin(2\pi t) - 0.6009 \sin(2\pi t)\delta(t - t_3),$$

where the join points  $(t_1, t_2) = (33.02, 36.02)$  correspond to (01/02/1993, 01/02/1996) and the knot point  $t_3 = 34.02$  corresponds to 01/02/1994. The criterion value  $AIC_C = -2.4531$ . The model fit and comparison with a simple linear regression are shown in Figure 5. In

the segmented model, the estimated rate of increase is 0.3647 m/year before 1993, the rate decreases at 0.5927 m/year ( $= 36.47 - 95.74$ ) in year 1994–1995, and then the rate increases again at 0.2290 m/year. The simple linear regression is  $\mu(t) = 8.2293 - 0.2978t$ , with  $AIC_C = 0.0193$ , giving an estimated rate of decrease of 0.2978 m/year.

## 5. Conclusions and discussion

Successfully modelling trends in groundwater level is not a simple matter. The examples presented above illustrate the power and flexibility of the segmented regression with constraints. The results to date are very promising, and our approach for modelling borehole data is being adopted by AgWest for routine use. A user-friendly S-PLUS<sup>®</sup> program can be obtained from the authors. The approach allows different linear trends for different segments of the data, and at the same time allows the amplitude(s) of the seasonal response in the data to vary.

Both the slopes and amplitudes (or fitted values at a nominated time) can be used to characterize the borehole data, and these values can then be related to appropriate characteristics of the landscape. The detection of change points is important in practice, because it encourages investigation into the causes; for example, variation in weather (rainfall and temperature) and land management. This information can then be integrated into the modelling of the borehole responses. Another potential use of our method is estimating the groundwater levels at the same time for all individual bores in a region so that the watertable at the region can be mapped using other statistical techniques such as spatial interpolation. For different spatial interpolators see e.g. Weber & Englund (1994).

The fundamental reason for using a statistical criterion such as AIC is to seek an objective measure to aid decision making, which very often has a subjective element such as experience and expert knowledge of the data concerned. Like many other statistical criteria, the modified AIC is ultimately based on large-sample properties. As a result, the statistically determined optimum needs to be balanced against expert judgement. An alternative model selection is based on hypothesis testing, as proposed by Gallant & Fuller (1973) for segmented polynomial regression models. It is very useful to compare the performance of different model selection methods.

Note that the borehole data were in fact irregular time series. However, our model is a segmented regression with time as covariate, and therefore possible autocorrelation errors are omitted. For example, in the model fitting for Bore DJ07I (see Figure 3), the systematically undershooting peaks and overlooking valleys reveal autocorrelation errors. Further, it is difficult in the segmented regression to assess uncertainties for model parameters, especially for the estimated change points. As alternatives, smoothing techniques, such as splines and non-parametric regression, can be used for fitting the bore data. However, the trends in groundwater levels cannot be easily evaluated by these techniques. Another disadvantage in our model is that the individual boreholes were considered separately. A possible extension of our approach would model all boreholes simultaneously. For example, random spline models, such as are discussed by Verbyla *et al.* (1999), may be employed to analyse the bore data and their relationship to hydrographical factors. More investigations are needed in modelling the trends in groundwater levels.

**Appendix: Parameter estimation for segmented regression with constraints**

The sampling dates are grouped as  $((t_{0,1}, \dots, t_{0,s_0}), \dots, (t_{j,1}, \dots, t_{j,s_j}))$  by potential join points. Define

$$X_{N \times 1}^{(J)} = (x_{0,1}, \dots, x_{0,s_0}, x_{1,1}, \dots, x_{1,s_1}, \dots, x_{j,1}, \dots, x_{j,s_j}),$$

$$T_{N \times (j+2)}^{(J)} = \begin{bmatrix} 1 & t_{0,1} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{0,s_0} & 0 & 0 & \cdots & 0 \\ 1 & t_{1,1} & t_{1,1} - J_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{1,s_1} & t_{1,s_1} - J_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{j,1} & t_{j,1} - J_1 & t_{j,1} - J_2 & \cdots & t_{j,1} - J_j \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{j,s_j} & t_{j,s_j} - J_1 & t_{j,s_j} - J_2 & \cdots & t_{j,s_j} - J_j \end{bmatrix}$$

and

$$\beta_{(j+2) \times 1}^{(J)} = (a, b_0, b_1, b_2, \dots, b_j).$$

Similarly, the sampling dates are grouped as  $((t_{0,1}, \dots, t_{0,s_0}), \dots, (t_{k,1}, \dots, t_{k,s_k}))$  by potential knot points. Define

$$X_{N \times 1}^{(K)} = (x_{0,1}, \dots, x_{0,s_0}, x_{1,1}, \dots, x_{1,s_1}, \dots, x_{k,1}, \dots, x_{k,s_k}),$$

$$T_{N \times (k+1)}^{(K)} = \begin{bmatrix} \sin(2\pi\omega_0 t_{0,1} + \theta) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \sin(2\pi\omega_0 t_{0,s_0} + \theta) & 0 & \cdots & 0 \\ \sin(2\pi\omega_0 t_{1,1} + \theta) & \sin(2\pi\omega_0 t_{1,1} + \theta) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sin(2\pi\omega_0 t_{1,s_1} + \theta) & \sin(2\pi\omega_0 t_{1,s_1} + \theta) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \sin(2\pi\omega_0 t_{k,1} + \theta) & \sin(2\pi\omega_0 t_{k,1} + \theta) & \cdots & \sin(2\pi\omega_0 t_{k,1} + \theta) \\ \vdots & \vdots & \ddots & \vdots \\ \sin(2\pi\omega_0 t_{k,s_k} + \theta) & \sin(2\pi\omega_0 t_{k,s_k} + \theta) & \cdots & \sin(2\pi\omega_0 t_{k,s_k} + \theta) \end{bmatrix},$$

and 
$$\beta_{(k+1) \times 1}^{(K)} = (\rho_0, \rho_1, \dots, \rho_k).$$

Note that  $X^{(J)} = X^{(K)} = X$ . The regression curve can be written as

$$X = [T^{(J)} \ T^{(K)}] \begin{bmatrix} \beta^{(J)} \\ \beta^{(K)} \end{bmatrix} + \varepsilon. \quad (\text{A1})$$

If the length of period  $\omega_0$  and the phase  $\theta$  are known (we can assume a yearly period in our applications, which means  $\omega_0 = 1$  and  $\theta$  is known), the above model is simply a multivariate linear regression model. Ordinary least squares can then be used to obtain the parameter estimates.

#### An iteration algorithm for parameter estimation when $\theta$ is unknown

When  $\theta$  is unknown, the following iteration can be used for parameter estimation. For easy notation, we assume again that there is no break point. For a fixed  $\theta$ ,  $\beta = (\beta^{(J)}, \beta^{(K)})$  can be estimated by using the set-up in (A1). For a fixed  $\beta$ , least squares is to minimize

$$\eta = \sum_{i=1}^N \left( x(t_i) - a - \sum_{i=0}^j b_i(t - J_i)_+ - \sum_{i=0}^k \rho_i \delta(t - K_i) \sin(2\pi(t - \theta)) \right)^2.$$

Letting  $\partial\eta/\partial\theta = 0$  and  $z = \sin\theta$ , simple algebra leads to

$$A_4 z^4 + A_3 z^3 + A_2 z^2 + A_1 z + A_0 = 0, \quad (\text{A2})$$

where 
$$A_4 = \left( \sum_{i=1}^N \rho_{s,t_i} \rho_{c,t_i} \right)^2 + \left( \sum_{i=1}^N \rho_{c,t_i}^2 \right)^2,$$

$$A_3 = -2 \left( \sum_{i=1}^N (x(t_i) - a - b_{t_i}) \rho_{c,t_i} \right) \left( \sum_{i=1}^N \rho_{c,t_i}^2 \right) - 2 \left( \sum_{i=1}^N (x(t_i) - a - b_{t_i}) \rho_{s,t_i} \right) \left( \sum_{i=1}^N \rho_{c,t_i} \rho_{s,t_i} \right),$$

$$A_2 = \left( \sum_{i=1}^N (x(t_i) - a - b_{t_i}) \rho_{s,t_i} \right)^2 + \left( \sum_{i=1}^N (x(t_i) - a - b_{t_i}) \rho_{c,t_i} \right)^2 - \left( \sum_{i=1}^N \rho_{c,t_i}^2 \right)^2,$$

$$A_1 = 2 \left( \sum_{i=1}^N (x(t_i) - a - b_{t_i}) \rho_{c,t_i} \right) \left( \sum_{i=1}^N \rho_{c,t_i}^2 \right), \quad \text{and} \quad A_0 = - \left( \sum_{i=1}^N (x(t_i) - a - b_{t_i}) \rho_{c,t_i} \right)^2,$$

with 
$$\rho_{c,t_i} = \sum_{i=0}^k \rho_i \delta(t_i - K_i) \cos(2\pi\omega_0 t_i),$$

$$\rho_{s,t_i} = \sum_{i=0}^k \rho_i \delta(t_i - K_i) \sin(2\pi\omega_0 t_i) \quad \text{and} \quad b_{t_i} = \sum_{i=0}^j b_i(t_i - J_i)_+.$$

The roots of the above biquadratic equation about  $z$  can be solved explicitly; see e.g. Beyer (1981 p. 12).

An iteration algorithm is based on (A1) and (A2) as follows.

- Step 1.** Initialize the parameter estimate of  $\theta$  as  $\theta^{[0]}$ . For example, let  $\theta^{[0]} = 0$ .
- Step 2.** In the  $m$ th iteration ( $m = 1, 2, \dots$ ), given  $\theta = \theta^{[m-1]}$ , estimate  $\beta$  by solving the multivariate linear regression (A1) and denote it by  $\beta^{[m]}$ .
- Step 3.** Given  $\beta = \beta^{[m]}$ , update the estimate of  $\theta$  by solving (A2) and denote it by  $\theta^{[m]}$ .
- Step 4.** Check the differences between the current estimates and the previous estimates. If the maximum difference is less than a predefined tolerance value, then stop the iteration. Otherwise, return to Step 2.

### References

- AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19**, 716–723.
- AN, H.Z. & GU, L. (1984). On the selection of regression variables. In *China–Japan Symposium on Statistics*, pp. 1–3. Beijing, China.
- AN, H.Z. & GU, L. (1986). *Statistical Models and Prediction Methods*. Beijing: Meteorology Publishing House. (In Chinese.)
- BARRY, D. & HARTIGAN, J.A. (1992). Product partition models for change point problems. *Ann. Statist.* **20**, 260–279.
- BEYER, W.H. (1981). *Standard Mathematical Tables*, 26th edn. Florida: CRC Press.
- GALLANT, A.R. & FULLER, W.A. (1973). Fitting segmented polynomial regression models whose join points have to be estimated. *J. Amer. Statist. Assoc.* **68**, 144–147.
- GEISSER, S. (1993). *Predictive Inference: An Introduction*. New York: Chapman & Hall.
- HEINRICH, L. & BENNETT, D. (1997–2000). *AgBores*, version 2.1a. Database Software Pty Ltd and AgWest Bunbury, Western Australia.
- HURVICH, M. & TSAI, C-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.
- HURVICH, M. & TSAI, C-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika* **78**, 499–509.
- LEE, C-B. (1997). Estimating the number of change points in exponential families distributions. *Scand. J. Statist.* **24**, 201–210.
- SEBER, G.A.F. & WILD, C.J. (1989). *Nonlinear Regression*. New York: John Wiley & Sons.
- SUGIURA, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist. Theory Methods* **7**, 13–26.
- TONG, H. (1980). Threshold autoregression, limit cycles and cyclical data, with discussion. *J. Roy. Stat. Soc. Ser. B* **42**, 245–292.
- VERBYLA, A.P., CULLIS, B.R., KENWARD, M.G. & WELHAM, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Appl. Statist.* **48**, 269–311.
- WANG, S.R. & AN, H.Z. (1984). Consistencies on the selection of regression variable. *J. Engrg. Math.* **1**, 13–22. (In Chinese.)
- WEBER, D.D. & ENGLUND, E.J. (1994). Evaluation and comparison of spatial interpolators. *Math. Geol.* **26**, 589–603.
- YAO, Y-C. & AU, S.T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A* **51**, 370–381.
- YIN, Y.Q. (1988). Detection of the number, locations and magnitudes of jumps. *Comm. Statist. Stochastic Models* **4**, 445–455.