Detecting Abrupt Leaks in Blended Underground Storage Tanks

Ryan S. Gill, University of Louisville

Jerome P. Keating, University of Texas at San Antonio

Michael I. Baron, University of Texas at Dallas

October 4, 2004

Abstract

We suggest a sequential multiple change-point algorithm for accurate detection of the onset of abrupt leaks in blended underground storage tanks, and we apply these methods to the Cary blended site data. In addition, we obtain a confidence set for the locations of the change points by inverting the related hypothesis test.

1 Introduction

The Charnock groundwater sub-basin provides the cities of Santa Monica and Culver City with an important source of drinking water. However, the wells operated by the City of Santa Monica were shut down in August 1996 because of MTBE (methyl tertiary butyl ether) pollution resulting from leaking USTs (underground storage tanks). The Southern California Water Company's wells were later shut down during October 1996 to avoid spreading the MTBE contamination.

Contamination of drinking water by MTBE is a major environmental health problem. Gasoline producers use MTBE as a fuel oxygenate which enhances gasoline combustion and reduces tailpipe emissions containing dangerous air pollutants such as carbon monoxide. MTBE has been used in low concentrations since the late 1970's, but it has been used in much higher concentrations since the early 1990's. While its more-widespread use has led to improvements in air quality, the danger posed to the water supply is a major concern. MTBE readily dissolves in water given its high solubility but degrades very slowly under natural conditions so that it forms long plumes which are much more likely to reach the water supply (Keller, Fernandez, Hitz, Kun, Peterson, Smith, Yoshioka, 1998). Even miniscule amounts of the water-soluble chemical can contaminate water supplies, and it is difficult to remove in water treatment. The EPA has classified MTBE as a potential human carcinogen, and MTBE has been linked to several other health problems including asthma and central nervous system irritation. Taste and odor of drinking water is also a major issue with MTBE contamination (Keller, Froines, Koshland, Reuter, Suffet, and Last, 1998).

The City of Santa Monica and the Southern California Water Company now buy replacement water after MTBE contaminated the city reservoir. The

Charnock MTBE Cleanup Project was initiated to identify the sources of contamination by MTBE and other gasoline-related pollutants and to oversee the cleanup efforts by the responsible parties. The responsible oil companies have reached a settlement in which they have agreed to pay for the cleanup of the Charnock wellfield as well as for related damages and legal fees. They have agreed to develop new technology and provide facilities necessary to remove the toxic impurities from the water supply, or they will acquire land to accommodate the new facility. This is done at a total cost of more than 300 million dollars.

The magnitude of the environmental and economic implications of the UST leaks in the Charnock Basin demonstrates the need for an accurate method of estimating the amount of gasoline which has leaked from a UST. Moreover, the propensity for leaks in underground storage tanks, as cited by the 1986 EPA *National Survey of Underground Storage Tanks*, intensifies the need for such methodology. In fact, this 1986 EPA study precipitated congressional action which was codified into leak detection protocol in *The Federal Register* (1988). These EPA guidelines require that every service station be equipped with a leak detection system and that the system must

1. detect leaks of 0.2 gallons per hour, with a probability of 95% and
2. maintain a probability of a false alarm of 5% per month.

In order to comply with these new standards, linear regression methodology can be used to detect leaks in USTs via a process patented by Keating, Dunn, and Dunn (1994). The patented process uses state of the art electronics with computer software, petrochemical information, and statistical methods to detect leaks according to the EPA protocol set out in 1988. This process based on mass reconciliation is far advanced from the industry standard of statistical inventory reconciliation (SIR). The statistical framework for this methodology can be found in Krueger, Keating, Kannan, and Mason (1996).

Many gasoline stations alleviate the problem of leaking UST's by reducing the number of tanks on site from three to two. See Figure 1. They still provide three grades of gasoline but the intermediate (or plus) grade is provided by blending the unleaded with the super blend. The blender produces gasoline with ratings between 87 (unleaded) and 92 (super). For example, a 60:40 blend will produce an intermediate grade of 89 (plus). The blending coefficient, γ, is therefore supposed to be preset to a certain designated amount such as 60%. However, in practice, the value tends to be off slightly, and thus we consider a more complicated estimation problem discussed by Keating and Mason (2000) which expands linear regression methodology to cover blended and manifolded USTs. Keating and Mason (2000) also considered a measurement error approach in which both the tank and meter readings are measured with error.

Consider two tanks and three banks of meters in the following setup. The *tank volumes* (masses) displaced are given by y_1, the volume (mass) displaced from the unleaded storage tank, and y_2, the volume (mass) displaced from the super unleaded storage tank. The *meter volumes* (masses) dispensed are given

by x_1, the volume (mass) of gasoline dispensed through the unleaded meters, x_2, the volume (mass) of gasoline dispensed through the unleaded plus meters, and x_3, the volume (mass) of gasoline dispensed through the super unleaded meters. The *intercepts* are denoted by δ's, where δ_1 is the amount of gasoline displaced from the unleaded UST when no gasoline has been dispensed through the unleaded or unleaded plus meters and δ_2 is the amount of gasoline displaced from the super unleaded UST when no gasoline has been dispensed through the super unleaded or unleaded plus meters. The intercepts, δ_1 and δ_2, are the *leak rates* for the respective tanks. The *slopes* $\beta_i = \frac{\Delta y}{\Delta x_i}$ are the changes in $y = y_1 + y_2$ for a given change in x_i. Thus for each gallon, $\Delta x_i = 1$, of gasoline dispensed Δy is the corresponding amount displaced from the UST. Thus the β's are actually the meter calibration values, which we hope are 1 but by law must fall between 1 ± 0.005 (Federal Register, 1988). Then the *blended UST leak model* is given by

$$\begin{aligned} y_{1j} &= \delta_1 + \beta_1 x_{1j} + \gamma\beta_2 x_{2j} + \epsilon_{1j} \\ y_{2j} &= \delta_2 + (1-\gamma)\beta_2 x_{2j} + \beta_3 x_{3j} + \epsilon_{2j} \end{aligned}$$

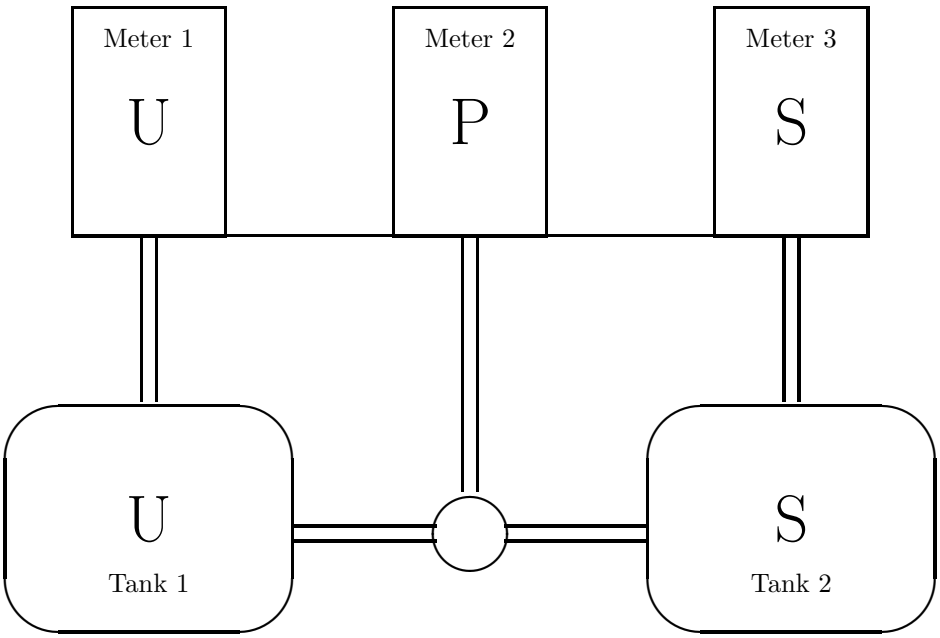

Figure 1: A typical blending site with three meters and two tanks. The plus grade(P) is produced through the blender by mixing gasoline from the unleaded(U) and super(S) tanks.

Table 1: *Estimates of the parameters for blended UST leak model.*

Parameter	δ_1	δ_2	β_1	β_2	β_3	γ	σ
Estimate	1.359	1.620	0.999	1.004	0.996	0.650	3.789

for $j = 1, \ldots, n$ where n is the number of observations and the ϵ's are independent and normally distributed with mean zero and unknown variance σ^2. Hence, we have two regression problems with the issue that both linear equations contain some common and some different predictors and parameters.

This model was used to analyze a sample of 54 observations taken from a blended site in Cary, North Carolina (see Keating and Mason 2000). The observations were taken at 12-hour intervals over 27 consecutive days of operation at the service station site. The least squares estimates of the δ's, β's, and γ's are given in Table 1, as well as an estimate of σ based on the unbiased estimator of the variance. The estimates of the β's and γ's under the measurement error model were nearly identical.

The studentized residual plot for this fit is given in Figure 2, with the residuals corresponding to tank 1 denoted by solid circles and the residuals from tank 2 denoted empty circles. Under our distributional assumptions, not only should all of the residuals be random with mean zero and constant variance but the residuals from the individual tanks should also be random with mean zero and constant variance. However, the residual plots in Figure 2 suggests that the residuals exchange signs around observation 30.

This motivates us to consider the following longitudinal model for *abrupt changes* at a blended site. Here we assume that

$$\begin{aligned} y_{1j}(t) &= \delta_1(t) + \beta_1 x_{1j} + \gamma\beta_2 x_{2j} + \epsilon_{1j} \\ y_{2j}(t) &= \delta_2(t) + (1-\gamma)\beta_2 x_{2j} + \beta_3 x_{3j} + \epsilon_{2j} \end{aligned} \tag{1}$$

for $j = 1, \ldots, n$ with the intercept being modeled as

$$\delta_i(t) = \delta_{i\ell} \text{ for } \nu_\ell < t \leq \nu_{\ell+1}, \ell = 0, 1, \ldots, \eta$$

for $i = 1, 2$, where η is the unknown number of change points, $\nu_0 = 0$, $\nu_{\eta+1} = n$, and $1 \leq \nu_1 < \cdots < \nu_\eta < n$ are the unknown *change-point* parameters. Since abrupt changes are unidentifiable between times of observations (see Gill 2004), we restrict the change-point parameters to be observation times and refer to them by their corresponding observation number. With this in mind, however, it should be understood that it is clearly not true that a change must occur at an observation time; instead, we interpret an estimate $\widehat{\nu} = T$ to mean that the change occurred sometime after observation T but before or at observation $T + 1$.

In this paper, we propose a sequential multiple change-point method for detection of abrupt leaks in blended sites based on the above model. Aspects involved in other sequential methods of estimation of (possibly) multiple change points are discussed in Montgomery (1997), Srivastava and Wu (1999), Kim et al.(2000), and Baron(2004). Here, our method treats the data as though they arrive sequentially. Initially assuming no change points, it uses the method of least squares to estimate the most likely location of a change point. If this estimate is significant, then we fix the location of the change point and continue through the data, searching for another change point. This initial sweep continues until we have cycled through all of the data and obtained an initial set of estimates for our change points. Since each value in the initial set of estimates of the change points is only based on a subset of the data, we update these estimates using an iterative algorithm which is repeated until convergence. Once this is complete, the method produces estimates of the locations of the change points as well as estimates of the other unknown parameters.

This paper is organized as follows. Section 2 gives a detailed mathematical description of our sequential method and derives the estimates and test statistics involved in this method. Section 3 derives a confidence set for the change points given that the number of change points is known. Section 4 analyzes the Cary blended site data based on the work in Sections 2 and 3. Finally, Section 5 summarizes the method and discusses its implications.

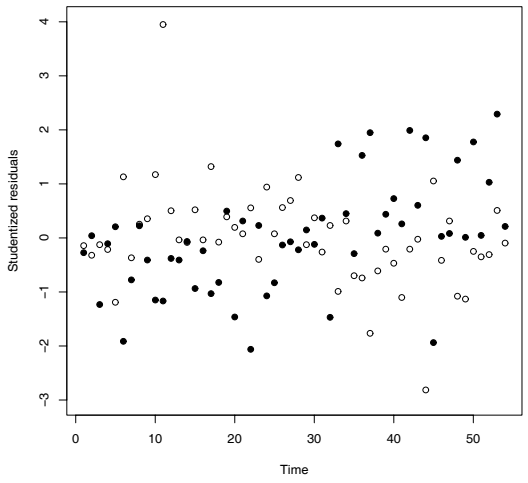

Figure 2: *Studentized residual plot for the blended UST leak model with tank 1 residuals indicated by* • *and tank 2 residuals indicated by* ∘.

2 Sequential multiple change-point algorithm

For a $(2d)$-dimensional vector $\mathbf{x} = [x_1, x_2, \ldots, x_{2d-1}, x_{2d}]'$, define the $(2d)$-dimensional vector

$$\widetilde{\mathbf{x}} = [x_2, x_1, \ldots, x_{2d}, x_{2d-1}]'$$

where the corresponding odd and even components of $\mathbf{x}$ are exchanged. In matrix form, the longitudinal model (1) for abrupt changes at a blended site can be expressed as

$$\mathbf{y} = \mathbf{W}_{\boldsymbol{\nu}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \mathbf{y}^{(n)} = [y_{11}, y_{21}, \ldots, y_{1n}, y_{2n}]',$$

$$\mathbf{W}_{\boldsymbol{\nu}} = \mathbf{W}_{\boldsymbol{\nu}}^{(n)} = [\mathbf{X}_1 \;\; \mathbf{X}_2 - \widetilde{\mathbf{X}}_2 \;\; \widetilde{\mathbf{X}}_2 \;\; \widetilde{\mathbf{X}}_3 \;\; \mathbf{1}_{1:\nu_1} \;\; \widetilde{\mathbf{1}}_{1:\nu_1} \;\; \cdots \;\; \mathbf{1}_{\nu_\eta:\nu_{\eta+1}} \;\; \widetilde{\mathbf{1}}_{\nu_\eta:\nu_{\eta+1}}],$$

$$\mathbf{X}_i = \mathbf{X}_i^{(n)} = [X_{i1}, 0, \ldots, X_{in}, 0]', \quad i = 1, 2, 3,$$

$$\boldsymbol{\beta} = [\beta_1, \gamma\beta_2, \beta_2, \beta_3, \delta_{10}, \delta_{20}, \delta_{11}, \delta_{21}, \ldots, \delta_{1\eta}, \delta_{2\eta}]',$$

$$\boldsymbol{\nu} = \{\nu_1, \ldots, \nu_\eta\},$$

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^{(n)} = [\epsilon_{11}, \epsilon_{21}, \ldots, \epsilon_{1n}, \epsilon_{2n}]',$$

and $\mathbf{1}_{i:j}^{(d)}$ is a $2d \times 1$ vector of ones in its ith through jth odd components and zeros elsewhere. By convention, the superscript will be left off $\mathbf{y}, \mathbf{W}, \mathbf{X}$, and $\boldsymbol{\epsilon}$ when it is (n), and it will be left off $\mathbf{1}_{i:j}$ when the dimension is clear. To estimate the set of change points $\boldsymbol{\nu}$, our method first obtains an initial estimate $\mathbf{V}_0$. Before obtaining the estimate $\mathbf{V}_0$, we first sequentially resample the data and estimate $\mathbf{V}_0^{(m)}$, the number of change points in the subsample from $i = 1$ to $i = m$ for $m = 5, \ldots, n$, conditional on fixed change points at the locations in $\mathbf{V}_0^{(m-1)}$. (Note that at least 5 observations are needed so that the model parameters are identifiable.) That is, beginning with $\mathbf{V}_0^{(4)} = \emptyset$, we estimate $V_0^{(m)}$ by momentarily treating the change points in $\mathbf{V}_0^{(m-1)}$ as fixed and testing for one more additional significant change point before observation m. Finally, we obtain $\mathbf{V}_0$ from $\mathbf{V}_0^{(n)}$ by applying the binary segmentation method (see Vostrikova 1981 and Kuo and Yang 2001).

Each value in $\mathbf{V}_0^{(n)} \subset \mathbf{V}_0$ is obtained using only the observations required to detect that respective change and is not based on any later observations. However, we do have all n observations, so after the initial estimates are obtained using the sequential procedure described above, we must go back and re-estimate $\boldsymbol{\nu}$. Consequently, if $\mathbf{V}_0$ is nonempty, the method iteratively cycles through the data obtaining a revised estimate $\mathbf{V}_i$ from $\mathbf{V}_{i-1}$ until there is no change by sequentially re-estimating and re-testing the $\widehat{\eta}_{i-1}$ change points in $\mathbf{V}_{i-1}$ one-at-a-time.

This method allows us to look for one change point at a time which makes the necessary work computationally feasible. To a large extent, it also avoids the inherent contamination problem in the binary segmentation method.

2.1 Initial estimates of $\boldsymbol{\nu}$

First, we obtain an initial estimate $\mathbf{V}_0$ of the set of change points $\boldsymbol{\nu}$. We begin with $\mathbf{V}_0^{(4)} = \emptyset$ and proceed sequentially for $m = 5, \ldots, n$ as follows. Given the set of change points $\mathbf{V}_0^{(m-1)}$, we wish to test for an additional change point. Specifically, we perform sequential α-level tests of the null hypothesis $H_{00}^{(m)}$ that there are $\widehat{\eta}_0^{(m-1)}$ change points at the points in $\mathbf{V}_0^{(m-1)}$ versus the alternative that there are $\widehat{\eta}_0^{(m-1)} + 1$ change points, $\widehat{\eta}_0^{(m-1)}$ of which are in $\mathbf{V}_0^{(m-1)}$ where $\widehat{\eta}_i^{(m)} = \|\mathbf{V}_i^{(m)}\|$.

Given the estimate of the set of change points $\mathbf{V}_0^{(m-1)}$ obtained from the first $m-1$ observations, denote the conditional least squares estimate of a new change point as

$$v_{0,m} = \underset{k \in \{2,\ldots,m-1\}}{\operatorname{argmin}} S(\mathbf{U}_{0,k}^{(m)}, \mathbf{y}^{(m)})$$

where $\mathbf{U}_{0,k}^{(m)} = \mathbf{V}_0^{(m-1)} \cup \{k\}$, $\mathbf{P}_{\mathbf{U}} = \mathbf{P}_{\mathbf{U}}^{(n)} = \mathbf{W}_{\mathbf{U}}^{(n)} \left((\mathbf{W}_{\mathbf{U}}^{(n)})' \mathbf{W}_{\mathbf{U}}^{(n)} \right)^{-1} (\mathbf{W}_{\mathbf{U}}^{(n)})'$, $\mathbf{I}_d$ is the $d \times d$ identity matrix, and

$$S(\mathbf{U}) = S(\mathbf{U}, \mathbf{y}) = \min_{\mathbf{b}} \|\mathbf{y} - \mathbf{W}_{\mathbf{U}}^{(n)} \mathbf{b}\|^2 = \|(\mathbf{I}_{2n} - \mathbf{P}_{\mathbf{U}})\mathbf{y}\|^2$$

where, by convention, the subscript is omitted from $\mathbf{P}$ when it is (n) and the second argument is omitted from S when it is $\mathbf{y}$. From our normality assumption, $v_{0,m}$ is also the conditional maximum likelihood estimate of a new change point in $\{1, \ldots, m\}$.

To test $H_{00}^{(m)}$, we consider the test statistic

$$\Lambda_0^{(m)} = \min_{k \in \{2,\ldots,m-1\}} S(\mathbf{U}_{0,k}^{(m)}, \mathbf{y}^{(m)}) / S(\mathbf{V}_0^{(m-1)}, \mathbf{y}^{(m)}) \tag{2}$$

which is equivalent to the generalized likelihood ratio test statistic, and we want to reject $H_{00}^{(m)}$ if it is less than or equal to the αth percentile of the distribution of $\Lambda_0^{(m)}$ under $H_{00}^{(m)}$. Unfortunately, the distribution of $\Lambda_0^{(m)}$ is complicated, so we instead perform a simulation-based test of the null hypothesis that the proportion of times that a value simulated from the distribution of $\Lambda_0^{(m)}$ under $H_{00}^{(m)}$ exceeds the observed value $\Lambda_0^{(m)}$ is α against the alternative that it is greater than α. When $H_{00}^{(m)}$ is true, we can rewrite (2) as

$$\Lambda_0^{(m)} = \min_{k \in \{2,\ldots,m-1\}} S(\mathbf{U}_{0,k}^{(m)}, \boldsymbol{\epsilon}^{(m)}/\sigma) / S(\mathbf{V}_0^{(m-1)}, \boldsymbol{\epsilon}^{(m)}/\sigma) \tag{3}$$

where $\boldsymbol{\epsilon}^{(m)}/\sigma$ follows a $2m$-dimensional normal distribution with mean vector $\mathbf{0}_{2m,1}$ and covariance matrix $\mathbf{I}_{2m}$. Hence, (3) is independent of $\boldsymbol{\beta}$ and σ, and we can approximate the distribution $\Lambda_0^{(m)}$ based on R simulations of $\boldsymbol{\epsilon}^{(m)}/\sigma$. Then let $B^{(m)}$ be the number of simulated values which are less than or equal to our observed $\Lambda_0^{(m)}$. If the true proportion is α, then $B^{(m)}$ follows a binomial

distribution based on R trials with probability of success α. So we reject the hypothesis of no new change if $B^{(m)}$ is less than or equal to the αth percentile of this binomial distribution; the ratio of this percentile divided by R approaches α as R approaches ∞. So at each step our updated set of change points is $\mathbf{V}_0^{(m)} = \mathbf{U}_{0,v_{0,m}}^{(m)}$ if $H_{00}^{(m)}$ is rejected, but remains the same otherwise.

After cycling through all of the data, we obtain $\mathbf{V}_0^{(n)}$. If $H_{00}^{(n)}$ is rejected, then we should test for additional significant change points. Here we test the null hypothesis H_{00} that there are $\widehat{\eta}_0^{(n)}$ change points at the points in $\mathbf{V}_0^{(n)}$ against the alternative that there are $\widehat{\eta}_0^{(n)} + 1$ change points, $\widehat{\eta}_0^{(n)}$ of which are at the points in $\mathbf{V}_0^{(n)}$, by considering the test statistic

$$\Lambda_0 = \min_{k \in \{2,\ldots,n-1\}} S(\mathbf{U}_{0,k}) / S(\mathbf{V}_0^{(n)})$$

where $\mathbf{U}_{0,k} = \mathbf{V}_0^{(n)} \cup \{k\}$. Using a test similar to that described above for $H_{00}^{(m)}$, we obtain our initial set of estimates of the change points $\mathbf{V}_0 = \mathbf{V}_0^{(n)}$ if we fail to reject H_{00}, or we repeat this final step with $\mathbf{V}_0$ replaced by $\mathbf{V}_0^{(n)} \cup \{v_0\}$ where

$$v_0 = \operatorname*{argmin}_{k \in \{2,\ldots,n-1\}} S(\mathbf{U}_{0,k})$$

if H_{00} is rejected.

2.2 Revision algorithm

The algorithm described below allows us to update $\mathbf{V}_{i-1}$ where i is any positive integer. Beginning with the earliest estimated change point in $\mathbf{V}_i^{(0)} = \mathbf{V}_{i-1}$, we use the following algorithm for the data set partitioned by the $\widehat{\eta}_{i-1} = \widehat{\eta}_i^{(0)}$ estimated change points. First, for $m = 1$, we remove the mth smallest element of $\mathbf{V}_i^{(m-1)}$, say $v_{(m)}$, and form the set

$$\mathbf{V}_i^{(-m)} = \mathbf{V}_i^{(m-1)} - \left\{v_{(m)}\right\}.$$

Then we test the null hypothesis $H_{0i}^{(m)}$ that there are $\widehat{\eta}_i^{(m-1)}$ change points, $\widehat{\eta}_i^{(m-1)} - 1$ of which are at the points in $\mathbf{V}_i^{(-m)}$, against the alternative that there are $\widehat{\eta}_i^{(m-1)}$ change points at the points in $\mathbf{V}_i^{(-m)}$ by considering the test statistic

$$\Lambda_i^{(m)} = \min_{k \in \{2,\ldots,n-1\}} S(\mathbf{U}_{i,k}^{(m)}) / S(\mathbf{V}_i^{(-m)})$$

where $\mathbf{U}_{i,k}^{(m)} = \mathbf{V}_i^{(-m)} \cup \{k\}$. Using a test similar to that described in Section 2.1 for $H_{00}^{(m)}$, our update set of change points is $\mathbf{V}_i^{(m)} = \mathbf{U}_{i,v_{i,m}}^{(m)}$ where

$$v_{i,m} = \operatorname*{argmin}_{k \in \{2,\ldots,n-1\}} S(\mathbf{U}_{i,k}^{(m)})$$

if $H_{0i}^{(m)}$ is rejected, but we remove $v_{(m)}$ to obtain $\mathbf{V}_i^{(m)} = \mathbf{V}_i^{(-m)}$ otherwise. We repeat this step for $m = 2, \ldots, \widehat{\eta}_{i-1}$ to obtain $\mathbf{V}_i^{(\widehat{\eta}_{i-1})}$.

If we reject $H_{0i}^{(\widehat{\eta}_{i-1})}$, then we proceed to test for additional change points as we did at the end of the initial cycle until none are significant. The test, say H_{0i}, is similar to H_{00} with $\mathbf{V}_0^{(n)}$ replaced by $\mathbf{V}_i^{(\widehat{\eta}_{i-1})}$. Finally, we obtain our updated estimate of the set of change points $\mathbf{V}_i$. If $\mathbf{V}_i = \mathbf{V}_{i-1}$, then our final estimate of the set of change points is $\widehat{\boldsymbol{\nu}} = \mathbf{V}_i$; otherwise, we must update $\mathbf{V}_i$.

2.3 Estimates of other parameters

Upon completion of updating the estimates of $\widehat{\boldsymbol{\nu}}$, we obtain estimates of the other parameters. The estimate of the number of change points is $\widehat{\eta} = \|\widehat{\boldsymbol{\nu}}\|$. The meter calibration estimates, $\widehat{\beta}_1, \widehat{\beta}_2$, and $\widehat{\beta}_3$, and the estimates of the leak rates, $\widehat{\delta}_{10}, \widehat{\delta}_{20}, \widehat{\delta}_{11}, \widehat{\delta}_{21}, \ldots, \widehat{\delta}_{1\widehat{\eta}}$, and $\widehat{\delta}_{2\widehat{\eta}}$, are elements of the vector

$$\widehat{\boldsymbol{\beta}} = \left((\mathbf{W}_{\widehat{\boldsymbol{\nu}}})' \mathbf{W}_{\widehat{\boldsymbol{\nu}}}\right)^{-1} \mathbf{W}_{\widehat{\boldsymbol{\nu}}} \mathbf{y},$$

which is the conditional least squares estimate of $\boldsymbol{\beta}$ given that there are $\widehat{\eta}$ change points. An ad hoc conditional estimate of the blending coefficient is

$$\widehat{\gamma} = \frac{\widehat{\gamma\beta}_2}{\widehat{\beta}_2}.$$

Finally, we estimate the variance by

$$\widehat{\sigma}^2 = \|\mathbf{y} - \mathbf{W}_{\widehat{\boldsymbol{\nu}}} \widehat{\boldsymbol{\beta}}\|^2 / (6 + 2\widehat{\eta}).$$

3 Confidence estimation of $\boldsymbol{\nu}$ when η is known

In this section, we assume the number of change points η is fixed, and we derive an approximate $100(1-\alpha)\%$ confidence set for $\boldsymbol{\nu}$; this confidence set $C = C(\mathbf{y}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ should approximately satisfy

$$\mathbb{P}(\boldsymbol{\nu} \in C \mid \eta) = 1 - \alpha.$$

We construct the largest possible $100(1-\alpha)\%$ confidence set by inverting the test of the null hypothesis $H_{0,\boldsymbol{\nu}_0}$ that the η change points are at locations in a set $\boldsymbol{\nu}_0$ which contains η elements versus the alternative that this is not the case. The test statistic for this hypothesis is

$$\Lambda_* = \min_{\mathbf{U} \in \mathcal{U}(\eta)} S(\mathbf{U}) / S(\boldsymbol{\nu}_0).$$

where $\mathcal{U}(\eta)$ is the set of all possible combinations of η change points among the n observations. In this case, the distribution of Λ_* depends on $\boldsymbol{\beta}$ and σ so that we must estimate these parameters under the assumption that $\boldsymbol{\nu} = \boldsymbol{\nu}_0$ and use these estimates $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\sigma}_0$ to simulate values of Λ_*. So we approximate the

distribution of the generalized likelihood ratio test statistic Λ_* by generating R simulated samples of $\boldsymbol{\epsilon}/\sigma$ and using it to calculate

$$\widehat{\Lambda}_* = \min_{\mathbf{U}\in\mathcal{U}(\eta)} S(\mathbf{U}, \mathbf{W}\widehat{\boldsymbol{\beta}}_0 + \boldsymbol{\epsilon})/S(\boldsymbol{\nu}_0, \mathbf{W}\widehat{\boldsymbol{\beta}}_0 + \boldsymbol{\epsilon}).$$

Then we test $H_{0,\boldsymbol{\nu}_0}$ by simulation similar to the method discussed in Section 2.1. Finally, the set C is comprised of all choices of $\boldsymbol{\nu}_0$ such that $H_{0,\boldsymbol{\nu}_0}$ is not rejected; that is, all tests for which the p-value of the test of $H_{0,\boldsymbol{\nu}_0}$ is greater than α.

4 Analysis of Cary blended site data

Now we apply the method in Sections 2 and 3 to the Cary blended site data. First, we use the sequential multiple change point algorithm to estimate the location(s) of significant change points. For practical purposes, we acknowledge that there will be some observations which do not fit the model for various reasons such as theft, and we do not want to include these points as they will affect our estimates of the parameters. In the end, we want the intercepts to represent leak rates or at least sources of consistent daily loss. Fortunately, our change-point algorithm identifies outliers by its own means. Next, we obtain a confidence set for the change point given that there is one change. All of the tests are performed at a 5% level based on $R = 100,000$ simulations so that the effective empirical level is 4.886%; that is, if B follows a binomial distribution based on R trials and probability of success 0.05, then $b = 4,886$ is the largest integer such that $\mathbb{P}(B \leq b) \leq 0.05$.

4.1 Estimation

Using the method described in Section 2.1, the initial estimate of the set of significant change points is $\mathbf{V}_0 = \{10, 11, 32, 44, 45\}$ where $v_{0,11} = 10$ and the empirical p-value of $H_{00}^{(11)}$ is $p = 0.02844$, $v_{0,13} = 11$ with $p = 0.01966$, $v_{0,33} = 32$ with $p = 0.00504$, $v_{0,45} = 44$ with $p = 0.00162$, and $v_{0,49} = 45$ with $p = 0.03567$. Then we attempt to revise $\mathbf{V}_0$ according to the revision algorithm in Section 2.2. However, there are no changes in the updated estimates $v_{i,m}$ of the change points and the respective empirical p-values of $H_{01}^{(i)}, i = 1, \ldots, 5$, are 0.00018, 0.00033, 0, 0.00255, and 0.01663. Hence, the estimate of $\boldsymbol{\nu}$ based on the entire data set is $\widehat{\boldsymbol{\nu}} = \mathbf{V}_0$. The estimates of the other parameters are given in Table 2 according to the formulas in Section 2.3.

Upon viewing the results for the regression coefficients based on these estimates, it appears that observations 11 and 44 do not fit the model because of large negative estimates of the intercepts. In any case where we have a change followed by an immediate reverse change, we must decide whether to treat the observation as an outlier and remove it or to treat it as two consecutive change points within the overall multiple change-point model. Here we claim these observations are outliers since the estimates of its associated leak rates cannot be

Table 2: *Regression coefficients for model with change points at 10, 11, 32, 44, and 45.*

Parameter	δ_{10}	δ_{20}	δ_{11}	δ_{21}	δ_{12}	δ_{22}
Estimate	-0.435	2.139	-2.494	17.607	-0.068	3.203
Parameter	δ_{13}	δ_{23}	δ_{14}	δ_{24}	δ_{15}	δ_{25}
Estimate	5.312	-0.941	-5.976	6.468	4.541	1.013
Parameter	β_1	β_2	β_3	γ	σ	
Estimate	0.999	1.001	0.999	0.653	2.668	

Table 3: *Regression coefficients for model with change points at 32, 43, and 45 after removing the outliers 11 and 44.*

Parameter	δ_{10}	δ_{20}	δ_{11}	δ_{21}	
Estimate	-0.026	2.456	5.131	-0.752	
Parameter	δ_{12}	δ_{22}	δ_{13}	δ_{23}	
Estimate	-6.025	6.011	4.668	0.569	
Parameter	β_1	β_2	β_3	γ	σ
Estimate	0.999	1.003	0.999	0.653	2.523

interpreted within the overall model. So we will re-apply the algorithm to the data with these observations removed. Notice that these two points were clearly outliers based on the simpler regression fit in Figure 2 as well.

With the outliers 11 and 44 removed, the initial estimate of the set of change points is $\mathbf{V}_0 = \{32, 43, 45\}$ where $v_{0,33} = 32$ and the empirical p-value of $H_{00}^{(33)}$ is 0.00382, $v_{0,45} = 42$ with $p = 0.00056$, and $v_{0,48} = 45$ with $p = 0.01696$. Again, the revision step produces no changes; the respective empirical p-values for $H_{01}^{(i)}, i = 1, 2, 3$ are 0, 0.00177, and 0.00587. The estimates of the other parameters are given in Table 3. However, we see that this produces another large negative estimate of an intercept at observation 45.

Therefore, we remove 45 in addition to 11 and 44, and we re-apply the algorithm. The initial estimate of the set of change points is $V_0 = \{32\}$ where $v_{0,33} = 32$ and the empirical p-value of $H_{00}^{(33)}$ is 0.00382. Again the revision step produces no changes; the empirical p-value for $H_{01}^{(1)}$ is 0. Finally, the estimates of the other parameters are given in Table 4 and the studentized residual plot of this fit is given in Figure 3. Neither the parameter estimates nor the residual plot indicate any clear deviations from our model.

4.2 Interpretation of estimates

In section 4.1, we see that the best estimate of the set of change points based on our model is $\widehat{\boldsymbol{\nu}} = \{32\}$. This indicates that the only significant change in the

Table 4: *Regression coefficients for model with a change point at 32 after removing the outliers 11, 44, and 45.*

Parameter	δ_{10}	δ_{20}	δ_{11}	δ_{21}	
Estimate	-0.034	2.542	4.920	-0.068	
Parameter	β_1	β_2	β_3	γ	σ
Estimate	0.998	1.004	0.998	0.652	2.515

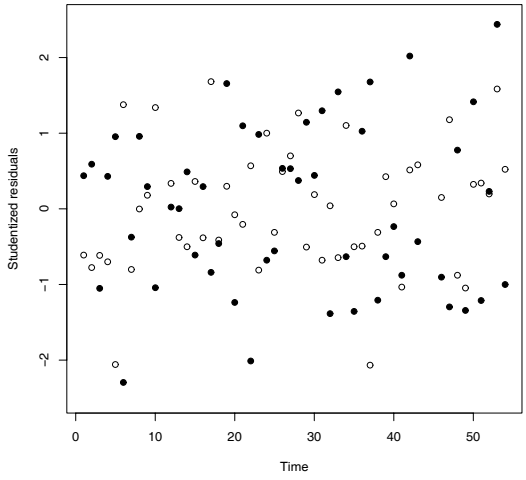

Figure 3: *Studentized residual plot for the model with a change point at 32 with tank 1 residuals indicated by* • *and tank 2 residuals indicated by* ∘.

consistent leak rates of the USTs occurred at observation 32. This is consistent with the change we noticed in the residual plot in Figure 2. These data indicate a leak rate of 2.542 gallons per half day from tank 2 in the first 32 observations and no leak after observation 32; also, these data indicate no leak through tank 1 in the first 32 observations but a leak rate of 4.920 gallons per half day after observation 32. These patterns are shown in Figure 4.

The data are from a demonstration test to check the efficacy of the leak detection methodology and simulate a worst case scenario. An "artificial" leak was induced by extracting .2 gph (see standards) out of tank 2 for the first 32 observations. Beginning on observation 33, the artificial extraction from tank 2 was stopped and a new one from tank 1 was initiated of .4 gph. As one can see, the proposed algorithm is most effective in

1. detecting and estimating change points

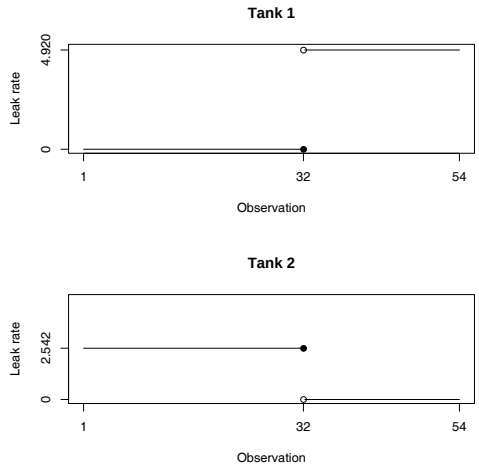

Figure 4: *Estimated tank leak rates for the Cary blended data.*

2. correctly estimating the magnitude of the leaks.

This methodology is useful in quantifying the total amount of gasoline that has leaked into the ecosystem. For example, in tank 2 we have an estimated spill of 2.542 gallons per half day for 32 observations for a total spill of 40.672 gallons. In tank 1, we have an estimated spill of 4.920 gallons/half day for 22 days for a total spill of 54.12 gallons. These quantities can be used by the EPA to assess damages and assign fines to polluters. This methodology provides a quntified estimate of the total spill by identifying the time of onset and the magnitude of the spill.

By some estimates, as much as 6,000 gallons of gasoline was spilled into the Charnock Basin by three service stations on Sepulveda Street in Santa Monica, CA. As the state of California and the EPA try to assess the total damage and assess punitive fines, companies that did not find leaks of a magnitude of 4.8 gpd within 30 days were clearly not compliant with EPA standards.

4.3 Confidence estimation

In this section, we use the method proposed in Section 3 to find an approximate 95% confidence set for $\boldsymbol{\nu}$ when $\eta = 1$. Clearly, observation 32 is in C since it is the maximum likelihood estimate so that our observed test statistic is 1. The observation with the next smallest empirical p-value is 30 with $p = 0.01751$. Consequently, this 95% confidence set is trivially $\{32\}$.

In fact, we can state this with an even stronger confidence level. Let's consider a different question. Based on our approximations related to simulations, what is the largest confidence level for which we do still obtain a

trivial set containing only the maximum likelihood estimate? To answer this, we determine the level α under which the largest observed empirical p-value (other than in testing observation 32) has probability α. In the Cary blended data, this corresponds to observation 30 so that the desired α is 0.018397 since $\mathbb{P}(B \leq 1,751) = 0.018397$ if B is a binomial random variable based on 100,000 trials and probability of success 0.018397. Thus, we can state with 98.1602% confidence that the change occurred after observation 32 but before or at 33.

5 Conclusions

In this paper, we have proposed a sequential algorithm for the detection of multiple change points in the longitudinal model for abrupt changes at a blended site. For a fixed number of change points, we have given a confidence set for the set of change points based on inverting the related hypothesis test. For the Cary blended site, we found that there was one change point after removing three outliers which did not fit the model. This observation is the only one in a 98% confidence set.

Given the data of the tank and meter readings, the algorithm proposed in this paper can be applied to the UST leak detection problem in any cases where we know of or suspect a leak such as in the Charnock Basin. Gasoline distributors by declaration of being compliant with the 1988 EPA guidelines have indirectly implied that they can provide such data. The environmental and economic implications of the leaks in the Charnock Basin demonstrate the need for accurate assessment of UST leaks. The sequential algorithm proposed in this paper provides much insight into the nature of a leak; that is, we can learn how large a leak is at a given time, where the leak is coming from, and when the leak started and/or changed.

References

[1] National survey of underground storage tanks. *EPA Report 560/5-86-013*, 1986.

[2] *The Federal Register.* (Friday September 23) 53(125):37134–37162, 1988.

[3] M. Baron. Sequential methods for multistate processes. In *Applications of Sequential Methodologies*, N. Mukhopodadhyay, S. Datta, S. Chattopadhyay, eds. Marcel Dekker, New York, 57–71, 2004.

[4] R. Gill. Maximum likelihood estimation in generalized broken-line regression. *The Canadian Journal of Statistics*, 32, 2004 (to appear).

[5] J. P. Keating, W. Dunn, and D. Dunn. Storage tank and line leakage detection and inventory reconciliation process. Patent No. 5,297,423, 29 March, 1994.

[6] J. P. Keating and R. L. Mason. Using statistical models to detect leaks in underground storage tanks. *Environmetrics*, 11:395–412, 2000.

[7] A. A. Keller, L. F. Fernandez, S. Hitz, A. Peterson, B. Smith, and M. Yoshioka. An integral cost-benefit analysis of gasoline formulations meeting california phase II reformulated gasoline requirements. *UC TSR&TP Report to the Governor of California*, 1998.

[8] A. A. Keller, J. Froines, C. Koshland, J. Reuter, I. Suffet, and J. last. Health & environmental assessment of MTBE, summary and recommendations. *UC TSR&TP Report to the Governor of California*, 1998.

[9] H.-J. Kim, M. P. Fay, E. J. Feuer, and D. N. Midthune. Permutation Tests for Joinpoint Regression with Applications to Cancer Rates. *Statistics in Medicine*, 19:335–351, 2000.

[10] C. D. Krueger, J. P. Keating, N. Kannan, and R. L. Mason. Calibrating gasoline flow meters. In *Statistics for Quality: A volume in Honor of Don Owen*, pages 49–76. Marcel Dekker, New York, 1996.

[11] L. Kuo and T. Yang. Bayesian Binary Segmentation Procedure for a Poisson Process. *Journal of Computational and Graphical Statistics*, 10:772–785, 2001.

[12] D. C. Montgomery. *Introduction to Statistical Quality Control, Third Ed.* Wiley, New York, 1997.

[13] M. S. Srivastava and Y. Wu. Quazi-stationary biases of change point and change magnitude estimation after sequential CUSUM test. *Sequential Analysis*, 18:203–216, 1999.

[14] L. J. Vostrikova. Detecting disorder in multidimensional random process. *Soviet Mathematics Doklady*, 24:55–59, 1981.