Document Review and Approval
Environmental Hazards Assessment Program
Department of Pesticide Regulation
830 K Street
Sacramento, CA 95814

Document Title: Statistical Approaches to Assessing Pesticide Concentrations in the DPR Surface Water Database

Author(s): Robert H. Shumway

Document Date: April 30, 2001

APPROVED: Patricia Dunn Date: 8-23-01
Senior Environ. Research Scientist
(Supervisor)

APPROVED: [signature] Date: 9/13/01
Senior Environ. Research Scientist
(Specialist)

APPROVED: [signature] Date: 9/17/01
Ag. Program Supervisor

APPROVED: John S. Sanders Date: 9/19/01
John Sanders, Ph.D.
Branch Chief

April 30, 2001

STATISTICAL APPROACHES TO ASSESSING PESTICIDE CONCENTRATIONS IN THE DPR SURFACE WATER DATABASE

Robert H. Shumway

Department of Statistics
University of California, Davis

Final Report
Environmental Protection Agency
Department of Pesticide Regulation
830 K. St., Room 200
Sacramento, CA 95814-3510

Agreement No. 99-0248
June 1, 2000 - December 31, 2000

Project Coordinators: Patricia Dunn and Keith Starner

STATISTICAL APPROACHES TO ASSESSING PESTICIDE CONCENTRATIONS IN THE DPR SURFACE WATER DATABASE

Robert H. Shumway

Department of Statistics
University of California, Davis

Abstract

Various studies on measured levels of pesticides in water have been completed and data from these studies are currently available on a surface water database maintained by the Department of Pesticide Regulation (DPR). We examine portions of two databases in this report and suggest statistical methodology for approaching questions of importance in monitoring pesticide levels in water. The question of primary interest is the merging of concentrations, measured over time, space and sampling technique, into a coherent pesticide signal from a given region. This involves handling time series that are irregularly observed, due to missing observations or different sampling rates We consider models over time and space that lead to prediction limits for the pesticide signal during episodes where the concentrations may be exceeding regulatory standards. A recommended exploratory data analysis and modeling procedure is developed and applied to concentrations of various pesticides, focusing on chlorpyrifos and diazinon, measured in a Dow Chemical Study on Orestimba Creek and diazinon and simazine in a U.S. Geological Survey study along the San Joaquin River.

Key Words: *Water quality regulation, pesticide concentrations, exploratory data analysis, missing data, state-space model, dynamic regression, signal extraction.*

1. Introduction: Review of Data

The Department of Pesticide Regulation (DPR) is currently creating a large database consisting of concentrations of various pesticides in surface water. For regulatory purposes, it would be useful to have some standard methodological techniques for accessing this data and for making assertions about the levels of pesticides in a given area and to the uncertainties that can be attributed to these levels.

The data used for this preliminary study are only a fraction of that currently available for analysis in the DPR surface water database. With the assistance of DPR personnel, two large files were accessed, containing measurements of pesticide levels on Orestimba Creek and the San Joaquin River at Vernalis, respectively.

The first database contains measurements taken in a study by Poletika and Robb (1998) in the Orestimba Creek tributary of the San Joaquin River on the pesticides chlorpyrifos, diazinon and methidathion during the period April, 1996 to May, 1997. Ideally, there would be 364 daily observations during this period at three locations, the State Highway 33 bridge, the Crow Creek drain and River Road. Measurements were taken hourly and merged into a daily composite. For the River Road location, weekly grab samples were available. This file contained roughly 3000 lines with location, sampling date, extraction and analysis dates, chemical code, detection limit, and sampling method by code and type. The observations are intermingled by date and there are many below the published detection limits. There were also dates with no available observations. Figure 1 shows a time plot for the pesticides chlorpyrifos and diazinon at the three locations, where the weekly grab samples also available from the River Road location. A complicating feature of this kind of data is the tendency for detections to involve very large oscillations of fairly short duration. This complicates the treatment of the data over long time periods and we concentrateon the local behavior of high concentration "episodes" .

The plots suggest a number of questions that will be evaluated in this report. First, there is the previously mentioned behavior over time, which tends to involve clusters of sporadic episodes that show relatively high concentrations. These bursts tend to occur simultaneously at all locations, implying that there might be consistent dynamics over space as well as time. The simultaneous occurrence of the different episodes involving two different pesticides over the same interval of time would be of interest, either at the same or different locations. For example, chlorpyrifos and diazinon appear to occur together at the River Road location.

The second database, collected in the U.S. Geological Survey of MacCoy et al (1995), involves two-day combined samples drawn from the San Joaquin River at Vernalis, over a 1202-day time period extending from January, 1991 to April, 1994. This file contained about 10,000 lines involving measurements on 23 pesticides. Figure 2 shows daily concentrations

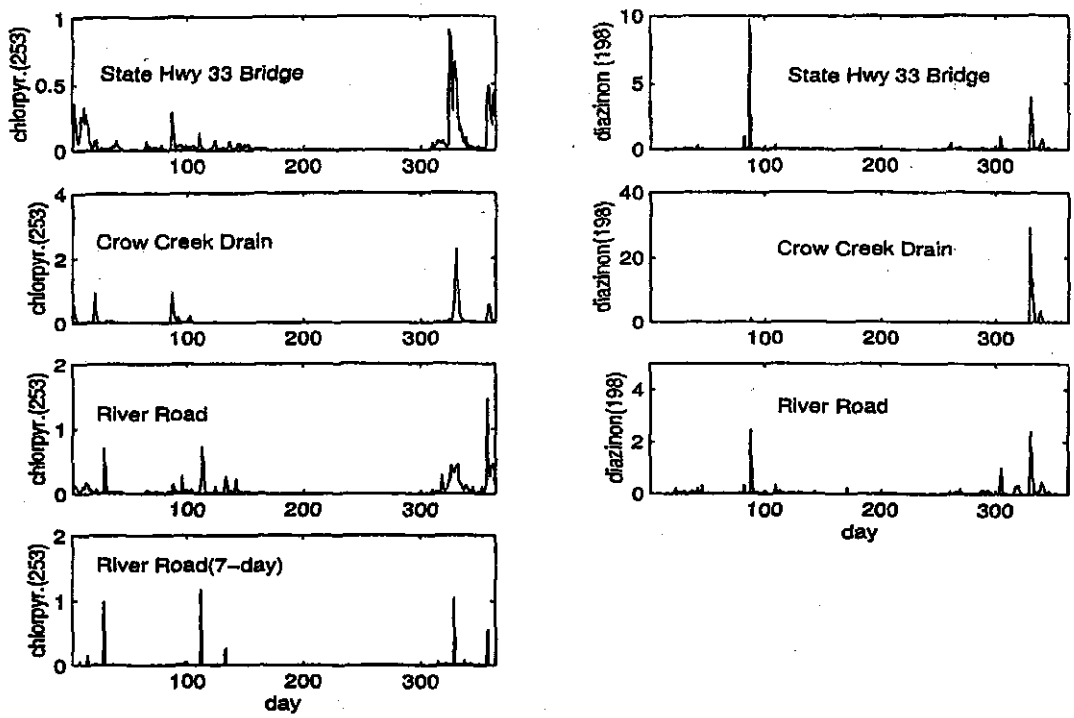

Figure 1: Daily chlorpyrifos and diazinon concentrations at three locations on Orestimba Creek (May,1996-May,1997). A seven-day grab sample is available for chlorpyrifos at the River Road location.

of two-day combined samples from the the pesticides diazinon, simazine, cyanazine and metolochlor; no chlorpyrifos was detected during this time period.

The plots suggest that high concentration episodes of diazinon and simazine were roughly concurrent but other questions of interest involving coincident locations could not be evaluated using this data. A single sampling method was applied so that there is no opportunity for comparing sampling methods and the absence of chlorpyrifos detections means that looking at its co-ocurrence with diazinon would not be possible. We can, however, look at the parallel behavior of measurements over time within a given episode.

The erratic behavior of the pesticide concentrations over time, as well as the high level of non-detections together suggest that exploratory data analysis will be an essential step for handling such data in the future. In Section 2, we discuss the use of transformations to scale down the large fluctuations and the application of various measures of correlation over time and space to indicate the kinds of models that might be applied to answer questions of interest for regulation and mitigation. Section 3 applies a dynamic regression model in space and time to suggest possible relations across those two dimensions. Section 4 uses the conclusions of Sections 2 and 3 to build a plausible model for common episodic signals over

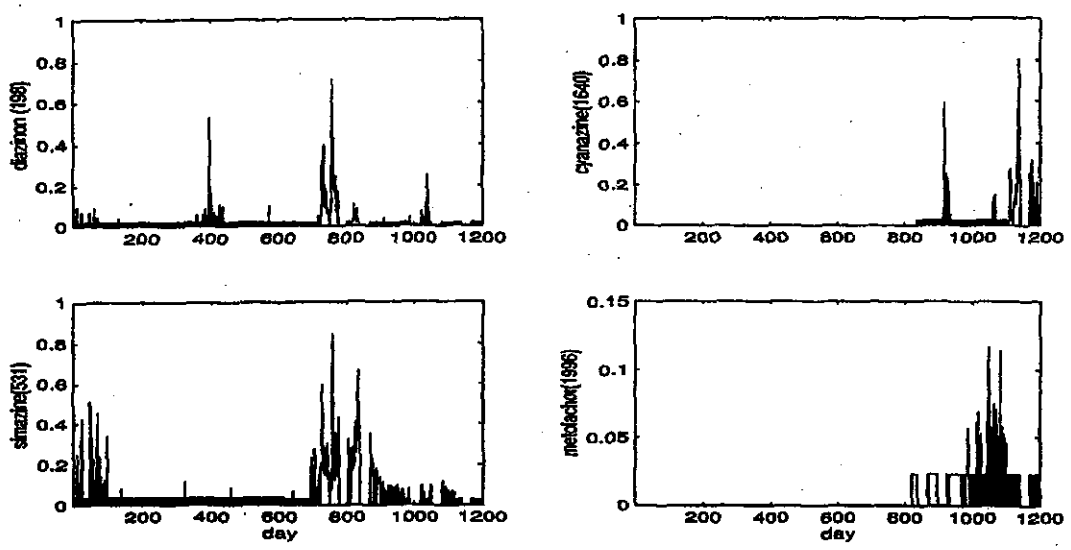

Figure 2: Daily concentrations of diazinon, simazine, cyanazine and metolachlor on the San Joaquin River at Vernalis (January, 1991- April 1994, 1202 days, USGS)

a given short time period. The model enables estimation of a common signal over space and its prediction limits. We use the distribution of the signal to give a lower bound for the probability of simultaneously exceeding some given standard for a dpecified number of time periods.

2. Exploratory Data Analysis

This section summarizes a number of possible exploratory data analysis techniques that may suggest ways of reducing the data to a form that is more susceptible to statistical modeling. First, we discuss the use of transformations, which can help the data conform to standard statistical assumptions made in modeling such as stationarity, linearity and approximate normality. Another important diagnostic for this particular data will be various forms of correlation, namely, simple correlation for relating two series and a scatter diagram, which shows possible nonlinearities in the relation. Correlations over various lags, such as the autocorrelation and cross correlation functions are important for evaluating the time varying behavior of the series. All measures are discussed in Shumway and Stoffer (2000, Chapter 1) and can computed via standard packages or through the package ASTSA that is available in McQuarrie and Shumway (1994); this package and documentation can be downloaded from the web site in the reference.

It is recognized that other statistical packages such as Minitab, SAS, SPSS and S-Plus will all have provisions for analyzing time series that are fully observed and not subject to detection limits, with S-Plus containing the most options and a modernized treatment.

However, the analysis undertaken here involves data that are sparse and subject to strong censoring, requiring that a pre-processing model-fitting procedure using Kalman filtering and smoothing be applied to produce a continuous record in time for input to standard packages. As a practical matter, the computations done in this report used MATLAB, which has superior computational and graphics capabilities for the purposes of reports and is satisfactory for manipulating large files.

2.1 Estimation of Frequency Distributions; Transformations

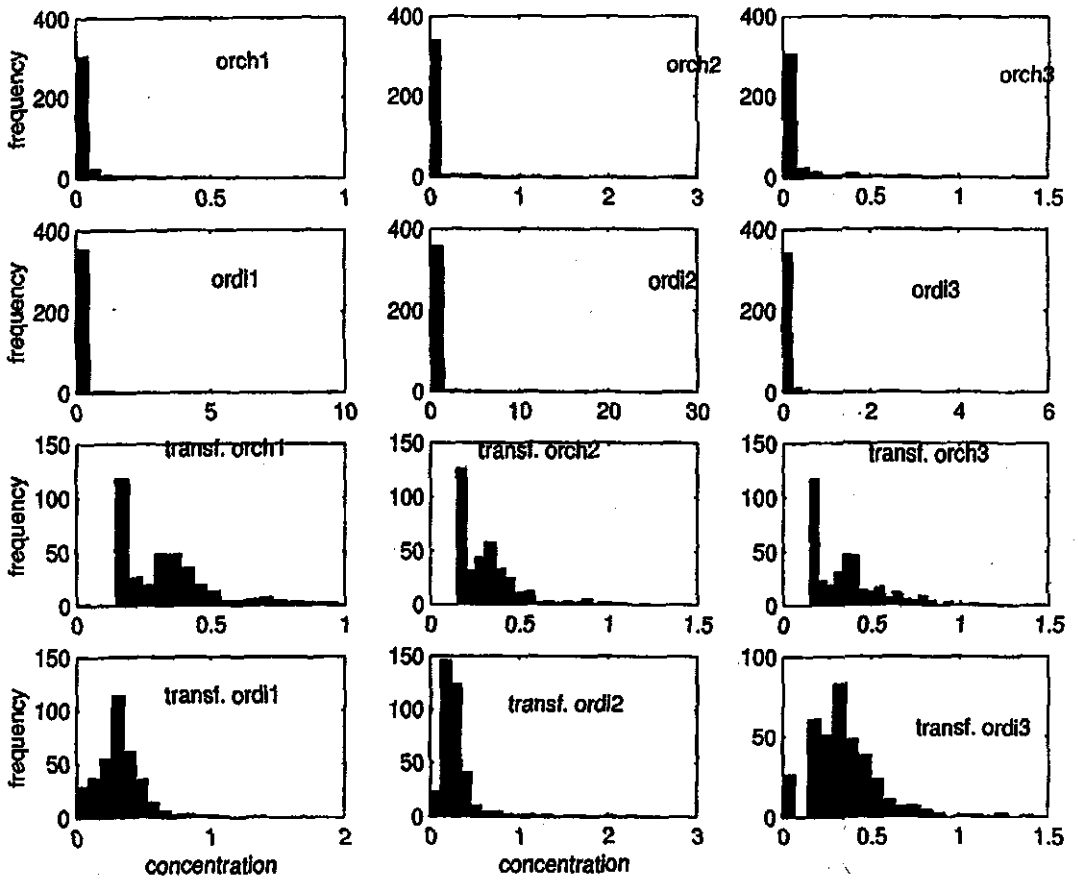

Figure 3: Histograms for original and transformed chlorpyrifos and diazinon concentrations at three locations on Orestimba Creek (May,1996-May,1997, Dow).

When there are large excursions such as are apparent in Figures 1 and 2, transformations may be useful for stabilizing possible relations between series and for improving the conformance to possible normality assumptions. Another complicating factor is the presence of large numbers of non detections or zeros. Common kinds of transformations applied to data

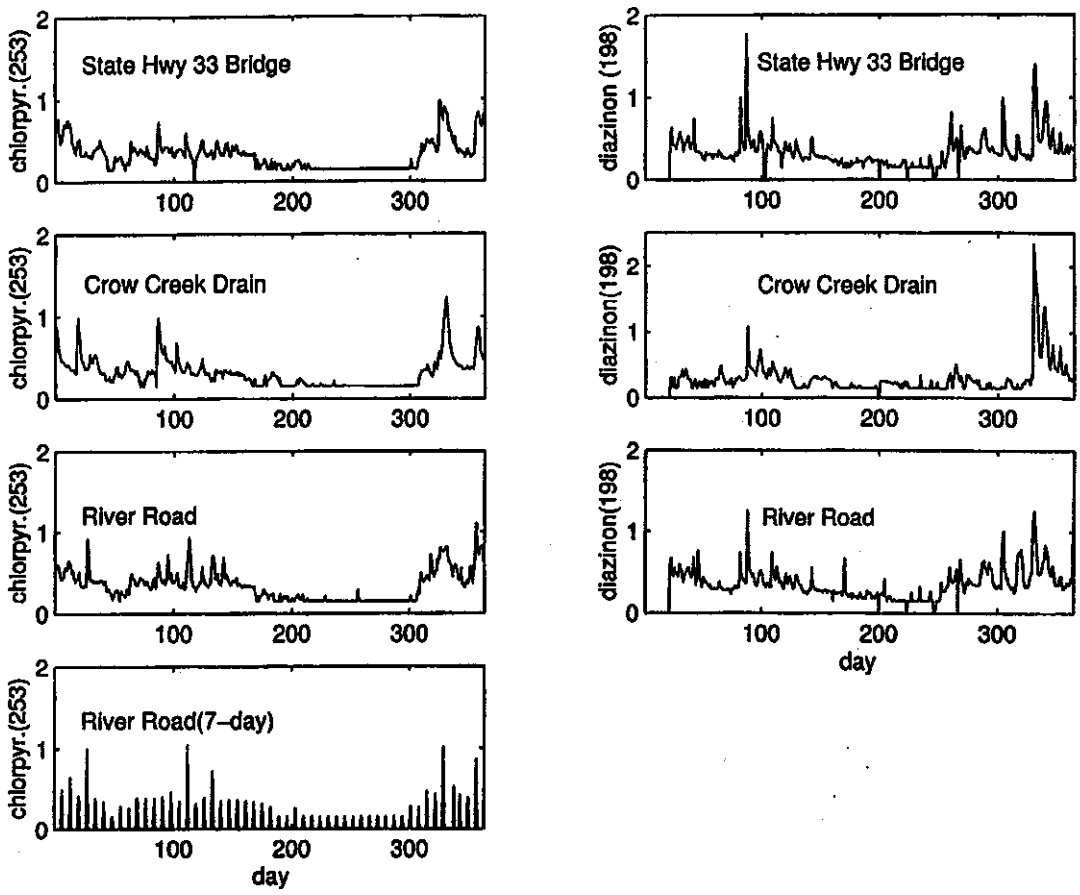

Figure 4: Daily transformed (4th root) chlorpyrifos and diazinon concentrations at three locations on Orestimba Creek (May,1996-May,1997, Dow). A seven-day grab sample is available for chlorpyrifos at the River Road location.

$y_t, t = 1, 2, \ldots, n$ are the logarithmic ($\ln y_t$) and power transformations of the form

$$z_t = \frac{y_t^\lambda - 1}{\lambda}$$

for $0 < \lambda \leq 1$ ($\lambda = 0$ gives the logarithmic). Censoring can be handled in the case where transformations are appropriate using results in Shumway et al (1989). The case where the y_t are correlated is more difficult and it may be appropriate to treat the censored data as being at the detection limit when the detection limit is low, which is the situation for the DPR data. In the present case, for simplicity, we set the values below the detection limit at the average of zero and the detection limit. The high density of non-detections and the low quoted limits combine to make a detailed treatment such as is given in Shumway et al (1988) of limited usefulness. Their treatment, which essentially replaces the values below the

detection limit with conditional expectations under the assumption of independence would not be appropriate in the correlated case considered here anyway.

Frequency histograms for the original and the transformed chlorpyrifos and diazinon concentrations on Orestimba Creek are shown in Figure 3. In this case logarithmic, square root and fourth root transformations were tried and still produced skewed distributions. The fourth-root transformed histograms, shown in the bottom half of the table, seem to improve the situation somewhat if the censored values are ignored, although the distribution is still distinctly non-normal. A time plot of the transformed data is shown in Figure 4 and the high values seem to be toned down enough so that series might be regarded as smooth trend plus a relatively stationary process. Hence, one may regard the transformation as primarily useful for stabilizing the variances and improving the approximation to stationarity.

Searching for a comparable transformation to apply to the USGS data did not lead to substantial improvements in the frequency histograms and it was decided that applying a transformation will not be helpful in this case.

2.2 Correlation and Scatter Diagrams

In order to evaluate the extent to which particular series, say y_{t1}, y_{t2} are linearly related to one another, it is conventional to compute the instantaneous correlation and to examine the scatter diagrams, obtained by plotting y_{t1} on the horizontal scale and y_{t2} on the vertical scale.

The correlation matrix for the transformed (fourth root) chlopyrifos and diazinon measurements at the three Orestimba Creek locations can be computed and we note that this yields the values in Table 1 below, where the order is the chlorpyrifos at the three locations followed by diazinon at the same three locations.

Table 1: Correlations for Chlorpyrifos at 3 Locations (orch1,orch2,orch3) and Diazinon (ordi1,prdi2,ordi3). The locations are 1. Highway 33 Bridge, 2. Crow Creek Drain and 3. River Road.

	orch1	orch2	orch3	ordi1	ordi2	ordi3
orch1	1	.81	.84	.18	.31	.17
orch2	.81	1	.74	.33	.49	.29
orch3	.84	.74	1	.24	.36	.26
ordi1	.18	.33	.24	1	.67	.79
ordi2	.31	.49	.36	.67	1	.62
ordi3	.17	.29	.26	.79	.62	1

Hence, the upper left hand 3×3 table shows the location intercorrelation between chlorpyrifos at the three locations. These seem to be uniformly high, indicating that the spatial

chlorpyrifos correlation is extremely high between the three locations. The lower right hand 3×3 matrix shows the same high spatial correlation for the diazinon values. It should also be noted that statistical significance (.01) can be declared when any cross correlation exceeds $2.33(1/\sqrt{364}) = .12$, so there will be significance for most of the relations. The upper right hand 3 matrix shows correlations between chlorpyrifos and diazinon between stations. The correlations between the two pesticides at the same station are higher, as one might expect.

A feature that complicates the interpretation of the correlations will be the censored and missing data; the latter have been coded as zeros. This effect can be observed in the scatter diagram, shown in Figures 5 and 6, which show the censored values as bands parallel to the horizontal and vertical axes. The behavior of the scatter, excluding the bands of censored (coded as the average of zero and the detection limit) and missing data (coded as zeros), shows that linearity is still not an unreasonable assumption. The missing data can be interpolated using the state-space model proposed later. Censoring in the presence of transformations can be treated for the independent case as in Shumway et al (1989) but the high rate of censoring present and the low detection limits present in these files partially justifies replacing censored values by either zero or some value between zero and the detection limit in the correlated case. Concentrating attention on higher values of chlorpyrifos and diazinon shows that there will be some predictability of one from the other, even at different locations.

At the San Joaquin River location, the correlation matrix given below in Table 2 shows strong associations only between diazinon and simazine. This file is mainly distinguished by the prevalence of non-detections and the correlation values are rather non-informative.

Table 2: Correlations for 4 Herbicides on the San Joaquin River at Vernalis.

	Diazinon	Simazine	Cyanazine	Metolachlor
Diaz.	1	.59	-.05	-.05
Sim.	.59	1	-.07	.07
Cyan.	-.05	-.07	1	.06
Metol.	-.05	.07	.06	1

2.3 Autocorrelation and Cross Correlation Functions

Autocorrelation and cross correlation functions, denoted respectively by ACF and CCF in what follows, extend the notion of association between stations or chemicals to some lag h. That is, the ACF measures the degree to which y_{t1} is correlated with its own past, say $y_{t-h,1}$, where the latter term denotes the pesticide series measured h days in the past. A plot of

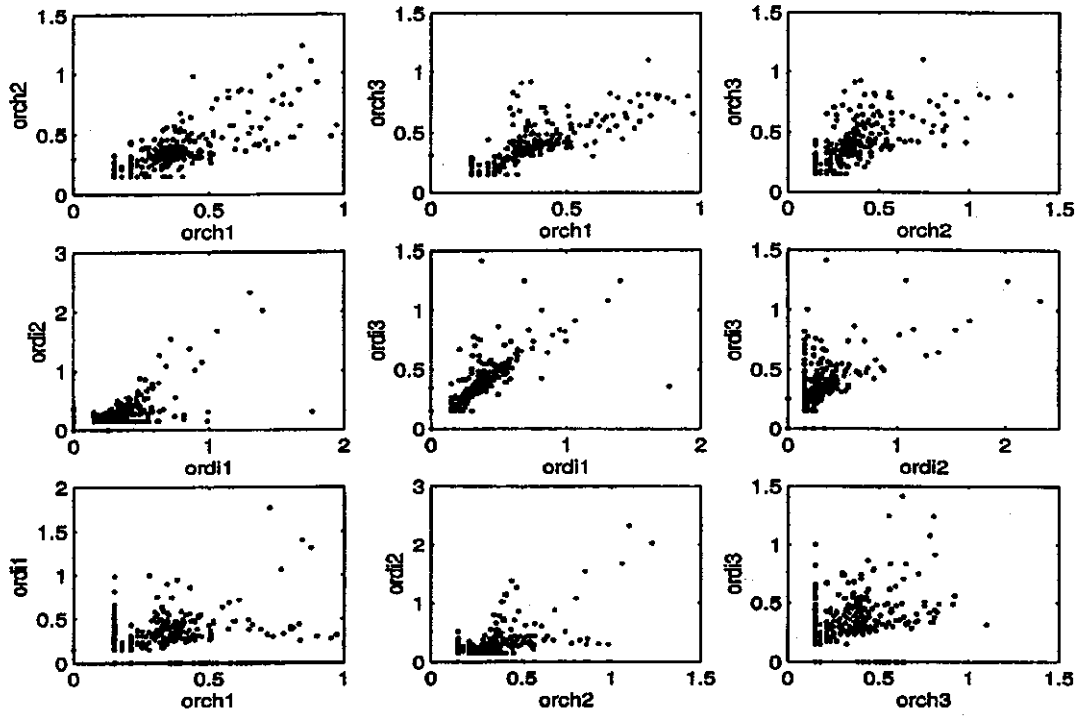

Figure 5: Scatter diagrams relating chlorpyrifos and diazinon levels between locations Orestimba Creek Data. Here, orch1, orch2, orch3, ordi1, ordi2 and ordi3 are as in Table 1.

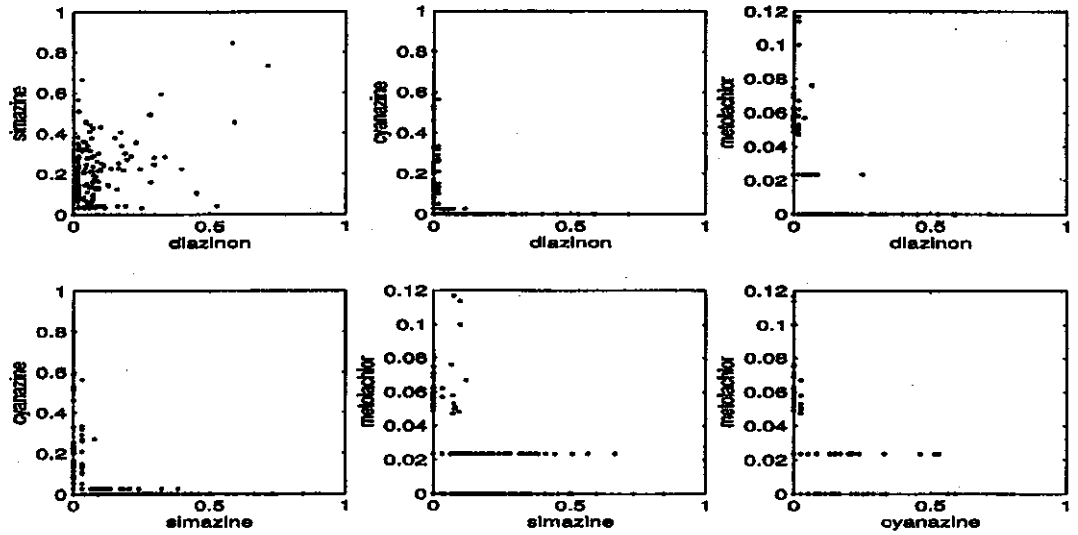

Figure 6: Scatter diagrams relating pesticide levels on the San Joaquin River at Vernalis.

this as a function of lag h displays graphically the correlation with all past values and is a measure of the predictability of the series from its own past.

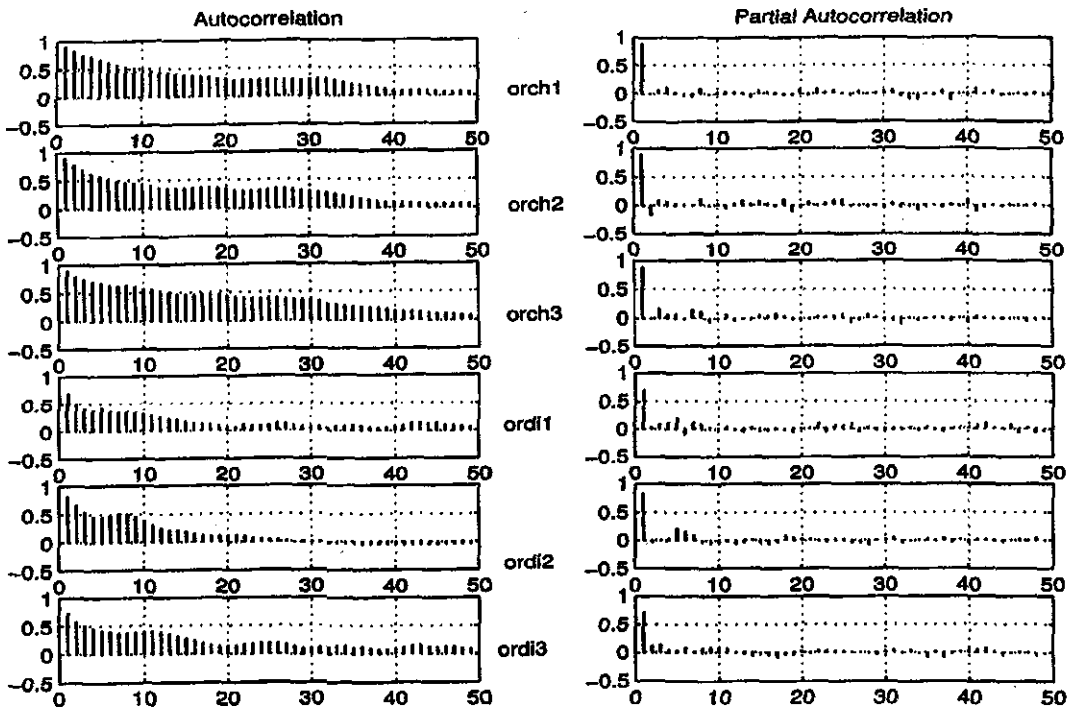

Figure 7: ACF and PACF of daily transformed (4th root) chlorpyrifos and diazinon concentrations at three locations on Orestimba Creek (May,1996-May,1997, Dow). Note the peak at lag one in the PACF, indicating that a first difference or first order autoregressive model might apply.

The CCF works the same for two series, correlating the first series $y_{t,1}$ with both past values, $y_{t-h,2}$, and future values $y_{t+h,2}$ of a second pesticide series. Again, the predictability of the first series from the second (or the second from the first) at lag h is measured. The value of the cross correlation function at lag $h = 0$ generates the measures shown in the correlation matrix, previously given in Table 1 and Table 2. Note that the CCF is not symmetric about zero; one would not expect the same predictability in both directions.

The partial autocorrelation function (PACF) correlates the series y_{t1} with itself at lag h, like the ACF, except that y_{t1} and $y_{t-h,1}$ are first adjusted for their regressions on the intervening values between the time points $t - h$ and t. It is analogous to the conventional partial autocorrelation between two variables, conditioned on a third possible variable to which both variables may be related. It is desirable to condition out the third variable in order to look at the pure correlation between the two of interest. In time series analysis, when the series can be expressed best as an autoregressive process, i.e., a regression on its p past values plus error, the PACF will have values out to lag $h = p$ and will be zero afterwards. This function is helpful in modeling the dynamic behavior of a single series over time.

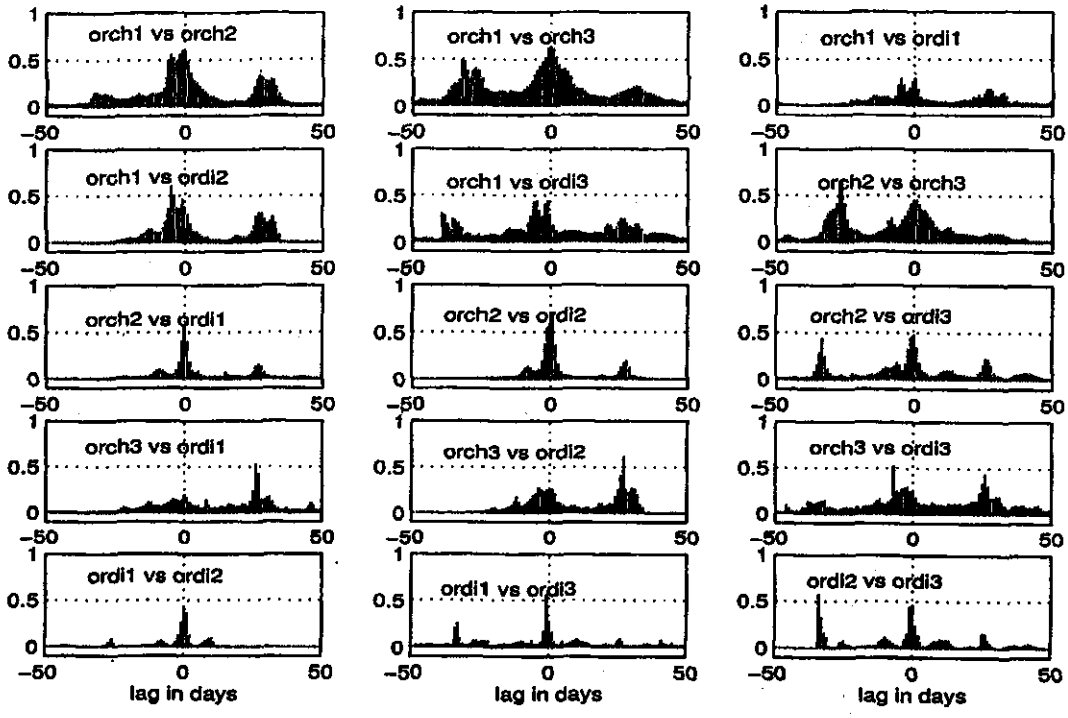

Figure 8: Cross correlation functions relating chlorpyrifos and diazinon levels between locations Orestimba Creek Data.

Figure 7 shows the ACF and PACF for the Orestimba transformed chlorpyrifos and the untransformed diazinon levels. The ACF is dominated by slow decay, indicating either a nonstationary or long memory process. It shows that there is substantial correlation for about 30 days and smaller correlations thereafter. The PACF indicates that some sort of first-order model might suffice; in particular, the values near 1 for the for unit lag PACF's in the chlorpyrifos series suggest that a model of the form

$$y_{ti} = y_{t-1,i} + e_{ti}$$

might be taken for series i, where the e_{ti} are independent and identically distributed errors with zero means and variance σ^2. We use this diagnostic to argue that the random walk model above will be appropriate for modeling the unobserved signal. The cross correlation functions (CCF's) are shown in Figure 8 and we note the large values at lag $h = 0$ mentioned before plus some peaks at $h = \pm 30$ days that may indicate a 30-day period for the episodes.

The ACF's and PACF's for the pesticides on the San Joaquin River, shown in Figure 9, reflect the fact that these are two-day samples so that two-day correlations will be expected, due to filling in zeros for the missing values. Hence, the first-order model, namely and autoregressive model with one lag, may still be adequate. We note the fairly strong correlation

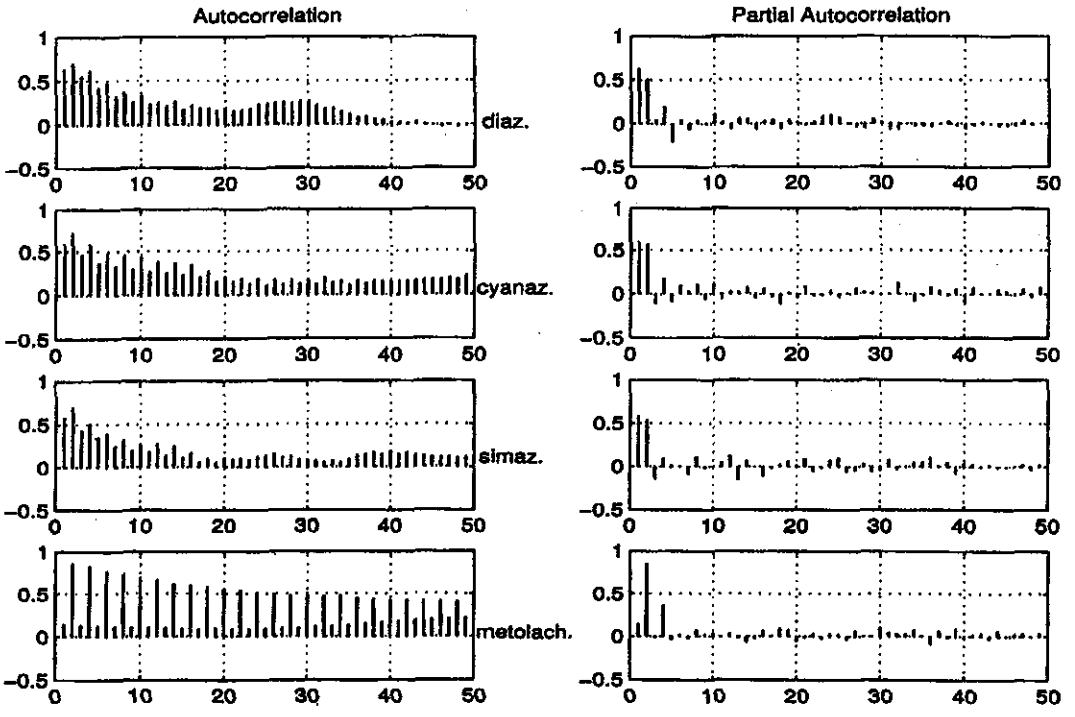

Figure 9: ACF and PACF of herbicide concentrations on the San Joaquin River at Vernalis. The zeros interspersed in the 2-day samples introduce distortion in the measures.

between simazine and diazinon, as before, with additional evidence of a 30-day peak.

3. Dynamic Regression

The exploratory data analysis of Section 2 suggests that there can be substantial correlations in both time and space that will be important for modeling the pesticide process. The detailed spatial structure will be difficult because the typical data will be measured at a limited number of stations, making any modeling that uses a general correlation model over space a difficult matter. A general class of models that seems to fit this situation is the dynamic regression situation that is represented by a class of state-space models that represent the observed data vector $\boldsymbol{y}_t = (y_{t1}, \ldots, y_{tq})'$ as a linear combination of an unobserved, i.e., unknown, signal vector $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$ and an error vector $\boldsymbol{v}_t$. The resulting model for the observed series is

$$\boldsymbol{y}_t = A\boldsymbol{x}_t + \boldsymbol{v}_t, \tag{1}$$

where A is a $q \times p$ matrix that converts the unobserved signal $\boldsymbol{x}_t$ into the observed data $\boldsymbol{y}_t$ and $\boldsymbol{v}_t$ are independent multivariate normal vectors with zero means and common, i.e.

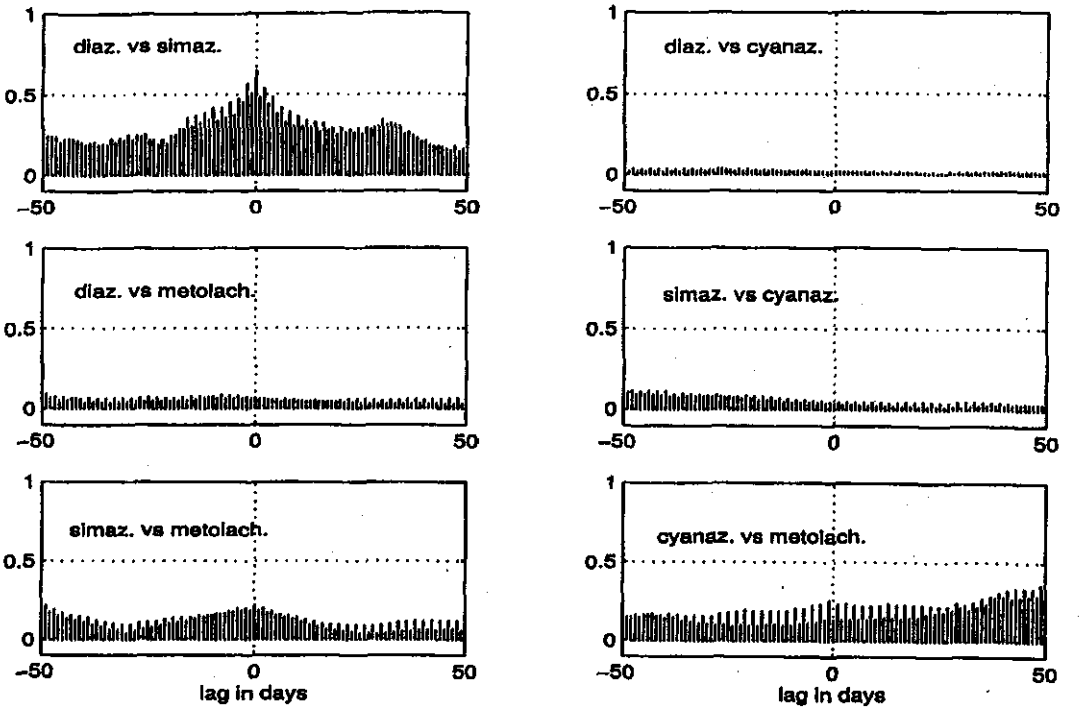

Figure 10: Cross correlation functions relating pesticide levels on the San Joaquin River at Vernalis.

identical, $q \times q$ covariance matrices R. The relation (1) is sometimes called the observation equation. The unobserved signal $\boldsymbol{x}_t$ is assumed to evolve through time and space according to the standard regression relation

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \boldsymbol{w}_t \tag{2}$$

where Φ is a $p \times p$ transition matrix that summarizes the space-time regresssion relation for the unobserved signal and $\boldsymbol{w}_t$ are independent normal noise vectors with a common $p \times p$ covariance matrix Q. The relation (2) is called the state equation. For further details relating to the state-space model and its other applications, see Shumway and Stoffer (2000, Chapter 4).

In the current context, the state-space model provides a convenient way of modeling the spatial and time connections for pesticide concentrations such as those found in the two study areas of this report. For example, the transformed chlopyrifos concentrations at three locations, shown in Figure 4, might be modeled as

$$y_{ti} = x_{ti} + v_{ti},$$

for the $i = 1, 2, 3$ locations at time t by taking the observation matrix A in (1) as the 3×3

identity matrix. The state equation (2) might be of the form

$$x_{ti} = \phi_{i1}x_{t-1,1} + \phi_{i2}x_{t-1,2} + \phi_{i3}x_{t-1,3} + w_{ti}$$

for $i = 1, 2, 3$, giving a set of transitions in space and time governing the evolution of the pesticide concentrations through time and space at the three locations.

There will be two problems of interest in treating the state-space model. The first is parameter estimation, including the transition matrix, Φ, the observation covariance matrix, R, and the model covariance matrix, Q. A second problem will be estimating the unobserved process $\boldsymbol{x}_t$ and its uncertainty, given values for the input parameters. This second problem is of less interest for the model given above but is of critical interest for the signal model given in the next section. The parameters are estimated by maximum likelihood whereas the unobserved process is estimated via the Kalman filters and smoothers (see Shumway and Stoffer, 2000, Chapter 4). Estimation of the parameters by maximum likelihood is covered in Sections 4.3 and 4.4, pp. 321-333. Software and documentation are available in McQuarrie and Shumway (1994) or in the MATLAB programs developed for this study.

In order to illustrate the results, consider the chlorpyrifos levels from the three locations on Orestimba Creek. Applying the computational procedure, we obtain

$$\hat{\Phi} = \begin{pmatrix} \underline{.82} & .06 & \underline{.11} \\ .05 & \underline{.83} & \underline{.10} \\ \underline{.34} & .03 & \underline{.65} \end{pmatrix}$$

as the estimated transition matrix; the standard errors on on the order of .05. Substituting the significant values into the station by station model yields

$$x_{t1} = .83x_{t-1,1} + .11x_{t-1,3} + w_{t1}$$

$$x_{t2} = .83x_{t-1,2} + .10x_{t-1,3} + w_{t2}$$

$$x_{t3} = .34x_{t-1,1} + .65x_{t-1,3} + w_{t3}$$

One interpretation of the above equations is that the current location value is most dependent on its value for the immediately preceding day for the Highway 33 and Crow Creek Drain series. These series also seem to depend weakly (coefficients are .10) on the River Road series. The River Road series depends on its own past (coefficient .65) and on the past of the Highway 33 Bridge series (coefficient .34). The other parameters in the system are the elements of the two covariance matrices R and Q. The standard deviations in the measurement matrix R are on the order of .02 whereas those in Q, representing model error on about .07. These two values enable a rough comparison to be made between model error (standard deviation .07) and observation error (standard deviation .02). Correlations between model errors, i.e. correlations relating the model errors $\boldsymbol{w}_t$ were about .5.

Similar patterns predominate for the three locations measuring diazinon, where the transition matrix is of the form

$$\hat{\Phi} = \begin{pmatrix} \underline{.51} & .10 & \underline{.35} \\ \underline{.28} & \underline{.78} & -.11 \\ \underline{.54} & -.09 & \underline{.54} \end{pmatrix}.$$

A slight difference is the additional communication that appears between the first and second locations, leading to the approximate model

$$x_{t2} = .28x_{t-1,1} + .78x_{t-1,2} - .11x_{t-1,3} + w_{t2}$$

The standard deviations in the measurement matrix are about .03 whereas those in Q are about .17, implying that the model error is larger relative to the measurement error in this case.

For the St. Joaquin River data, the transitions only involve past values of the same pesticides, measured for chlopyrifos and diazinon so that the transition matrix is estimated by

$$\hat{\Phi} = \begin{pmatrix} .87 & .02 \\ .00 & .97 \end{pmatrix}.$$

The simplified model becomes

$$x_{t1} = .87x_{t-1,1} + wt1$$

$$x_{t2} = .97x_{t-1,2} + w_{t2}$$

and we see that the pesticides are basically disconnected. The form of the relation suggests that a random walk model might work separately for each pesticide. In the next section, we use this form for estimating a common signal rather than the second order AR model that might have been indicated by the PACF in Figure 9. It should be noted that for the two-day samples, intervening values will be automatically interpolated by the Kalman filters and smoothers.

4. Signal Extraction

The most interesting applications of the preceding will be to the problem of assessing the levels of pesticide concentration in a given area. For this to happen, we need a model that expresses observed series such as those given in Figures 1 and 2 in terms of a common signal. The data seem to be characterized by the occurrence of relatively short-lived episodic fluctuations that may exceed some prescribed regulatory standards. Problems that occur will created by irregular sampling and long sequences of observed values that are below detection limits. Procedures for merging irregularly observed and episodic fluctuations into a common

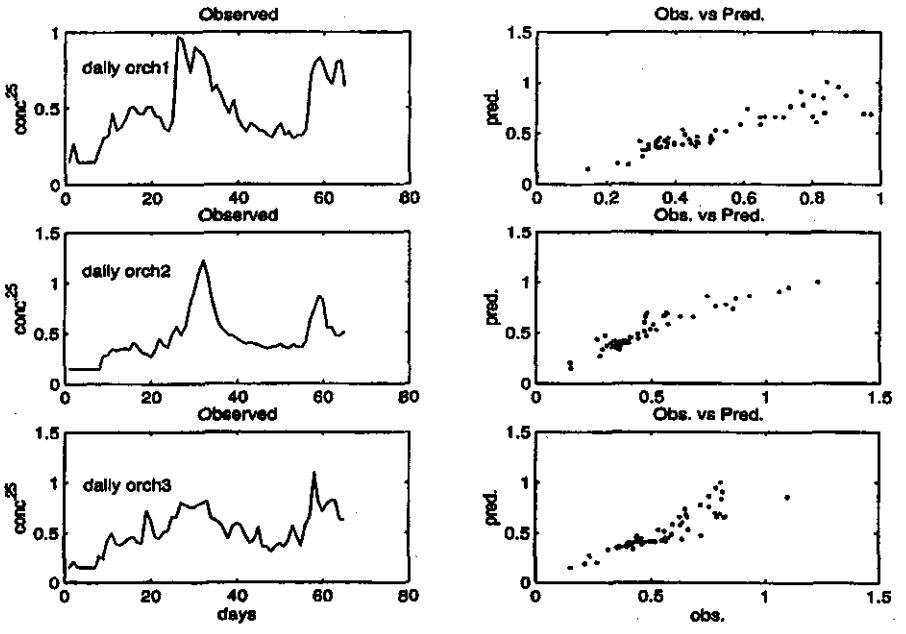

Figure 11: Observed transformed chlorpyrifos concentrations at three locations on the Orestimba Creek for the last 64 days in Figures 1 and 4. Observed series on left and scatter diagram on the right relating observed to predicted for common signal model.

signal and probability limits for that signal are of interest for determining if standards have been exceeded.

The state-space model defines the signal x_t in terms of the equations given by (1) and (2). The signal is estimated by the Kalman smoothed values

$$x_t^n = E\{x_t|y_1, \ldots, y_n\}, \tag{3}$$

i.e., the conditional expectation, given the observed data. The uncertainty of the smoothed values is expressed as the mean square covariance, say

$$P_t^n = E\{(x_t^n - x_t)(x_t^n - x_t)'|y_1, \ldots, y_n\}, \tag{4}$$

which is also the unconditional covariance. Under the assumption that the errors v_t and w_t are normally distributed, prediction intervals, at any given probability level, are available from (3) and (4). The filtering and smoothing equations for estimating the unobserved process are given as Properties 4.1-4.3 in Section 4.2, pp. 312-317.

For pesticide data considered here, it seems sensible to hypothesize a simple model for the pesticide signal x_t of the form

$$x_t = x_{t-1} + w_t, \tag{5}$$

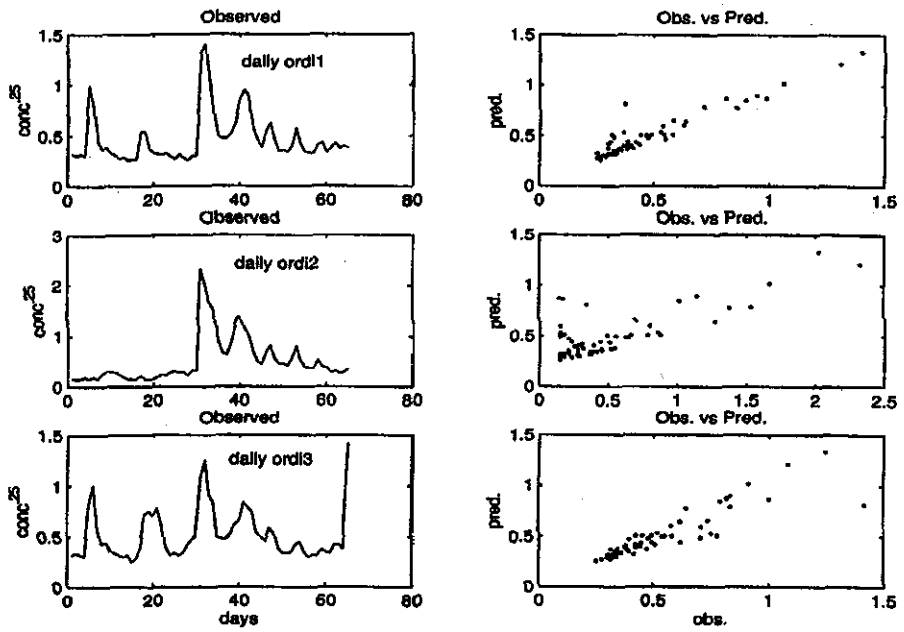

Figure 12: Observed transformed diazinon concentrations at three locations on the Orestimba Creek. Observed series on left and scatter diagram on the right relating observed to predicted for common signal model.

which appears to be common to a given subset of observation series. This is the state equation corresponding to (2). The observed pesticide levels at a number of series, say $i = 1, \ldots, q$ containing the signal could be assumed to be of the form

$$y_{ti} = x_t + v_{ti}, \tag{6}$$

which corresponds to the observation equation (1). The observation covariance matrix is R and we take the model covariance σ^2 as given. This special form of the model means that a common signal x_t is observed on each of the series and that the common signal satisfies the random walk assumption (5).

The above model is in state-space form with A, the $q \times 1$ vector of ones, where q is the number of series that contain the common signal. Note that R will be the $q \times q$ matrix of measurement error variances and covariances and $\Phi = 1$ will be the simple scalar one that generates (5). The signal variance will be just q_{11}, since $p = 1$ in the general state space model (2). The first step in this procedure is estimating the unknown parameters R and q_{11} and we accomplish this by maximum likelihood, using the EM algorithm (see Shumway and Stoffer, 2000). The final process is to produce the estimator (3) and its variance (4) for the signal, say x_t^n and P_t^n, evaluated at the maximum likelihood estimators $\hat{R}$ and $\hat{q}_{11}$. For this particular case, the normality assumptions yields approximate prediction limits of the

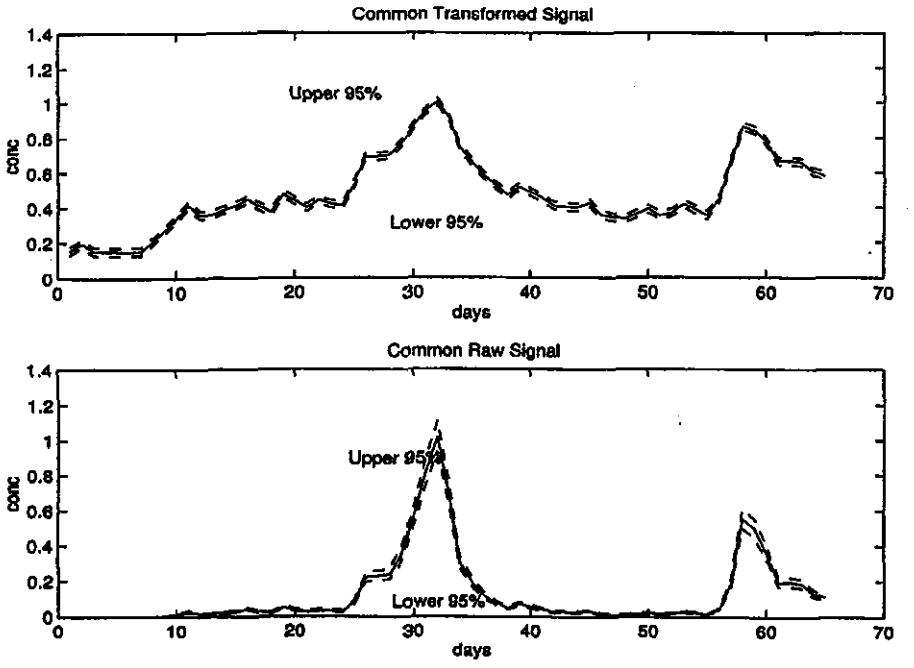

Figure 13: Estimated chlorpyrifos signal and 95 percent prediction limits for episode at end of data using all three locations.

form $x_t^n \pm zP_t^n$ where the multiplier z is taken from the normal distribution to obtain the appropriate probability level. We will discuss this particular point further in Section 5.

For the Orestimba Creek data, it is natural to assume that all three locations measure about the same signal and to investigate a model that assumes a common signal. Because of the high concentration of non-detections in the data, we took a typical episode occurring in the March to April, 1997 time interval as the base data and developed an estimator for the common signal and its variance under the special state-space formulation given by (6) and (5). This was repeated for both the chlorpyrifos (ch1-ch3) and diazinon (di1-di3) data for the last 64 days of data. Table 3 gives the estimated parameters for the two cases and we note the small variances associated with both the measurement and model errors.

Note that the diazinon observation variance at the second location is higher, confirming the visual difference in Figure 4 between the Crow Creek Drain and the other two locations. Figures 11 and 12 show the observed transformed (fourth root) chlorpyrifos and diazinon concentrations over the last 64 days and gives an indication of the fit of the single signal model. The right panels plot the observed vs the predicted values for the single signal model at all three locations and we see that there is fairly good agreement between the two. Figures 13 and 14 show the estimated common profile chlorpyrifos and diazinon signals

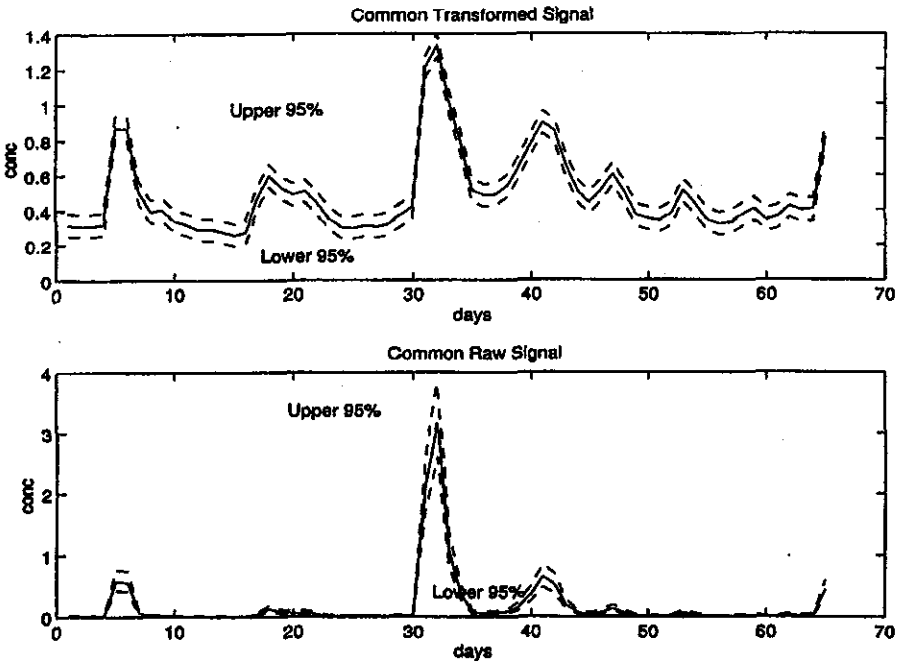

Figure 14: Estimated diazinon common signal and 95 percent prediction limits using all three locations.

for the three locations, along with 95% prediction limits. We see that the limits are quite narrow, indicating good predictability and high confidence in the values for separate points. Not that the major peak in the two pesticide concentrations occurs at about the same time, i.e. 32 days into the record. There are two smaller peaks at the end of the chlorpyrifos and beginning of the diazinon signals that do not agree.

A second question of interest is the relation between the more intensive daily averages and the seven day grab samples, which were both available only at the River Road location for chlopryrifos levels. Figure 15 shows both sets of measured concentrations and it is clear that the general trend of the two series agree although there were only 10 weekly samples available over the full 64 day period. The scatter diagram relating the profile signal estimates to the

Table 3: Estimation of covariance parameters in matrices R and Q for the common signal model for the Orestimba Creek episode where ch4 denotes the weekly grab samples.

Series	r_{11}	r_{12}	r_{13}	r_{22}	r_{23}	r_{33}	q_{11}
ch1-ch3	.0068	-.0040	.0012	.0062	-.0058	.0080	.0051
ch3-ch4	.0031	-.0012		.0052			.0075
di1-di3	.0074	.0110	-.0072	.0993	-.0166	.0125	.0292

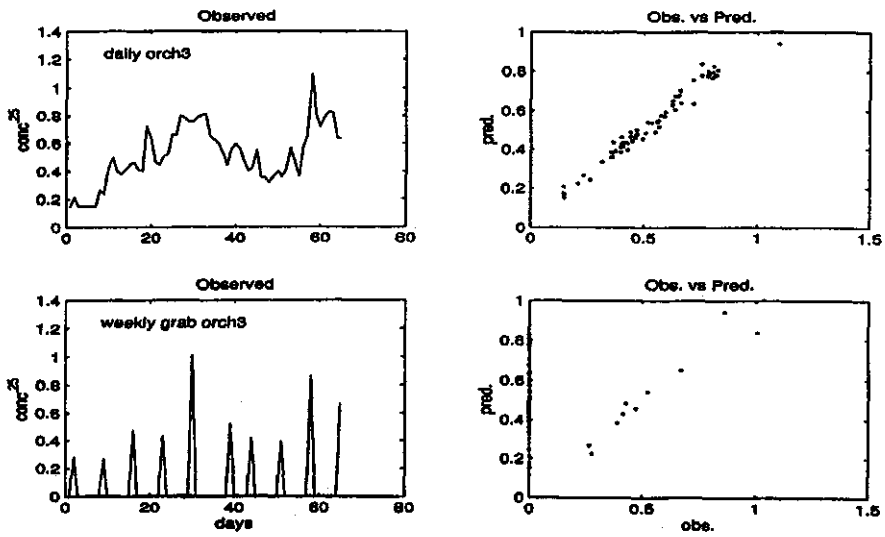

Figure 15: Observed transformed chlorpyrifos concentrations at River Road location compared to weekly grab samples at the same location. Observed series are on the left, showing zeros for missing values in the grab sample series. The scatter diagrams on the right relate the observed data on the horizontal scale to the predicted data on the vertical scale using the common signal model (9).

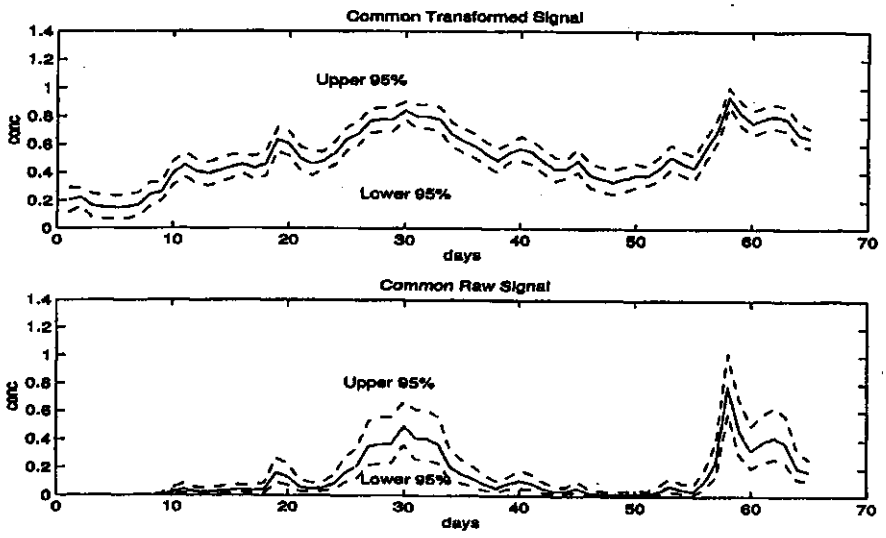

Figure 16: Estimated chlorpyrifos signal and 95 percent prediction limits using only River Road location with benchmark data.

data is quite good for both the daily and the weekly values. Figure 16 shows the common signal, again obtained by Kalman smoothing, and we note that the prediction intervals are

Table 4: Estimation of covariance parameters in matrices R and Q for the common signal model for the San Joaquin River episode.

Series	r_{11}	r_{12}	r_{22}	q_{11}
Diazinon-Simazine	.0072	.0001	.0134	.0059

slightly larger than would be obtained using all three locations, as in the upper panel of Figure 13. The shape is very similar, although the second peak now shows larger that the first peak, a result of the difference that seem to be unique to the third location (see figure 11). The estimated parameters are shown in Table 3 and we note that the measurement errors are smaller, due to the common location, but the model error is larger, which will be due to one less data series for estimation.

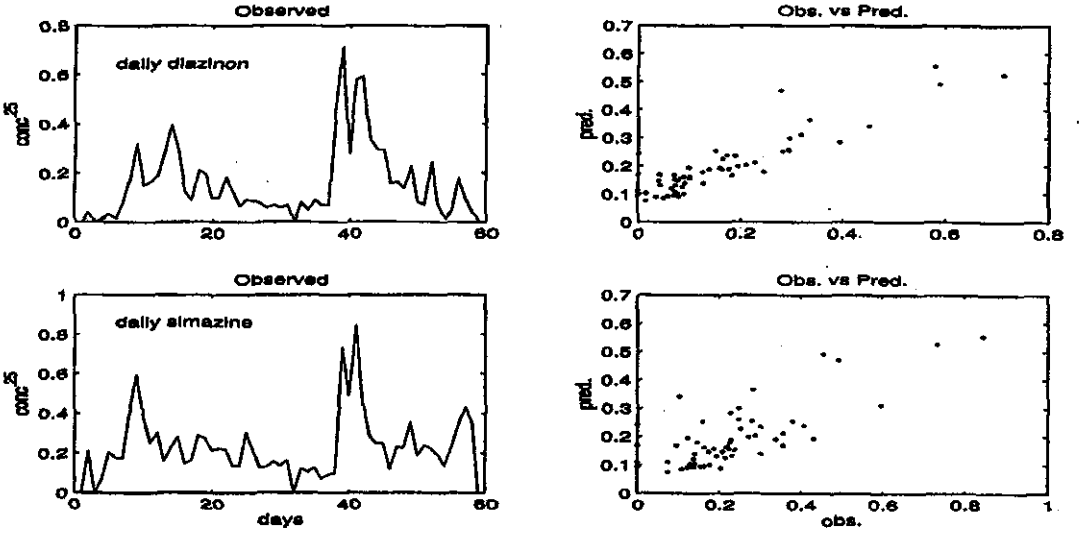

Figure 17: Observed diazinon and simazine concentrations on the San Joaquin River at Vernalis for days 721-790 in Figure 2. Observed series on left and scatter diagram on the right relating observed to predicted for common signal model.

For the USGS data from the San Joaquin River at Vernalis, there were only significant detections of the four herbicides shown in Figure 2 and the best possible episode of interest occurred in the diazinon and simazine series between days 721 and 790. A plot of this signal is shown in Figure 17 and we note that the profiles of the two herbicides are quite similar over this range. The episode would have occurred in in about January of 1993. The covariance parameters are given in Table 4 and we see that the measurement and model errors are roughly compatible with those from Orestimba Creek in Table 3.

The common signal model shows scatter in figure 17 that is consistent with good pre-

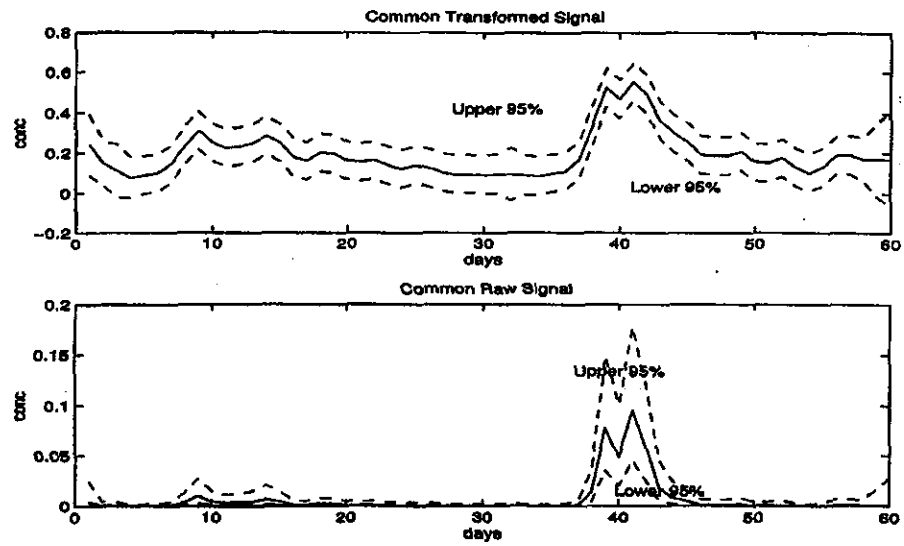

Figure 18: Estimated diazinon-simazine common episode signal and 95 percent prediction limits using using both measurements.

dictability. The estimated common signal and its' 95% prediction limits are shown in Figure 18 and here we note that the intervals are slightly wider than those done earlier.

5. Implications for Regulation/Mitigation and Future Study

The question remains as to how best to utilize the large files that are available. The fact that most sampling consists of daily samples with substantial numbers of non-detections suggests concentrating on the high concentration episodes and utilizing the prediction profiles developed in the preceding section. The strong correlation over time observed in most episodes suggests that signals can be estimated with reasonable accuracy from single series. The spatial correlation over locations along the same tributary in the Orestimba Creek study allowed more accurate profiles to be developed, assuming common signals at all locations. This also suggests a general regional contamination rather than one that is confined to a single location. The study has also shown how to combine irregularly observed samples into a daily profile signal, using the grab and daily coomposite samples in the Orestimba Creek study and using a mixture of 1-day and 2-day samples in the San Joaquin River study.

For deciding whether a given episode is in violation of a given regulatory standard, we suggest examining a profile signal and its prediction limits at a very high level of confidence, as measured by the prediction limits. As an example, consider Figure 19, which shows the estimated chlorpyrifos signal with 99.8% prediction limits. Interpreting this in a slightly different fashion we note that the lower dotted line is a lower bound on the concentration

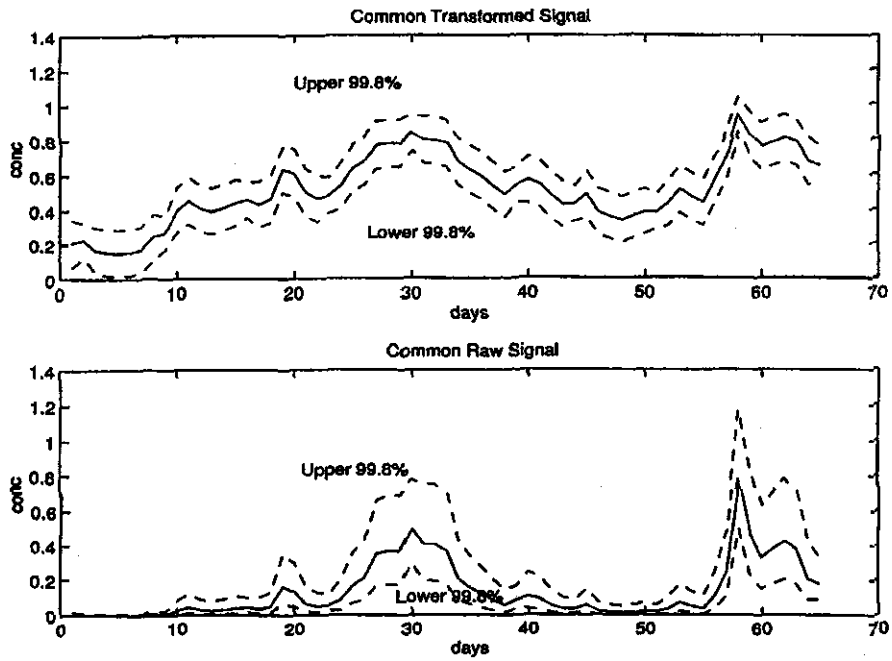

Figure 19: Estimated chlorpyrifos signal and 99.8 percent prediction limits for episode at end of data using all three locations.

level and there is only about one chance in a thousand that the true concentration is below that limit on any given day.

Formally, let the allowable concentration be denoted by L and suppose that we define a violation as exceeding this threshold for m days in a row. Since the difference between the estimated signal and its prediction day t, say, $x_t^n - x_t$, will be normally distributed with mean 0 and variance P_t^n for any given sample, the probability of being in compliance on that day is approximately

$$P\{x_t \leq L\} = 1 - \Phi\Big(\frac{x_t^n - L}{\sqrt{P_t^n}}\Big), \tag{7}$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution. The probability of being in violation for for m days consecutively can be bounded below by

$$P\Big\{\cap_{t=1}^m (x_t > L)\Big\} \geq 1 - \sum_{t=1}^m \Big(1 - \Phi\Big(\frac{x_t^n - L}{\sqrt{P_t^n}}\Big)\Big), \tag{8}$$

applying Bonferonni's inequality. Suppose that the probability in (7) is .001 and that the standard was exceeded for 10 days. Then, using (8), the probability that this would happen by chance is bounded below by $1 - 10(.001) = .99$, i.e., there is at least a 99% chance that the standard was exceed on all 10 days.

There are a number of other factors such as rainfall, pesticide use, evapotranspiration and streamflow data which could provide additional predictive power for the underlying observed concentrations. Such covariates may be added to the state vector and we might consider the additional problem of estimating the scaling matrix A in the observation equation (1). The influence of fixed effects due to differences in stations or regions can also be estimated by adding a term that incorporates covariates in a vector, say z_t, to the observation equation so that

$$y_t = \Gamma z_t + A x_t + v_t \tag{9}$$

becomes the new model and we may need to estimate both Γ and A. Finally, the steps described in this report, namely, exploratory data analysis, dynamic regression and signal extraction can be applied to additional data sets to verify whether regulatory standards are being consistently violated in a given area during a specified period of time.

6. Acknowledgements

I am pleased to acknowledge the assistance of Dr. Frank Spurlock of DPR whose careful and insightful reading of the draft report lead to a number of corrections and revisions. Remaining errors or lack of clarity are solely my responsibility.

7. References

MacCoy, D., K.L. Crepeau and K.M. Kuivila (1995). Dissolved pesticide data for the San Joaquin River at Vernalis and the Sacramento River at Sacramento, California, 1991-1994. USGS Report 95-110, California Dept. of Pesticide Regulation Surface Water Database.

McQuarrie, A.D.R. and R.H. Shumway (1994). *ASTSA for Windows.* Available for download as freeware on **www.stat.ucdavis.edu/ shumway/tsa.html.**

Poletika, N.N. and C.K. Robb(1998). A monitoring study to characterize chlorpyrifos concentration patterns and ecological risk in an agriculturally dominated tributary of the San Joaquin River. Dow AgroSciences LLC Study ENV96055, California Dept. of Pesticide Regulation Surface Water Database.

Shumway, R.H., Azari, A.S. and Johnson, P. (1989). Estimating mean concentrations under transformation for environmental data with detection limits. *Technometrics*, 31, 347-356.

Shumway, R.H. (2000). Dynamic mixed models for irregularly observed time series. *Resanhas*,4, 433-456.

Shumway, R.H. and D.S. Stoffer (2000). *Time Series Analysis and Its Applications.* New York: Springer Verlag.