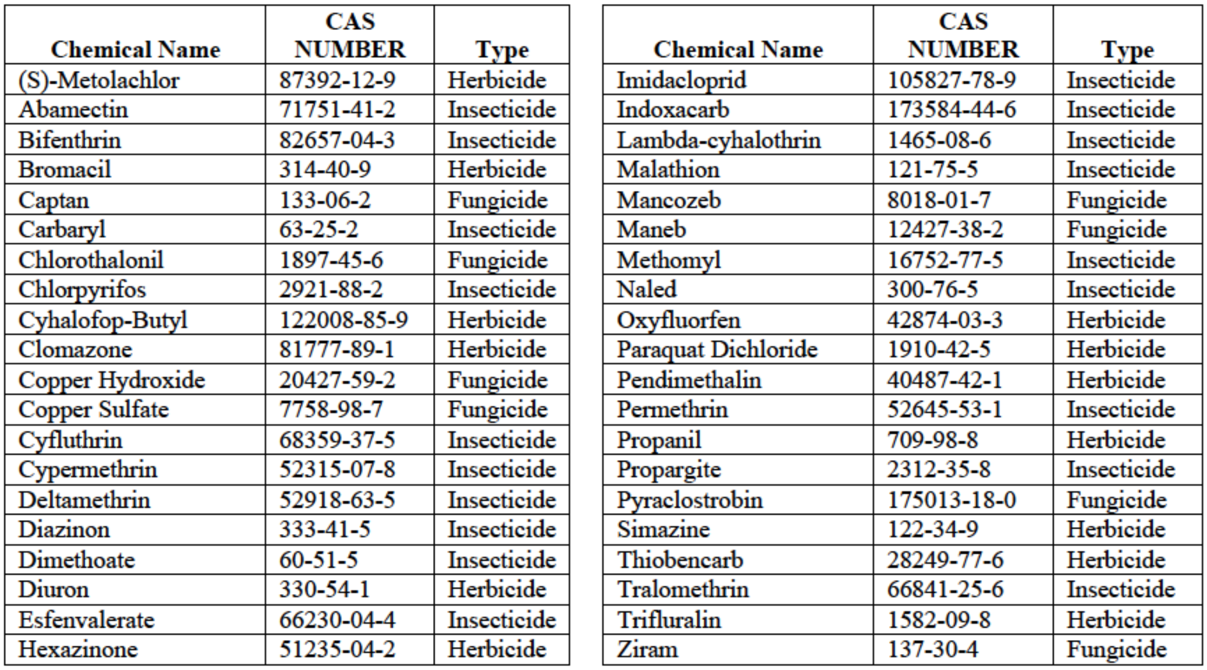
Predicting Missing Observations in Pesticide Data

Daniel Iong

Introduction

From 2007 to 2011, researchers at the University of California, Davis and Waterbourne Environmental, Inc conducted a study to determine the relative ranking of potential areas of concern with respect to pesticide exposure to sensitive and endangered aquatic species in the Sacramento River watershed, San Joaquin River, and Bay-Delta estuary in California. The study involved the use of simulation modeling, historical water quality monitoring, and Geographic Information System (GIS) analysis in a weight-of-evidence context. A list of 40 pesticides published by the Central Valley Regional Water Quality Control Board in 2009 estimated to pose the highest risk to aquatic life (presented below) was used in their analysis.



Chemical Name	CAS NUMBER	Type
(S)-Metolachlor	87392-12-9	Herbicide
Abamectin	71751-41-2	Insecticide
Bifenthrin	82657-04-3	Insecticide
Bromacil	314-40-9	Herbicide
Captan	133-06-2	Fungicide
Carbaryl	63-25-2	Insecticide
Chlorothalonil	1897-45-6	Fungicide
Chlorpyrifos	2921-88-2	Insecticide
Cyhalofop-Butyl	122008-85-9	Herbicide
Clomazone	81777-89-1	Herbicide
Copper Hydroxide	20427-59-2	Fungicide
Copper Sulfate	7758-98-7	Fungicide
Cyfluthrin	68359-37-5	Insecticide
Cypermethrin	52315-07-8	Insecticide
Deltamethrin	52918-63-5	Insecticide
Diazinon	333-41-5	Insecticide
Dimethoate	60-51-5	Insecticide
Diuron	330-54-1	Herbicide
Esfenvalerate	66230-04-4	Insecticide
Hexazinone	51235-04-2	Herbicide

Chemical Name	CAS NUMBER	Type
Imidacloprid	105827-78-9	Insecticide
Indoxacarb	173584-44-6	Insecticide
Lambda-cyhalothrin	1465-08-6	Insecticide
Malathion	121-75-5	Insecticide
Mancozeb	8018-01-7	Fungicide
Maneb	12427-38-2	Fungicide
Methomyl	16752-77-5	Insecticide
Naled	300-76-5	Insecticide
Oxyfluorfen	42874-03-3	Herbicide
Paraquat Dichloride	1910-42-5	Herbicide
Pendimethalin	40487-42-1	Herbicide
Permethrin	52645-53-1	Insecticide
Propanil	709-98-8	Herbicide
Propargite	2312-35-8	Insecticide
Pyraclostrobin	175013-18-0	Fungicide
Simazine	122-34-9	Herbicide
Thiobencarb	28249-77-6	Herbicide
Tralomethrin	66841-25-6	Insecticide
Trifluralin	1582-09-8	Herbicide
Ziram	137-30-4	Fungicide

Figure 1: Pesticide selected for analysis by the Central Valley Regional Water Quality Control Board

Daily pesticide loads were estimated using 10-years of historical pesticide use data obtained from the California Department of Pesticide Regulation's (DPR) Pesticide Use Reporting (PUR) database. Loadings from agricultural uses were predicted using the Pesticide Root Zone Model (PRZM) and the Rice Water Quality (RICEWQ) model but for our study, we only consider the outputs of the PRZM model. More details on the PRZM model are provided in the next section.

PRZM Model

The Pesticide Root Zone (PRZM) model was originally developed by the US Environmental Protection Agency (EPA) to simulate the transport and transformation of agriculturally applied pesticides in the crop root zone. The most recent version of the PRZM model is a one-dimensional, dynamic, compartmental model used to simulate the movement of chemicals, not limited to pesticides, in unsaturated soil systems within and immediately below the plant root zone. It allows users to perform simulations of potentially toxic chemicals that are applied to soil or to plant foliage. (Suarez 2005)

In a study by researchers at the University of California, Davis and Waterborne Environmental, Inc that lasted from 2007 to 2011, over 8.7 million individual PRZM simulations were conducted with each simulation representing a unique combination of weather conditions, soil, crop, irrigation type, and pesticide application history to produce "edge-of-field" estimates of pesticide runoff. The soil parameters were identified from the soil survey geographic database. 169 different crops are present in the PUR database in which the chemicals of interest were applied on. After accounting for similarities in crop canopy, rooting depth, and other factors that affect crop growth, only 29 different crops were included in the model. Although 10 years (2000-2009) of historical weather data was available, only the data for the year of application and the following year were included in the simulations to reduce runtime of the PRZM model. Chemical environmental fate properties used in the PRZM simulations include: solubility, organic carbon-water partition coefficient (Koc), and aerobic soil metabolism. Pesticide use data were obtained from the PUR database, which contains information about chemical application dates, amounts, and types at a 1 square mile (one PLSS section) resolution. Results from a detailed survey conducted by the California Department of Water Resources to characterize irrigation methods at the county-level were used in the simulation. The different irrigation methods include: flood, furrow, drip, sprinkler, etc.

For our study, we only examined the daily pesticide loadings predicted by the PRZM model for the herbicide Trifluralin in the year 2008. We chose this combination of pesticide and year because it had an adequate amount of observations within the year and it is particularly toxic.

Overview of Papers Considered

We considered methods for handling missing observations in time series by implementing the methods discussed in two papers: Shumway et. al. (1982) and Robinson (1980). Shumway et. al. (1982) proposes an approach to smoothing and forecasting for time series with missing observations, which involves using an EM algorithm in conjunction with Kalman smoothed estimators to estimate parameters in a state-space model by maximum likelihood. In a general state-space model, the $p \times 1$ vector series of interest x_t is observed as a component in the random regression model

$$y_t = M_t x_t + v_t, t = 1, 2, \ldots, n$$

where M_t is a known $q \times p$ design matrix and the noise terms $v_t, t = 1, \ldots, n$, are uncorrelated and distributed as N(0, R). Here, R is a $q \times q$ covariance matrix. We only observe $y_t, t = 1, \ldots, n$ but we are mainly interested in the random series x_t, which is modelled as a first-order multivariate process of the form

$$x_t = \Phi x_{t-1} + w_t, t = 1, \ldots, n$$

where Φ is a $p \times p$ transition matrix which describes the way x_t changes over successive time periods. The initial value, x_0, is assumed to to be a Gaussian distributed random vector with mean vector μ and $p \times p$ covariance matrix Σ. The noise terms, w_t, are centered uncorrelated normal vectors with covariance matrix Q. Shumway et. al. considers smoothing and forecasting time series with missing observations as a problem of estimating the random process in the state space context described above. To obtain estimates for μ, Σ, Φ, Q, and R, Shumway et. al. minimizes the log likelihood below by an expectation-maximization (EM) algorithm.

$$
\begin{aligned}
logL = \frac{1}{2}log|\Sigma| - \frac{1}{2}(x_o - \mu)'\Sigma^{-1}(x_o - \mu) \\
-\frac{n}{2}log|Q| - \frac{1}{2}\sum_{t=1}^{n}(x_t - \Phi x_{t-1})'Q^{-1}(x_t - \Phi x_{t-1}) \\
-\frac{n}{2}log|R| - \frac{1}{2}\sum_{t=1}^{n}(y_t - M_t x_t)'R^{-1}(y_t - M_t x_t)
\end{aligned}
$$

Robinson (1980) proposes the following method for analyzing Gaussian time series containing censored and/or missing observation. Let X_t be a Gaussian process and define $\mu_t = E(X_t)$, $\gamma_{tu} = E(X_t X_u) - \mu_t$. Let $\mathbf{X}$ be a vector containing realizations of X_t at each time t, $\mathbf{X_O}$ be a vector containing the observed values of $\mathbf{X}$, and $\mathbf{X_C}$ be a vector containing the censored/missing values of $\mathbf{X}$. Define $\mu = E(\mathbf{X})$, $\mu_{\mathbf{O}} = E(\mathbf{X_O})$, $\mu_{\mathbf{C}} = E(\mathbf{X_C})$, and $\mathbf{\Sigma} = (\sigma_{tu})$. Let $\mathbf{P} = (\mathbf{P_C} : \mathbf{P_0})$ be a permutation matrix such that $\mathbf{X_C} = \mathbf{XP_C}$ and $\mathbf{X_O} = \mathbf{XP_O}$.

Since X_t is a Gaussian process,

$$
\mathbf{X} \sim N(\mu, \mathbf{\Sigma}) \tag{1}
$$

and $\mathbf{XP} \sim N(\mu, \mathbf{P}'\mathbf{\Sigma}\mathbf{P})$.

Moreover, $\mathbf{X}_c | \mathbf{X}_o = \mathbf{x}_o \sim N(\mathbf{v}, \mathbf{\Delta})$ with

$$
\mathbf{v} = \mu_c + (\mathbf{x}_o - \mu_o)(\mathbf{P}_o'\mathbf{\Sigma}\mathbf{P}_o)^{-1}\mathbf{P}_o'\mathbf{\Sigma}\mathbf{P}_c \tag{2}
$$

and

$$
\mathbf{\Delta} = \mathbf{P}_c'\mathbf{\Sigma}\mathbf{P}_c - \mathbf{P}_c'\mathbf{\Sigma}\mathbf{P}_o(\mathbf{P}_c'\mathbf{\Sigma}\mathbf{P}_o)^{-1}\mathbf{P}_o'\mathbf{\Sigma}\mathbf{P}_c \tag{3}
$$

When $\mathbf{X}_c$ contains only missing observations, the best predictor of the j^{th} element of $\mathbf{X}_c$ is the j^{th} element of of $\mathbf{v}$. When $\mathbf{X}_c$ contains censored observations, it is not the best predictor but can still be used.

Analysis of Trifluralin in 2008

We applied the methods described above to Trifluralin levels in 2008 by assuming $\mathbf{X}$ in (1) is an AR(1) process and estimating $\mathbf{\Sigma}$ using the EM algorithm in Shumway et. al (1982). Then, (2) is our prediction for the missing observations in $\mathbf{X}$ and a 95% confidence interval is constructed using (3). PRZM model outputs of Trifluralin in 2008 from April to November within a certain 6-mile-square townships determined by the Public Land Survey System (PLSS) are plotted below, along with predictions of missing observations in blue and the corresponding 95% confidence interval in grey.

The methods described above work well when the missing observations are scattered throughout the year (as in Figure 2) and when the variation is not large (as in Figure 3). When there are many consecutive observations (as in Figure 4), the predictions seem to follow a perfectly linear trend. When the variation is large (as in Figure 5), the confidence intervals of the predictions are large. In addition, there are many extreme values in this data like the one in mid-April in Figure 3. The confidence intervals also do not account for seasonal differences. One would expect pesticide levels to be much lower in the fall and winter months but the confidence intervals for predictions during these months in Figure 5 do not account for this.

Moving Forward

Although Shumway (1982) and Robinson (1980) are seminal papers in the time series literature, the methods discussed can be improved to predict missing observations in daily pesticide levels more adequately by extending the methods described above to account for spatial dependence structures, extreme values, and seasonal differences.

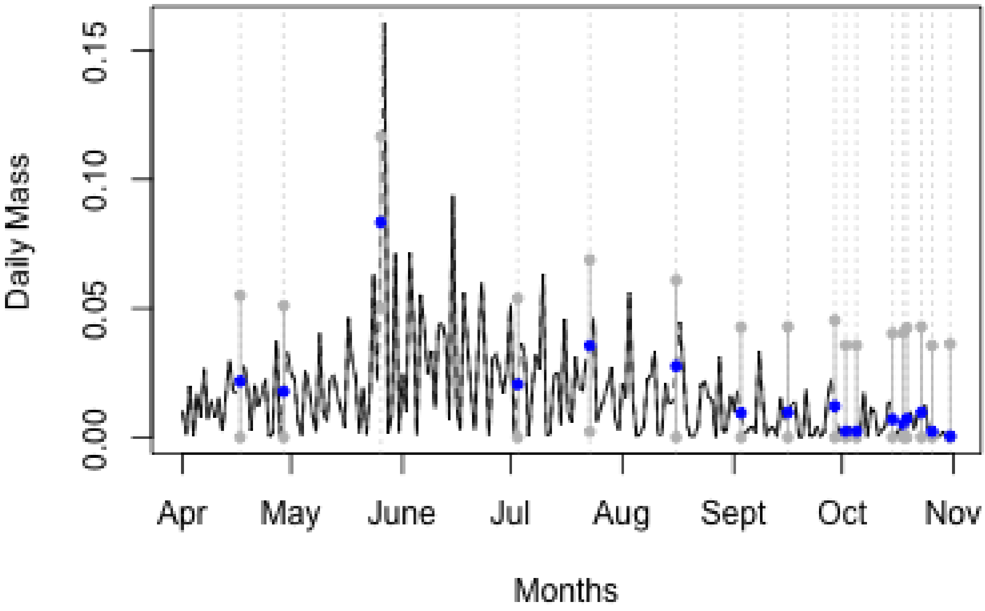


Figure 2: 2008 Trifluralin levels in Dos Palos, CA

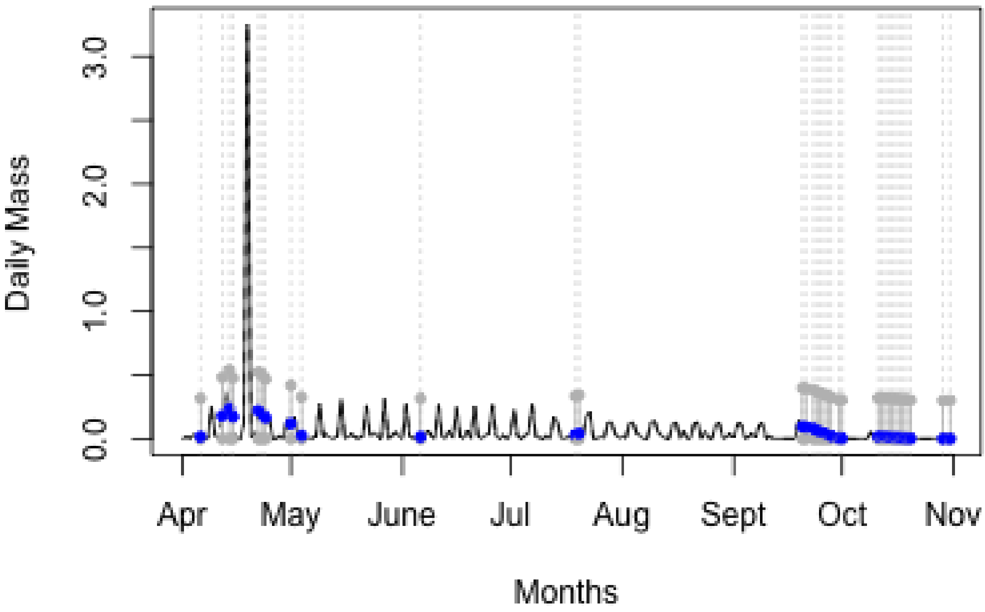


Figure 3: 2008 Trifluralin levels in Glenn, CA

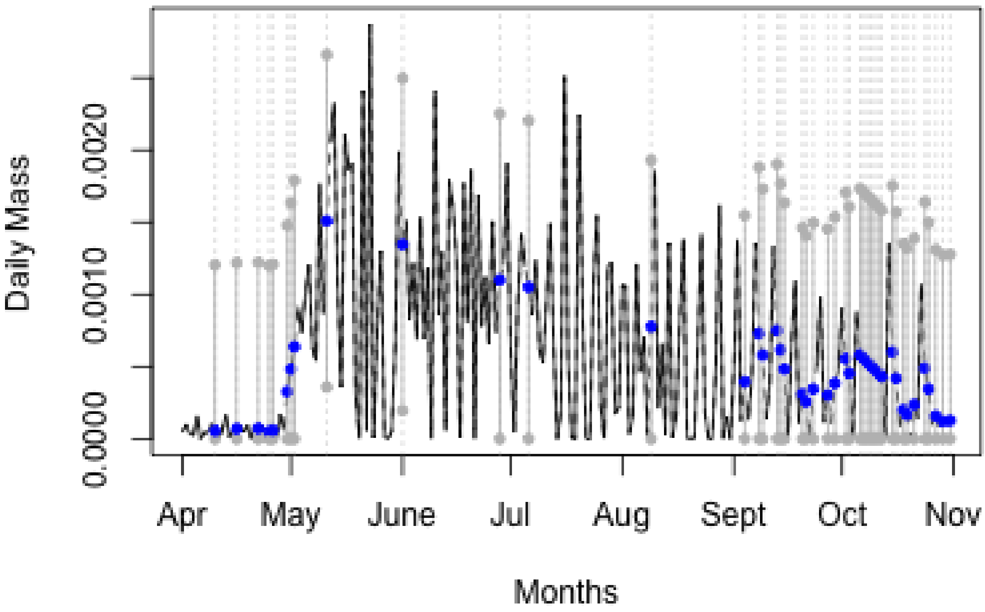


Figure 4: 2008 Trifluralin levels in Rancho Cordova, CA

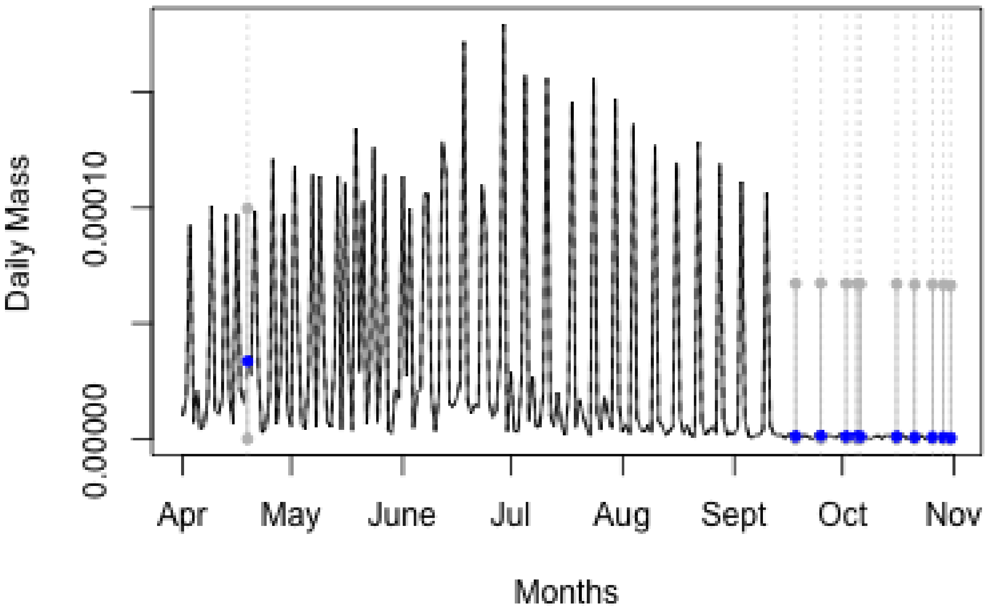


Figure 5: 2008 Trifluralin levels in Stockton, CA