Monitoring Pesticide Concentrations: Database Time Series Analysis

Huong Vu

Submitted in partial fulfilment of the requirements for Highest Honors for the degree of

Bachelor of Science

in

MATHEMATICAL ANALYTICS AND OPERATIONS RESEARCH

in the

College of Letters and Science

of the

University of California,

Davis

Faculty Advisor:

Alexander Aue

June 2018

Contents

Abstract

This thesis considers analyzing pesticide concentrations in CA surface water. There are two major components. The first major component is the creation of a database, pulling information from various sources including the Environmental Protection Agency and the CA Department of Pesticide Regulation. This information is both on monitoring records and on physical attributes. The main characteristic feature of the data is its high degree of missingness as observations on pesticide concentrations are sparse both in time and space. The second major component is the analysis of such pesticide concentrations with common time series models for missing data. It is found that these standard methods do not lead to satisfactory. Future work should look into extensions of time series methods better adapted to pesticide monitoring data.

Acknowledgements

I would like to express my deep gratitude to Professor Alexander Aue, my supervisor, for his patient guidance, enthusiastic encouragement and useful critiques of this research work. Alex has provided me a lot of opportunities and guidance to grow personally and academically. I really appreciate Alex's care and supervision during the time I worked with him. I would also like to thank Ozan Sonmez, Andrew Blandino and Shuhao Jiao, my colleagues in this project, for giving me insightful advice. I enjoyed working and meeting with everyone. My grateful thanks are also extended to Ruriko Imai, my "stat" buddy, for always giving me food, cheering me up through my hardship and for always checking on me.

1 Introduction

According to the Environmental Protection Agency (EPA), pesticide is defined as any substance or mixture of substances intended for preventing, destroying, repelling or mitigating any pest. Pesticide products contain active and inert ingredients. An active ingredient is a component of a pesticide product that controls target pests and inert ingredients are important for performance and usability. By definition, these chemicals kill living things. Using pesticides has an impact on both humans and environment. Humans can be exposed to pesticides directly while applying, formulating or working with pesticides. In addition, oral exposure can happen through consuming food containing pesticides. In the environment, pesticides are present in soil, water, turf or other vegetation. Not only killing insects or weeds, pesticides can be toxic to birds, fish, beneficial insects, and non-target plants. Some pesticides are highly hazardous due to the acute toxicity even with small exposure levels. Other pesticides are persistent in the environment and cannot be washed off easily.

In California, 28 percent of area is farmland and in 2015, 213 million pounds of active ingredients contained in applied pesticides were reported to EPA. With a large amount of pesticides used every year and many harmful consequences on humans and the environment, the tasks of regulating pesticides and monitoring the amount of pesticides in environment become very crucial. Ideally, we would like to measure the concentration directly from streams and rivers. However, due to high cost of direct monitoring, only few direct measurements are taken from streams and rivers. Therefore, models that predict concentrations of pesticides based on available data are important for monitoring pesticides in environment. Predictions of pesticide concentrations cannot replace direct measurements from monitoring sites, but will give us an idea of what to expect for pesticide concentrations at certain locations and act as an indicator for which areas are contaminated and require extra monitoring.

We will focus on modeling concentrations of bifenthrin and imidacloprid in water. We will explore natural and artificial factors such as pesticides usage, runoff rate, climate, etc. that would affect pesticide concentrations in water. Once the model is established, we hope to use it to predict pesticide concentrations at other locations with similar characteristics.

2 Data Overview

For the DPR project, we use Catchment Attributes, Monitoring Data and Pesticide Usage Reporting datasets. The Catchment Attributes dataset is collected from United States Environmental Protection Agency (US EPA) database. Both Monitoring Data and Pesticide Usage Reporting datasets are collected from California Department of Pesticide Regulation (CDPR) database. Because the datasets are provided by different agencies, the ways how each dataset is stored and sorted are also different. One of the essential tasks in this project is to figure out how to access the databases, download necessary datasets and clean the datasets. The goal of this step is to ensure a connection between the three datasets.

2.1 Catchment Attributes

In this thesis, catchments represent the portion of the landscape, where surface flow drains directly into a stream segment, excluding any upstream contributions. A stream segment is a section of contiguous stream or river between upstream and downstream tributaries, except where the segment is a headwater or terminal stream. A watershed is a set of hydrologically connected catchments, consisting of all upstream catchments that contribute flow into any catchment in the watershed.

The Stream-Catchment (StreamCat) database describes natural and anthropogenic landscape features for both individual stream catchments and cumulative upstream watersheds, based on the National Hydrography Dataset Plus Version 2 (NHDPlusV2) geospatial framework. Our job is to extract the attributes of monitoring stations located within the HUC14 system and their upstream watersheds. HUC is a system of unique hydrologic unit codes to identify the United States hydrologic units. Senior Environmental Scientist at DPR, Dan Wang, provided us an Excel document containing a list of California Central Coast monitoring stations and their ComID. ComID is a unique identifier for a stream segment that can be linked to NHDPlusV2 stream or catchment shapefiles. The list contains 255 stations located in 188 catchments. Some stations do not have ComID or are coded differently from the rest because they locate in small sea and are not considered in stream network. From the EPA website[1], attributes for California can be downloaded. Attributes are categorized and grouped into 51 data files based on their similarities. We used R to match given comIDs of Central Coast catchments with the comIDs listed in attribute data file (see the Appendix for the R code). From the StreamCat database, 692 attributes were collected for 188 catchments

[1] ftp://newftp.epa.gov/EPADataCommons/ORD/NHDPlusLandscapeAttributes/StreamCat/States/

along the California Central Coast and were compiled to output the Catchment Attributes dataset. Out of 693 attributes, Ruoyu Wang, postdoc at Department of Land, Air and Water Resources, UC Davis suggested that 258 attributes are useful in predicting the pesticide concentrations. In those 258 attributes, four attributes are essential data from NHDPlus, 220 are land use and land cover, 10 are soil, 22 are weather and two are processed results from CalFed physical model.

CalFed physical model determines the relative ranking of potential areas of concern with respect to pesticide exposure to sensitive and endangered aquatic species. First, the study area was characterized based on the physical landscape. Then, for each pesticide considered in the model, environmental-fate properties were collected, along with historical pesticide applications. CalFed model also used PRZM and RICEWQ environmental fate models to predict edge-of-field mass and pesticide runoff. To compute a daily concentration, the daily mass loadings (sum of daily pesticide mass loadings from agricultural fields, rice paddies, and urban areas) were "mixed" into the streams and rivers. If the estimated concentration exceeded ecotoxicological benchmarks, it was considered an indicator event and used in the co-occurrence analysis. In co-occurence analysis, frequency of indicator events is calculated and used in conjunction with monthly estimates of species presence to determine co-occurrence. Area with high co-occurence was considered as potential candidate for extra monitoring.

2.2 Monitoring Data

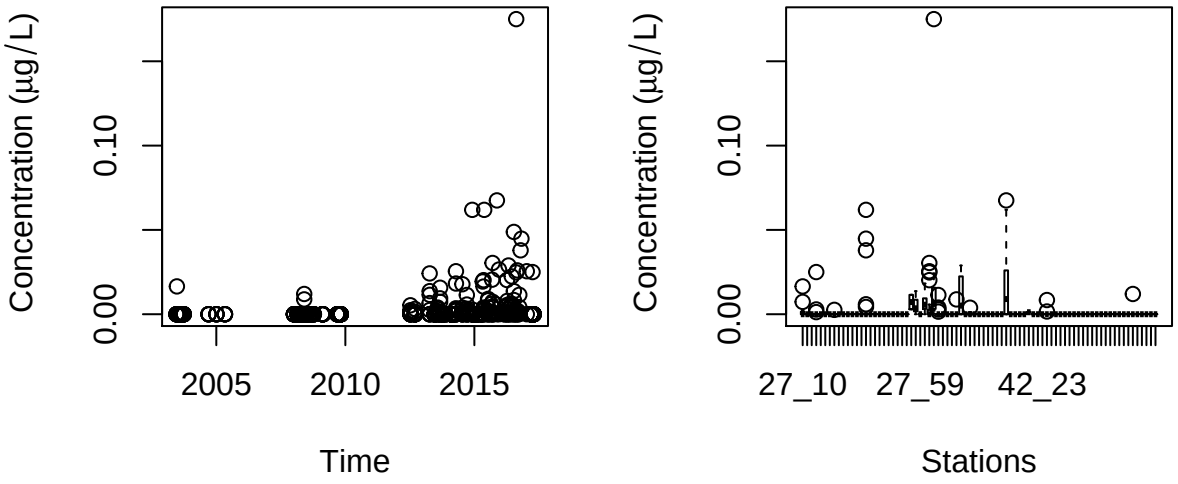
Department of Pesticide Regulation provided monitored concentrations of two important and commonly used pesticides, bifenthrin and imidacloprid, in Central Coast. The common structure for pesticide concentration datasets includes location information (site code, county, longtitude, latitude, etc.), sampling date, measured concentration, sampling method (limit of quantitation, agency, sampling code, etc.), and data source. For all pesticides, the monitoring data is extremely sparse in space and time. There are a lot of missing observations and censored data, which are measurements below the LOQ.

2.2.1 Bifenthrin

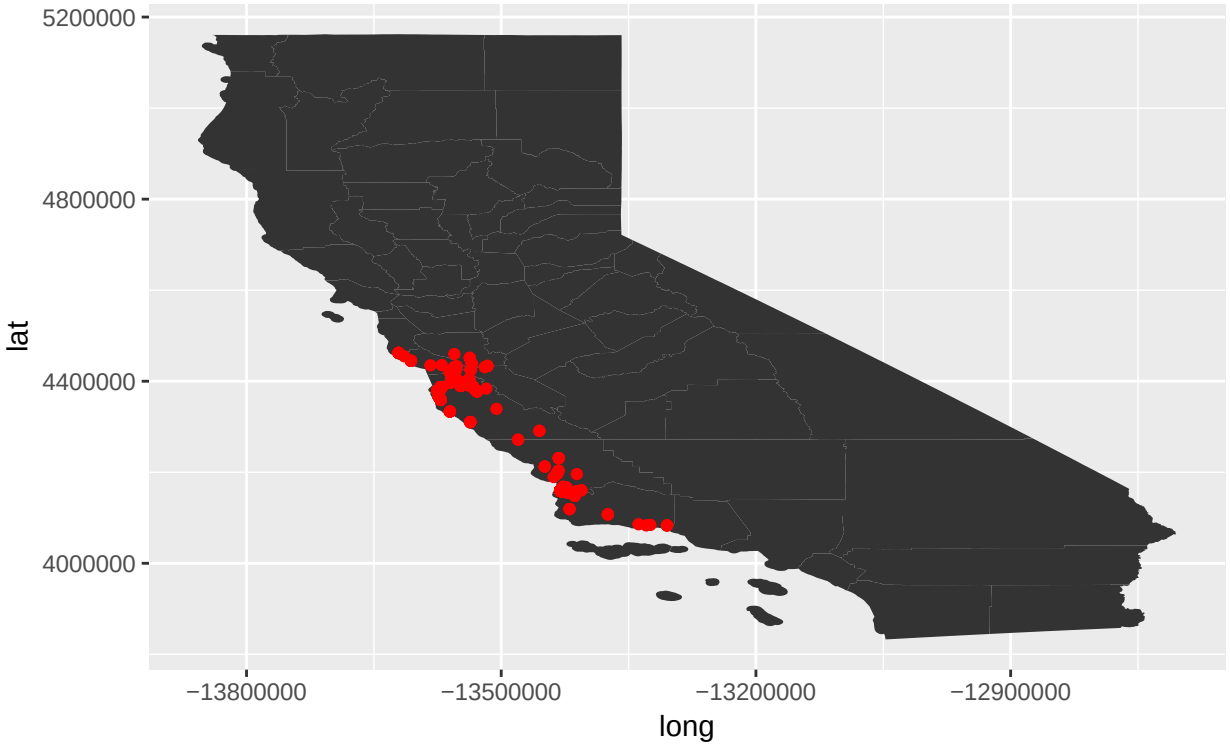
Bifenthrin is an insecticide that is used on various agricultural crops and in homes. In the United States, bifenthrin is contained in over 600 products in forms of sprays, granules and aerosols. Bifenthrin is higly toxic to fish and small aquatic organisms and is classified as carcinogen. Since bifenthrin binds tightly to soil, its concentration is measured in both water

and sediment.

- Bifenthrin concentration in water: The sampling period for bifenthrin pesticide in water is from June 16, 2003 to April 20, 2017. The samples are taken from 79 unique stations on 148 unique dates. There is no clear pattern in sampling frequency at the monitoring stations because some years have too few observations. However, April, May and June are sampled the most compared to other months. The averages of concentrations measured from stations that are in the same catchment are shown in Firgure 2.1b. As we can see from the figure, the monitoring stations locate throughout Central Coast.

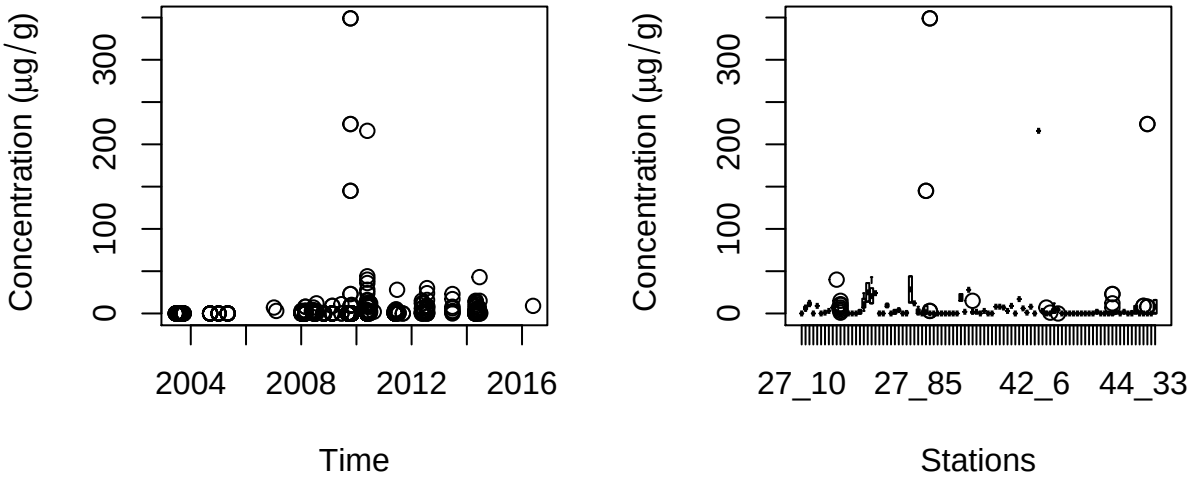


[Figure 2.1(a,b): Scatter plot of bifenthrin concentration in water in space and time]

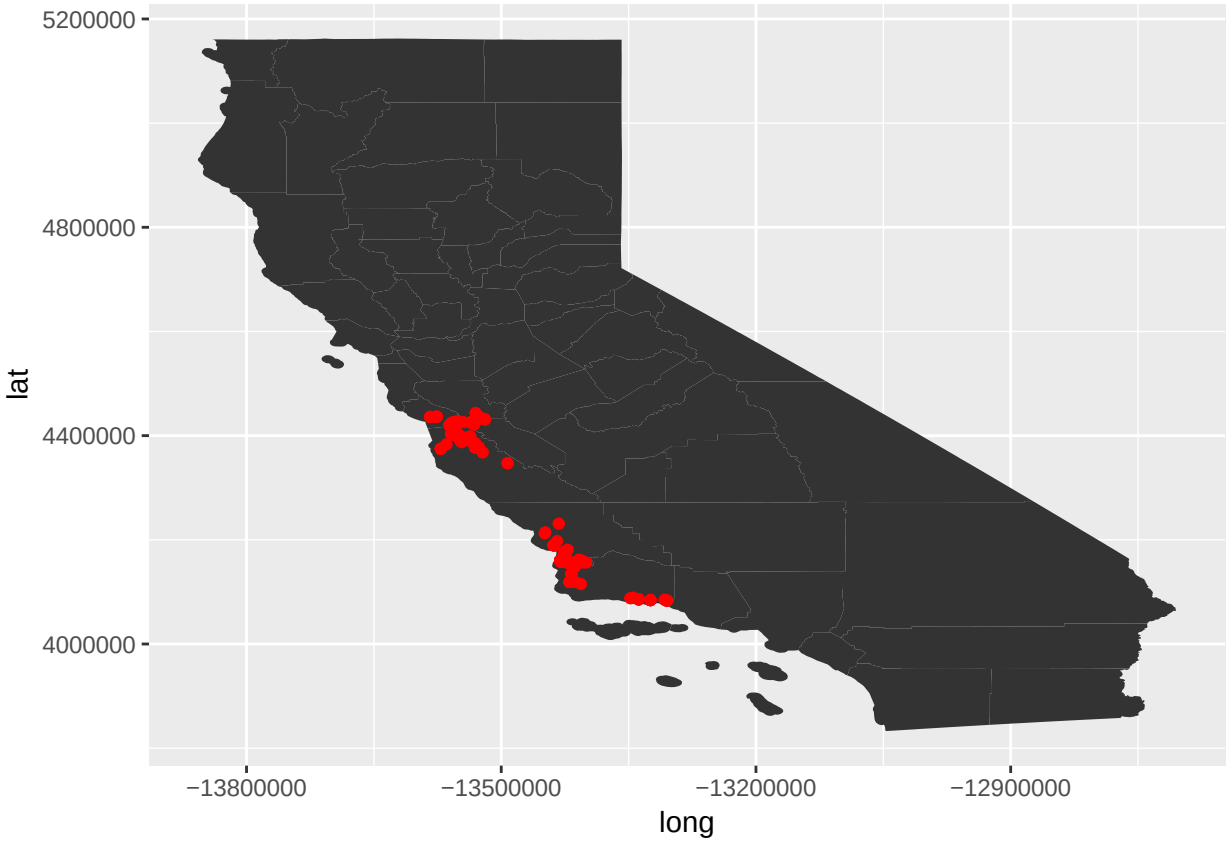


[Figure 2.2: Map of bifenthrin in water monitoring stations]

- Bifenthrin concentration in sediment: The sampling period for bifenthrin concentration in sediment is from June 16, 2003 to May 26, 2016. Samples are taken from 92 unique stations with 88 unique sampling dates. There is no clear pattern of sampling frequency among the stations. Monitoring stations for bifenthrin in sediment locate mainly in Santa Cruz, Watsonville, Salinas (in Santa Cruz and Monterey county), Arroyo Grande, Santa Maria (in San Luis Obispo county) and Santa Barbara (in Santa Barbara county).



[Figure 2.3(a,b): Scatter plot of bifenthrin concentration in sediment in space and time.]

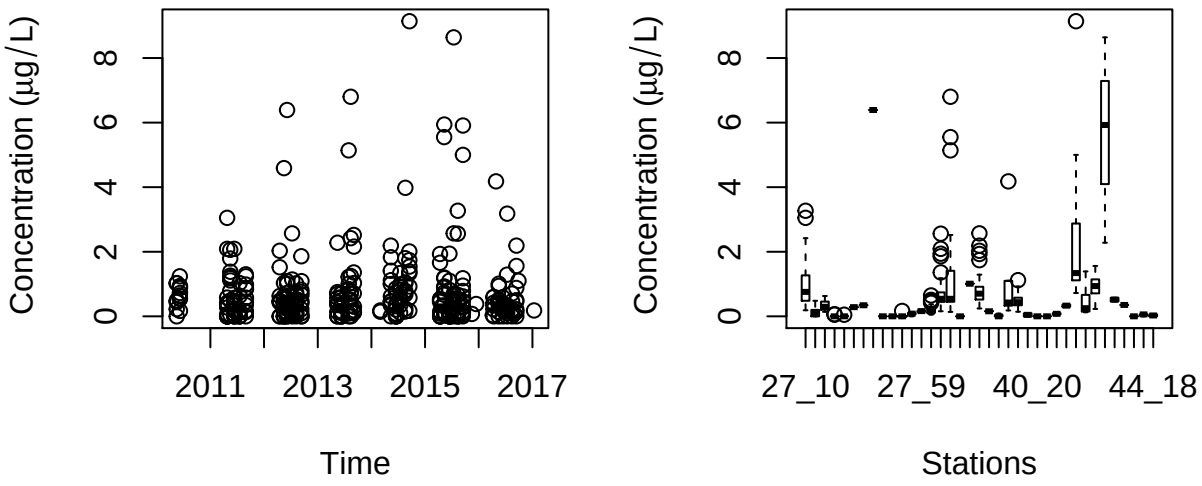


[Figure 2.4: Map of bifethrin in sediment monitoring stations]

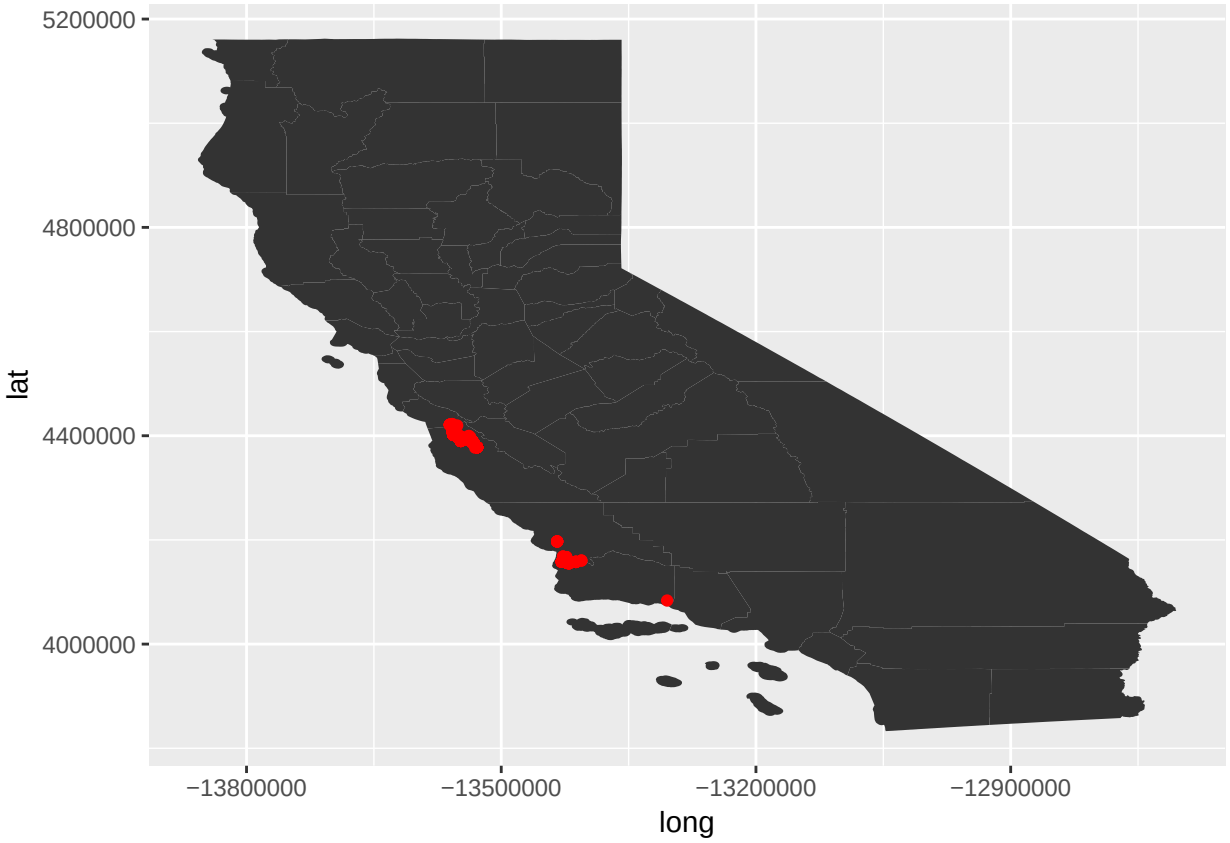
2.2.2 Imidacloprid

Imidacloprid is an insecticide that is used to control sucking insects, termites, some soil insects, and flees on pets. Over 400 products in the United States in forms of liquids, granules, dusts, and packages that dissolve in water contain imidacloprid. The effect of exposure to imidaclorpid in fish varies by species. Some native fish in California Central Coast streams and rivers are salmon, steelhead and trout and they are commonly consumed by humans. Therefore, high levels of imidacloprid concentrations in water have a direct impact on humans.

The concentration dataset of imidacloprid contains 366 observations which were taken from 37 unique stations on 78 unique dates. The sampling period of imidacloprid is shorter than bifenthrin pesticide, only from 17 May, 2010 to 12 October, 2017. Samples are most likely taken once every month, and April to September have the highest frequency of sampling. The monitoring stations gather around Watsonville, Salinas and Santa Maria.



[Figure 2.5(a,b): Scatter plot of imidacloprid concetration from in space and time.]



[Figure 2.6: Map of imidaclopid monitoring stations]

2.3 Pesticide Usage Reporting

California Department of Pesticide Regulation (CDPR) publishes pesticide usage in agriculture and urban land every year since 1978[2]. The Pesticide Usage Reporting (PUR) datasets list agricultural and non-agricultural pesticide use sorted by county number. Agricultural use of pesticide is required to report to DPR weekly. Each report provides information about amount of used pesticide products, amount of active chemicals in used pesticide products, date of application, area of land that is treated and land location identifier (CO_MTRS). Each CO_MTRS represents a one square meter land. The county, meridian, township, range and section of a treated land are coded in CO_MTRS that can be linked to a longitude latitude coordinate in the sshapefile provided by the CDPR. Non-agricultural pesticide usage reports are made monthly and do not provide specific spatial coordinates where the pesticides are applied. Therefore, non-agricultural pesticide usage is stored separately from the agricultural pesticide usage dataset. Note that CO_MTRS does not uniquely identify a spatial location in some instances. This happens when there is a stream/river present in the area that is linked to COM_MTRS. Hence, the area belongs to different catchments. However, the reported pesticide use still reflects the total amounnt of pesticide applied on the area.

[2]link:http://www.cdpr.ca.gov/docs/pur/purmain.htm

3 Time Series

In this section, we will discuss stationary time series and how to model a stationary time series with ARMA processes.

A time series is a collection of random variables, x_t, indexed by $t \in T$, where $T \subseteq \mathbb{Z}$. A time series is composed of deterministic trend, seasonal component and random fluctuations:

$$X_t = m_t + s_t + Y_t, \quad t \in T, \tag{1}$$

where $(m_t)_{t\in T}$ denotes the trend function, $(s_t)_{t\in T}$ the seasonal effects and $(Y_t)_{t\in T}$ the residuals. Time series are used widely in many different fields such as economics, social science, biology and physics.

Time series modeling steps include 1) removing trend and seasonal components from the series, 2) assessing the residuals, 3) modeling the residuals.

3.1 Trend and Seasonal Components Removal

- Assuming $s_t = 0$ for all $t \in T$, then $X_t = m_t + Y_t$ with $E[Y_t] = 0$. There are three methods to estimate the trend of this time series.

  1. Least Squares Estimation
  2. Smoothing with Moving Averages
  3. Differencing

- If $s_t \neq 0$ such that $s_{t+d} = s_t$ and $\sum_{j=1}^{d} s_j = 0$ where d is the period of the seasonal component, then

  $$X_t = m_t + s_t + Y_t, \quad t \in T$$

  with $E[Y_t] = 0$. We also have three methods to remove trend and seasonal components from a given time series.

  1. Small Trend Method
  2. Moving Average Estimation
  3. Differencing at lag d

3.2 Stationary Time Series

Before modelling residuals, we need to make sure that the residuals are stationary. There are two classes of stationarity, strict stationarity and weak stationarity.

Definition:(Strict Stationarity) A stochastic process $(X_t)_{t\in T}$ is called strictly stationary if, for all $t_1, \ldots, t_n \in T$ and h such that $t_1 + h, \ldots, t_n + h \in T$, it holds that

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{\mathcal{D}}{=} (X_{t_1+h}, \ldots, X_{t_n+h})$$

That is, the so-called finite-dimensional distributions of the precess are invariant under time shifts. Here, $\stackrel{\mathcal{D}}{=}$ indicates equality in distribution.

Definition:(Weak Stationarity) A stochastic process $(X_t)_{t\in T}$ is called weakly stationary if

- the second moments are finite: $E[X_t^2] < \infty$ for all $t \in T$;
- the means are constant: $E[X_t] = m$ for all $t \in T$;
- the covariance of X_t and X_{t+h} depends on h only:

$$\gamma(h) = \gamma_x(h) = Cov(X_t, X_{t+h}), \quad h \in T \text{ such that } t + h \in T,$$

is independent of $t \in T$ and is called the autocovariance function (ACVF). Moreover,

$$\rho(h) = \rho_X(h) = \frac{\gamma(h)}{\gamma(0)}, \quad h \in T$$

is called the autocorrelation function (ACF).

3.3 ARMA Processes

After obtaining a stationary time series, we can model the series using ARMA models. By the decompostion theorem, any stationary series can be approximated by stationary ARMA model.

Definition:(ARMA processes) (a) A weakly stationary process $(X_t)_{t\in Z}$ is called an autoregressive moving average time series of order (p, q), abbreviated by ARMA(p, q), if it satisfies the following equation

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t + \phi_1 Z_{t-1} + \cdots + \phi_q Z_{t-q}, \quad t \in \mathbb{Z}, \tag{2}$$

where $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are real constants, $\phi_p \neq 0 \neq \theta_q$ and $(Z_t)_{t\in\mathbb{Z}} \sim \text{WN}(0, \sigma^2)$.

(b) A weakly stationary stochastic process $(X_t)_{t\in\mathbb{Z}}$ is called an ARMA(p,q) time series with mean μ if the precess $(X_t \sim \mu)_{t\in\mathbb{Z}}$ satisfies the equation system above.

We define the autoregressive polynomial and the moving average polynomial by

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p, \quad z \in \mathbb{C}$$
$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q, \quad z \in \mathbb{C}$$

Together with the backshift operater, B, we can write equation (2) in a concise form as

$$\phi(B)X_t = \theta(B)Z_t, \quad t \in \mathbb{Z}$$

Definition (Linear processes): A stochastic process $(X_t)_{t\in\mathbb{Z}}$ is called linear process or MA(∞) time series if there is a sequence $(\psi_j)_{j\in\mathbb{N}_0}$ with $\sum_{j=0}^{\infty} |\psi_j| < \infty$ such that

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \quad t \in \mathbb{Z},$$

where $(Z_t)_{t\in\mathbb{Z}} \sim \text{WN}(0, \sigma^2)$. We can also write a linear process as $X_t = \psi(B)Z_t$ where $t \in \mathbb{Z}$.

3.3.1 Causality and Invertibility Properties of ARMA Processes

An MA(q) model is always stationary without conditions on the coefficients $\theta_1, \ldots, \theta_q$. An AR(1) process can be written as

$$X_t = \phi X_{t-1} + Z_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$$

with $|\phi| < 1$. This is a linear process. Hence, AR(1) is stationary. If $|\phi| > 1$, the autoregressive processes of order one are called explosive. From the above cases of AR(1) models, we would like to define the notion of causality which means the process $(X_t)_{t\in\mathbb{Z}}$ has a representation in terms of the white noise $(Z_s)_{s\leq t}$ and is hence independent of the future as given by $(Z_s)_{s>t}$.

Definition (Causality): An ARMA(p,q) process given by (2) is causal if there is q sequence $(\psi_j)_{j\in\mathbb{N}_0}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \quad t \in \mathbb{Z}$$

Theorem: Let $(X_t)_{t\in\mathbb{Z}}$ be an ARMA(p,q) process such that the polynomials $\phi(z)$ and $\theta(z)$ have no common zeroes. Then $(X_t)_{t\in\mathbb{Z}}$ is causal if and only if $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ with $|z| \leq 1$. The coefficients $(\psi_j)_{j\in\mathbb{N}_0}$ are determined by the power series expansion

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \quad |z| \leq 1$$

Another concept closely related to causality is invertibility.

Definition (Invertibility): An ARMA(p, q) process given in (2) is invertible if there is a sequence $(\pi_j)_{j\in\mathbb{N}_0}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \quad t \in \mathbb{Z}$$

Theorem: Let $(X_t)_{t\in\mathbb{Z}}$ be an ARMA(p, q) process such that the polynomials $\phi(z)$ and $\theta(z)$ have no common zeros. Then $(X_t)_{t\in\mathbb{Z}}$ is invertible if and only if $\theta(z) \neq 0$ for all $z \in \mathbb{C}$ with $|z| \leq 1$. The coefficients $(\pi_j)_{j\in\mathbb{N}_0}$ are determined by the power series expansion

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \quad |z| \leq 1$$

3.3.2 The PACF of a Causal ARMA Process

Definition (Partial autocorrelation function): Let $(X_t)_{t\in\mathbb{Z}}$ be a weakly stationary stochastic process with zero mean. Then, we call the sequence $(\phi_{hh})_{h\in\mathbb{N}}$ given by

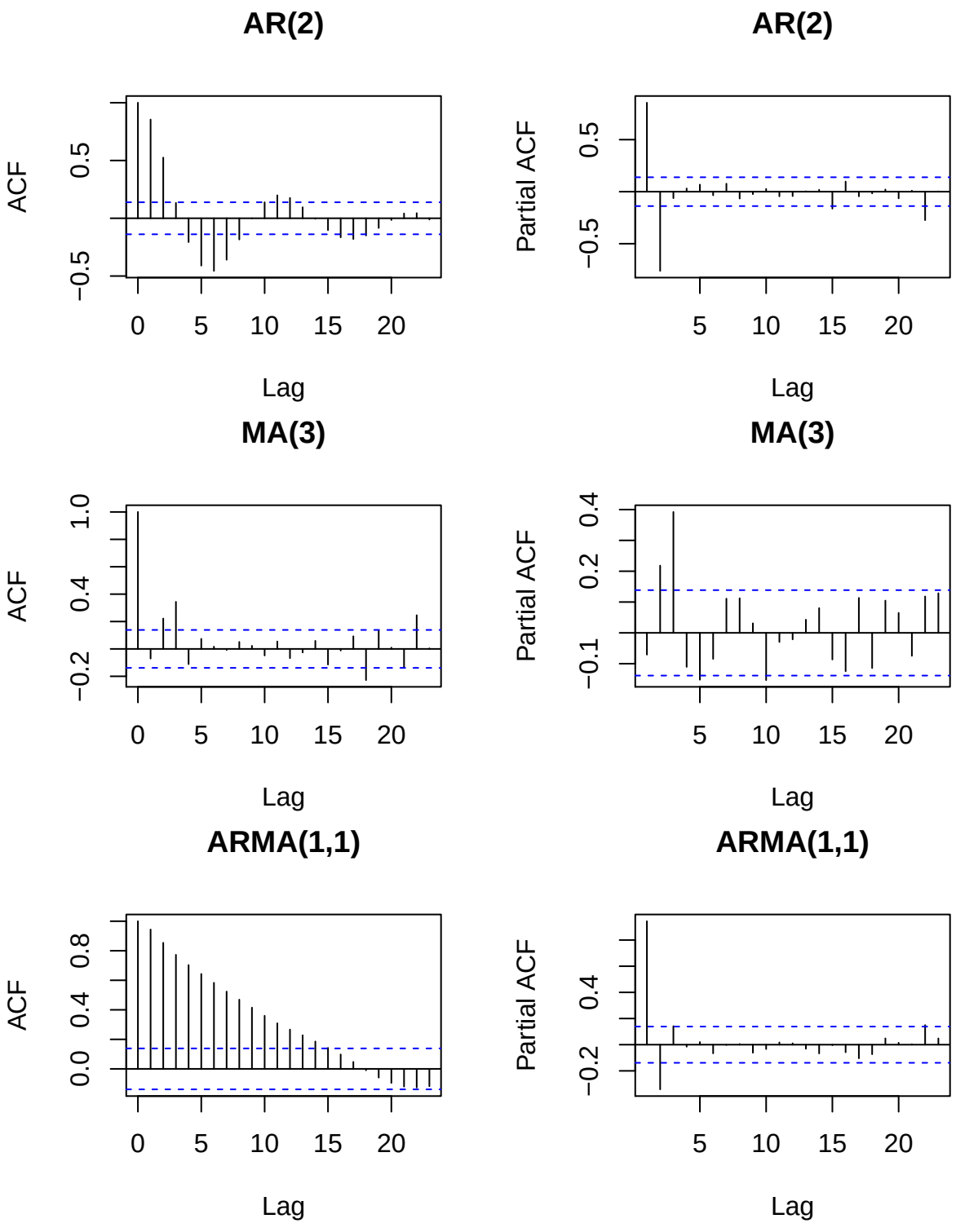
$$\begin{aligned}
\phi_{11} &= \rho(1) = Corr(X_1, X_0), \\
\phi_{hh} &= Corr(X_h - X_h^{h-1}, X_0 - X_0^{h-1}), \quad h \geq 2,
\end{aligned}$$

the partial autocorrelation function (PACF) of $(X_t)_{t\in\mathbb{Z}}$. Therein,

$$\begin{aligned}
X_h^{h-1} &= \text{regression of } X_h \text{ on } (X_{h-1}, \ldots, X_1) \\
&= \beta_1 X_{h-1} + \beta_2 X_{h-2} + \cdots + \beta_{h-1} X_1 \\
X_0^{h-1} &= \text{regression of } X_0 \text{ on } (X_1, \ldots, X_{h-1}) \\
&\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{h-1} X_{h-1}
\end{aligned}$$

	AR(p)	MA(q)	ARMA(p,q)
ACF	tails off	cuts off after lag q	tails off
PACF	cuts off after lag p	tails off	tails off

[Table 1: The behavior of ACF and PACF for AR, MA, and ARMA processes]



[Figure 3.1: ACFs and PACFs of an AR(2)-top, MA(3)-middle and ARMA(1,1)-bottom]

3.3.3 Forecasting

Definition (One-step Best Linear Predictors (BLP)): Given the observed variables $X_1, \ldots, X_n$ of a weakly stationary time series $(X_t)_{t\in\mathbb{Z}}$, one-step best linear predictors are linear combinations

$$\hat{X}_{n+1} = \phi_{n0} + \phi_{n1}X_n + \cdots + \phi_{nn}X_1$$

of the observed variables $X_1, \ldots, X_n$ that minimize the mean-squared error

$$E[X_{n+1} - g(X_1, \ldots, X_n)^2]$$

for functions g of $X_1, \ldots, X_n$.

Theorem (Best linear prediction): Let $(X_t)_{t\in\mathbb{Z}}$ be a weakly stationary stochastic process of which we observe $X_1, \ldots, X_n$. Then, the one-step BLP $\hat{X}_{n+1}$ of X_{n+1} is determined by the equations

$$E[(X_{n+1} - \hat{X}_{n+1})X_{n+1-j}] = 0 \tag{3}$$

for all $j = 1, \ldots, n+1$ where $X_0 = 1$.

Assume $(X_t)_{t\in\mathbb{Z}}$ has mean 0 ie. $\phi_{n0} = 0$. Then, with the ACVF γ of $(X_t)_{t\in\mathbb{X}}$, the equation (3) in theorem above can be expressed as

$$\sum_{l=1}^{n} \phi_{nl}\gamma(j-l) = \gamma(j), \quad , j = 1, \ldots, n \tag{4}$$

In matrix notation, let $\mathbf{\Gamma}_n = (\gamma(j-l))_{j,l=1,\ldots,n}$, $\boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nn})^T$ and $\boldsymbol{\gamma}_n = (\gamma(1), \ldots, \gamma(n))^T$. With these notations, equation (4) becomes

$$\mathbf{\Gamma}_n\boldsymbol{\phi}_n = \gamma_n \iff \boldsymbol{\phi}_n = \mathbf{\Gamma}_n^{-1}\boldsymbol{\gamma}_n, \tag{5}$$

provided that $\mathbf{\Gamma}_n$ is nonsingular.

Let $\mathbf{X}_n = (X_n, X_{n-1}, \ldots, X_1)^T$, then $\hat{X}_{n+1} = \boldsymbol{\phi}_n^T\mathbf{X}_n$. To access the quality of the prediction, the mean-squared error can be computed as following:

$$\begin{aligned} P_{n+1} &= E[(X_{n+1} - \hat{X}_{n+1})^2] \\ &= \gamma(0)\boldsymbol{\gamma}_n^T\mathbf{\Gamma}_n^{-1}\boldsymbol{\gamma}_n \end{aligned}$$

For large n, it is an expensive process to compute the inverse matrix $\mathbf{\Gamma}_n^{-1}$. There are recursive prediction methods that help us avoid taking the inverse of $\mathbf{\Gamma}_n$. They are Durbin-Levinson algorithm, innovations algorithm and prediction based on the infinite past.

3.3.4 Parameter Estimation

For a causal and invertible ARMA(p, q) process with zero mean, let $\boldsymbol{\beta} = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)^T$ be the parameter vector. Three methods of parameter estimation are method of moments, maximum likelihood estimation and least square estimation. Method of moments works best in case of pure AR processes. For the general ARMA processes, maximum likelihood and least square methods provide more efficient estimators.

4 Missing Values Estimation

4.1 Statistical Approaches to Assessing Pesticide Concentrations in the DPR Surface Water Database

Robert H. Shumway was a Statistics professor at University of California, Davis. Time Series Analysis is one of his research interests, and he had worked on DPR Pesticide Database in 2000 and 2001. One of the datasets that Shumway analyzed contained measurements taken in the Orestimba Creek tributary of the San Joaquin River (locations include State Highway 33, Crow Creek drain and River Road) on pesticides chlorpyrifos, diazinon and methidathion during the period April, 1996 to May, 1997. This dataset had similar issues of sparsity and strong cencoring as discussed above. Therefore, Shumway used Kalman filtering and smoothing to pre-process data before model fitting. In addition, a fourth root transformation was applied to stabilize variances and to improve the approximation to stationarity. Since the analysis was done on two different pesticides at three different locations, Shumway calculated the correlation matrix that reflected the spatial correlation between different pesticides at different locations. The ACF plot indicated a non-stationary or long memory process while PACF plot showed an AR(1) model would suffice.

Shumway also used the state model

$$\overrightarrow{Y}_t = \overrightarrow{x}_t + \overrightarrow{v}_t \tag{6}$$

where A is a $q \times p$ matrix and Y_t is an independent multivariate normal vector with mean 0 and common (identical) $q \times q$ matrix R.

The unobserved data is assumed to evolve through time and space. Then, the state equation is given as

$$\overrightarrow{x}_t = \Phi \overrightarrow{x}_{t-1} + \overrightarrow{w}_t \tag{7}$$

where Φ is a $p \times p$ matrix that summarizes the space-time regression relation for the unobserved signal and $\overrightarrow{w}_t$ are independent normal noise vectors with a common $p \times p$ covariance matrix Q.

Transformed diazinon concentrations at three locations are modelled as

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

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

matrix and the state equation is

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

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

Since the purpose of the paper is to predict the pesticide concentrations in a given area, we need a model that expresses observed series in terms of a common signal. To deal with the problem of irregular sampling and long sequences of observed values that are below detection limits, irregularly observed and episodic fluctuations are merged into a common signal. Probability limits for that signal are of interest for determining if standards have been exceeded.

In the example of diazinon concentration mentioned above, the first step of the procedure is to measure the transition matrix Φ, the observation covariance matrix, R, and model covariance matrix Q. Those parameters can be estimated by maximum likelihood method. Then, Kalman filters and smoothers are used to estimate the unobserved process, $\overrightarrow{x}_t$ and its uncertainty by the following equations.

The state-space model defines the signal $\overrightarrow{X}_t$ in terms of the following equations

$$\overrightarrow{Y}_t = A\overrightarrow{x}_t + \overrightarrow{v}_t$$
$$\overrightarrow{x}_t = \phi\overrightarrow{x}_{t-1} + \overrightarrow{w}_t$$

The signal is estimated by the Kalman smoothed values

$$x_t^n = E\{x_t|y_1, \ldots, y_n\} \tag{8}$$

The uncertainty of the smoothed values is expressed as the mean square covariance

$$P_t^n = E[(x_t^n - x_t)(x_t^n - x_t)'|y_1, \ldots, y_n] \tag{9}$$

Under the assumption that the errors $\overrightarrow{y}_t$ and $\overrightarrow{w}_t$ are normally distributed, prediction intervals, at any given probability level, available from two equations (8) and (9) above.

4.2 Interpolating Missing Values in a Time Series

In time series, a period of missing values can be estimated by forecasting from previous values and backcasting from later values. Eivind Damsleth introduced an algorithm that combined forecasts and backcasts into between-forecasts with a minimum forecast error. This method

requires the data to have sufficient length for parameters and the model to be estimated correctly. We also presume that the parameters of the ARIMA-model are known.

First, the notations used for this algorithm are defined as following:

- A forward representation of an ARIMA(p, d, q) process is defined as

$$\phi(B)(1-B)^d X_t = \theta(B)E_t$$

where B is the backward shift operator and

$$\begin{aligned}\phi(B) &= 1 - \phi_1 B - \cdots - \phi_p B^p \\ \theta(B) &= 1 - \theta_1 B - \cdots - \theta_q B^q, \quad \text{such that} \quad \phi(B)\theta(B) = 0\end{aligned}$$

and $\{E_t\}$ is a sequence of independent identically distributed normal random variables with mean zero and variance σ_z^2.

- Given the observations $x_q, x_{q-1}, \ldots$ the minimum mean square error for forward forecast of $x_{q+l} (l \geq 1)$ at time q is

$$\hat{e}_q(l) = \sum_{j=0}^{l-1} \psi_j E_{q+l-j}, \quad \psi_0 = 1$$

where ψ_j are defined by $\theta(B) = (1 + \psi_1 B + \psi_2 B^2 + \cdots)\phi(B)(1-B)^d$.

- A backward representation of an ARIMA(p, d, q) model is expressed as

$$\phi(F)(1-F)^d x_t = \theta(F)c_t$$

where F is the forward shift operator such that $F^j x_t = x_{t+j}$ and $\{c_t\}$ is a sequence of indpedent identically distributed normal random variables with mean zero and variance σ_c^2 with $\sigma_z^2 = \sigma_c^2$.

- Given $x_{q+m+j} (j \geq l)$ the minimum mean square error for backward forecast x_{q+l} at time $q + m + l$ is

$$\tilde{e}_s(m+1-l) = \sum_{j=0}^{m-l} \psi_j c_{q+l+j}, \quad s = q + m + l$$

The algorithm for finding the best linear combination between forecasting and backcasting is given as

i) Calculate the optimal forecast $\hat{e}_q(l)$ using $\phi(B)(1-B)^d X_t = \theta(B)E_t$ and the optimal backcast $\tilde{e}_s(m+1-l)$ using $\phi(F)(1-F)^d x_t = \theta(F)c_t$.

ii) Calculate the coefficients π_j of B^j in the polynomial expansion of $\frac{\theta(B)}{\phi(B)}$ for $j = 0, 1, \ldots, \max(l-1, m-l)$.

iii) Calculate the cross covariance function $\gamma_{ae}(j)$ for $j = 0, 1, \ldots, m-l$, using

$$\theta(F)(1 - \sum_{i=1}^{p+j} \phi_j B^j)\gamma_{ae}(j-s) = \pm\alpha_j\sigma^2, \quad -p \le j \le 0,$$
$$\theta(B)(1 \sum_{i=1}^{q-j} \theta_i F^i)\gamma_{ae}(j-s) = \pm\alpha_j\sigma^2, \quad 0 \le j \le q$$

This gives $p+q+1$ linear equations in $p+q+1$ unknowns $\gamma_{ae}(-p-s), \gamma_{ae}(-p-s+1), \ldots, \gamma_{ae}(q-s-1, \gamma_{ae}(q-s)$, which can be solved. The solutions will provide starting values for the difference equation $\theta(F)\phi(B)\gamma_{ae}(j-s) = \pm\alpha_j\sigma^2$, $j = \ldots, -1, 0, 1, \ldots$ where α_j is the coefficient of B^j in $\theta(B)\phi(B^{-1})$, given by the following:

$$\alpha_j = \begin{cases} 0 & j < -p \\ -\phi_{-j} + \sum_{j=1}^{min(p+j,q)} \phi_{i-j}\theta_i & -p \le j \le -1 \\ 1 + \sum_{i=1}^{min(p,q)} \phi_i\theta_i & j = 0 \\ -\theta_j + \sum_{i=1}^{min(p,q-j)} \phi_i\theta_{i+j} & 1 \le j \le q \\ 0 & j > q \end{cases}$$

iv) Calculate

$$v(\hat{e}) \text{ from } (\sum_{j=0}^{l-1} \psi_j^2)\sigma_z^2,$$
$$v(\tilde{e}) \text{ from } (\sum_{j=0}^{m-1} \psi_j^2)\sigma_z^2, \quad \text{and}$$

$$\sigma_{12} = \sum_{i=0}^{l-1}\sum_{k=0}^{m-l} \psi_i\psi_j\gamma_{ae}(i+j)$$

where $\gamma_{ae}(k)$ is the cross covariance function between $\{a_k\}$ and $\{e_l\}$.

v) If $\{X_t\}$ is stationary then calculate c and d as the solutions to

$$(Ex^2 - \sigma_1^2)c + (Ex^2 + \sigma_{12} - \sigma_1^2 - \sigma_2^2)d = Ex^2 - \sigma_1^2$$
$$(Ex^2 + \sigma_{12}^2 - \sigma_1^2 - \sigma_2^2)c + (Ex^2 - \sigma_2^2)d = Ex^2 - \sigma_2^2$$

In the non-stationary case, we calculate c from

$$c = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} \quad \text{and } d = 1 - c$$

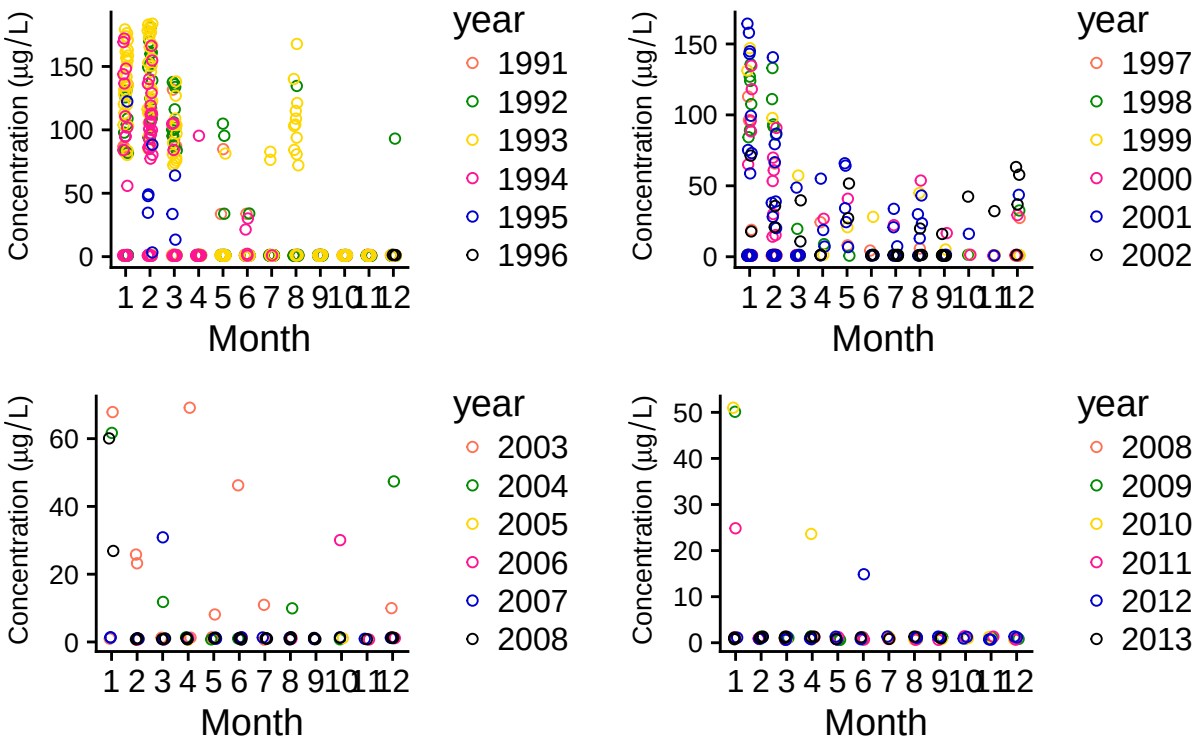
vi) The optimal between-forecast of x_{r+l} is then given by

$$c(\hat{e}_q(l) + Ex) + d(\tilde{e}_{q+m+l}(m + 1 - l) + Ex)$$

4.3 Interpolating Missing Values for Diazinon Dataset

4.3.1 Data Exploration

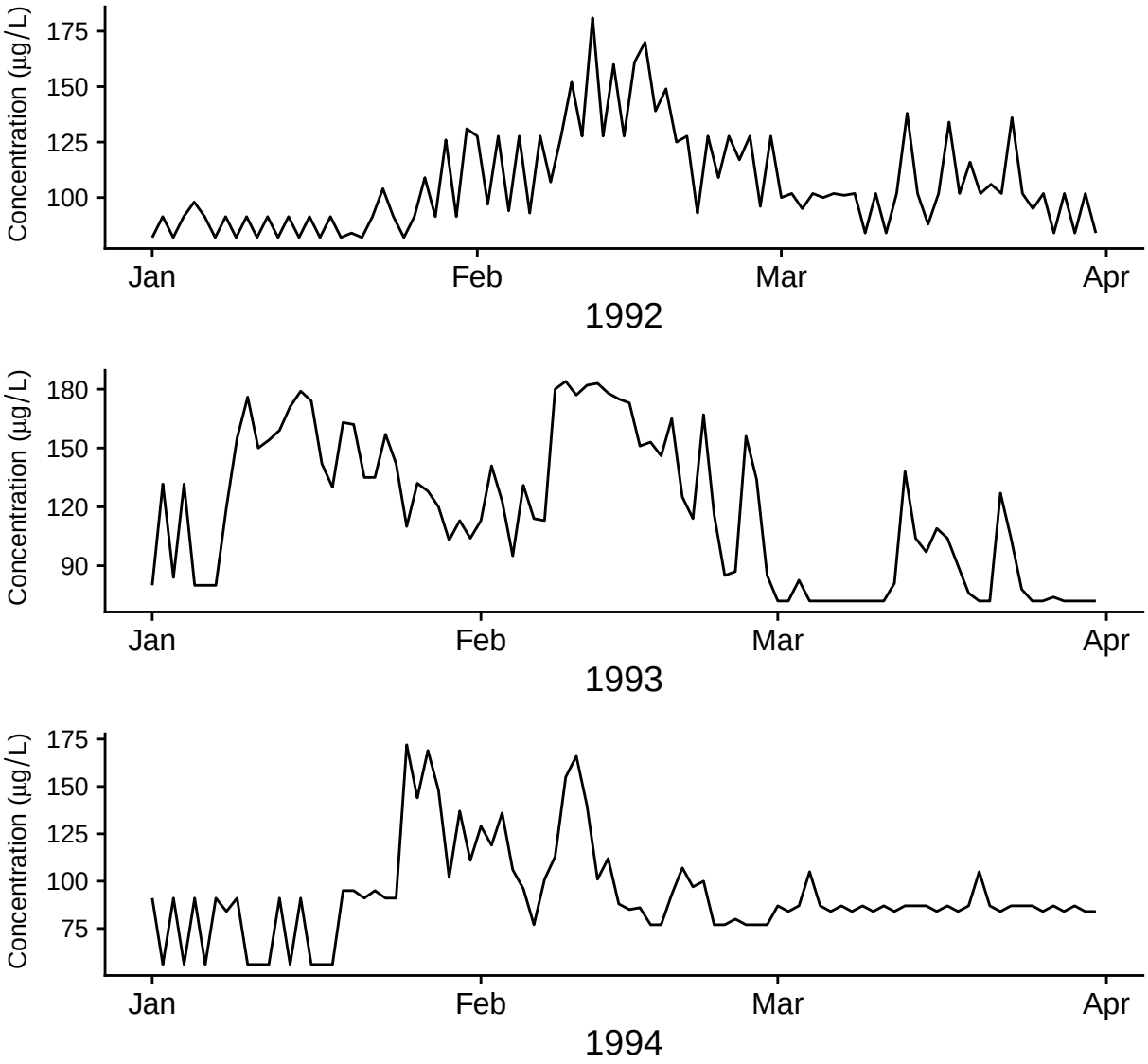
In this section, we will try to implement Damsleth's algorithm of interpolating missing values. Out of 339 different pesticides, diazinon has the largest number of observations. For diazinon, San Joaquin River station near Vernalis (code: 39-17) has the most observations. Therefore, we will try to implement Damsleth's algorithm to interpolate missing values in this dataset. This is the same dataset that Shumway used in his analysis. However, he subsetted the data to contain observations from April 1996 to May 1997 only. This dataset contains 8277 observations from 1991 to 2013.



[Figure 4.1: Plot of diazinon concentrations over twelve months from 1991 to 2013]

We plot the diazinon concentrations over twelve months for all the years and observe that only from 1991 to 2002, we have a good amount of observations. From 2003 to 2013, the numbers of observations in each year are too limited, hence, not enough for us to analyze. Therefore, we subset the dataset to contain only observations from 1991 to 2002. Moreover, January, February and March tend to have the most observations compared to other months in all years. This suggests that diazinon is usually applied during the time near those months.

Out of all the years, 1992 to 1994 have the most observations, so we take a closer look at observations taken in January, February and March in those years to explore if there is any pattern during those periods. In this subset data, there is a large amount of days that have concentrations measured at 1 $\mu g/L$. This value seems to be unreasonable when all other observations are above 50 $\mu g/L$. Hence, we decided to impute days with concentrations as 1 $\mu g/L$ by the lowest concentration of the month. In addition, we also imputed the missing days with the average of the month.

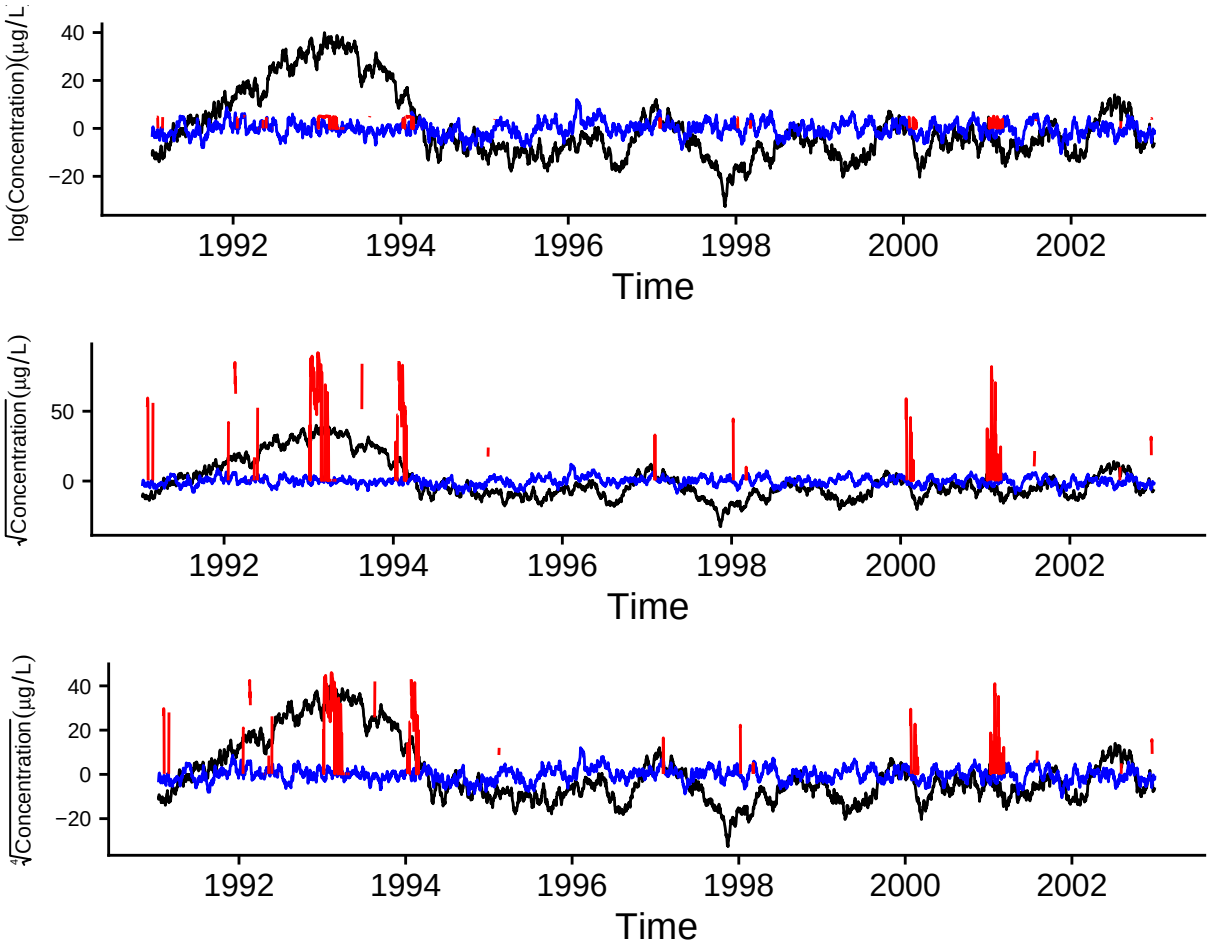


[Figure 4.2: Diazinon concentrations during January, February and March in 1992 to 1994]

Figure 4.2 shows the patterns of diazinon concentration from January to March in 1992, 1993 and 1994. The range of diazinon concentration during those periods in the three years are within 75 - 175 $\mu g/L$. The patterns in those three years are not exactly the same. However, we can see that during those months, diazinon concentration went up twice. For 1992, the concentration went up in mid February and again mid March. For 1993, the concentration went high in mid January and again in mid Februray. In 1993, the spikes in concentration happened closer to each other than the previous years, at the end of January and mid February. We could conclude that farmers apply diazinon pesticide twice at the begining of the year.

4.3.2 Interpolating

Damsleth's interpolating algorithm assumes the model of the time series is known. Hence, first, we will find the most appropriate model for diazinon concentrations by estimating the parameters for AR(1), AR(2), MA(1) and ARMA(1,1) models by least squared method using the available data. For the below plots, I used log, square root and fourth root to transform the monitoring data.



[Figure 4.3: Fitted values of AR(1)-black, AR(2)-blue against the transformations of real data-red]

In Figure 4.3, log transformed diazinon concentration is best modelled with AR(2) and fourth root diazinon concentration is best modelled with AR(1). Even so, the RSS of the two models are still high. The RSS from using AR(1) to model fourth root of diazinon concentration is 348516.7 and the RSS from using AR(2) to model log of diazinon concentration is 12484.82. The RSS reduces significantly with AR(2) model compared to AR(1) model.

5 Future Work

The initial goal of this project is to apply time series analysis to model pesticide concentration in California water system. However, in order to model a time series with ARMA model, at least 80% of data of that series should be present. However, we only have at most 20% of the data. Due to the extensive missingness and sparsity in data, we cannot apply time series analysis directly on the data. One solution for missing value problem is to interpolate the time series with the algorithm provided by Eivind Damsleth. Before we interpolate missing values in a time series, we have to examine the patterns of missing values. However, with the datasets that we have at hand, the missing values occur mostly randomly. Therefore, the interpolation method does not give an accurate result in this project. The missingness and sparsity in data that we face in this project is a complicated problem and need more time to solve. Another proposed approach for future work is to use Bayesian Inference with physical model CalFed predictions as our prior knowledge to build a statistical model. Then we can use the moniotring data to update our model and find the posterior distribution of the pesticide concentrations. Then, we can analyze the posterior distribution and infer the pesticide concentrations. However, this approach requires a good understanding of the physical model before we can build our statistical model.

A Appendices

A.1 R Code

A.1.1 Map of Monitoring Stations

```
library(rgdal)
library(ggplot2)
Imi = read.csv("~/Desktop/DPR/Imidacloprid.csv")
#map for Imidacloprid
mapdat1 = data.frame(Imi$LONGI_NAD8, Imi$LAT_NAD83,
                     Imi$SITE_CODE,Imi$comID)
colnames(mapdat1) = c("Long_NAD83", "Lat_NAD83","site",  "comID")
mapdat1$Lat_NAD83 = as.numeric(as.character(mapdat1$Lat_NAD83))
mapdat1$Long_NAD83 = as.numeric(as.character(mapdat1$Long_NAD83))
mapdat1$site = as.character(mapdat1$site)
dsn = system.file("vectors", package ="rgdal")[1]
cali = readOGR(dsn=dsn, layer = "CA_Counties_TIGER2016")
#need to put shp file in to vector folder of rgdal package
#class(mapdat1) #data.frame
coordinates(mapdat1) = ~Long_NAD83+Lat_NAD83
#class(mapdat1) # "SpatialPoints"
              #attr(,"package")
              #"sp"
#proj4string(mapdat1) #NA
proj4string(mapdat1) = CRS("+proj=longlat +datum=WGS84")
#WGS84 is obtained from .prj file read by texteditor
mapdat1 = spTransform(mapdat1, CRS(proj4string(cali)))
#identical(proj4string(mapdat1), proj4string(cali))#TRUE
mapdat1 = data.frame(mapdat1)
names(mapdat1)[names(mapdat1)=="Long_NAD83"] = "x"
names(mapdat1)[names(mapdat1)=="Lat_NAD83"] = "y"
ggplot()+geom_polygon(data=cali, aes(x=long, y = lat, group =group))+
  geom_point(data = mapdat1, aes(x=x,y=y), color = "red")+
  ggtitle( "Station plot of Imidacloprid")
```

A.1.2 Extracting Catchment Attributes

```
##matching COMID
library(gdata)

#read all the attribute files
filenames = list.files("/Users/huongvu/Desktop/DPR/attributes",
                       pattern="*.csv", full.names=TRUE)
all.file = lapply(filenames, read.csv)

#read in list of Site names
Site = read.xls("~/Desktop/DPR/CentralCoastSiteComID.xlsx",
                header = TRUE, skip = 1, sheet = 1)
Station = unique(Site[,2]) #get unique ComID
#function to extract attributes in ComID order from CA files
extract = function(file) {
  file = file[which(file[,1]%in%Station),]
}
#function order
sort.COMID = function(file){
  print(file)
  file = file[order(file[,1]),]
}

#extract COMID
data.file = lapply(all.file, extract)
#test if all COMIDs are extracted
for (i in 1:50){
  if (nrow(data.file[[i]]) == nrow(data.file[[i+1]]))
    print("TRUE")
}

#ordering data according to COMID
data.file = lapply(data.file,sort.COMID)
#test if sorting is right
for (i in 1:50){
```

```
  result = all(data.file[[i]][,1] == data.file[[i+1]][,1])
  print(result)
}

#combining all tables
big.table = data.file[[1]]
for (i in 2:51){
  big.table = cbind(big.table, data.file[[i]][,-1])
}
write.csv(big.table, "Big Table.csv")
```

References

[1] Damsleth, Eivind. (1980)."Interpolating Missing Values in a Time Series." *Scandinavian Journal of Statistics* Vol. 7, No. 1 (1980), pp. 33-39.

[2] Shumway, Robert H.(2001). *Statistical Approaches to Assessing Pesticide concentrations in the DPR Surface Water Database*.

[3] Hill, Ryan A., Marc H. Weber, Scott G. Leibowitz, Anthony R. Olsen, and Darren J. Thornbrugh, 2015. The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. *Journal of the American Water Resources Association* (JAWRA) 1-9. DOI: 10.1111/1752-1688.12372

[4] Fung, David S. (2006). *Medthods for the Estimation of Missing Values in Time Series*. Edith Cowan Unniversity.

[5] Aue, Alexander. (2010). *Applied Time Series Analysis*. University of California, Davis.

[6] Shumway, Robeert H., Stoffer, David S. (2016). *Time Series Analysis and Its Applications With R Examples*. Switzerland: Springer.

[7] "Mapping in R using the ggplot2 package". ZevRoss. July 16, 2014. zevross.com.