Detection and Analysis of Sludge Bulking Events Using Data Mining and Machine Learning Approach

Yuanhao Zhao
Marquette University

DETECTION AND FORECASTING OF SLUDGE BULKING EVENTS USING DATA MINING AND MACHINE LEARNING APPROACH

by

Yuanhao Zhao, B.E.

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin

May 2012

ABSTRACT
DETECTION AND FORECASTING OF SLUDGE BULKING EVENTS USING DATA MINING AND MACHINE LEARNING APPROACH

Yuanhao Zhao, B.E.

Marquette University, 2012

Sludge bulking is the most notable cause of activated sludge plant failure (i.e. exceeding discharge permit quality limits) worldwide. Numerous mathematical methods have been applied to detect and provide warning for the prevention of sludge bulking. However, these models often fail to reliably forecast sludge bulking events because they focus on the point-by-point "curve-fitting" strategy, while the number of bulking event data points is relatively small in comparison with the large amount of data in the time series. Therefore, three machine learning approaches which focus on detecting the temporal pattern data before the sludge bulking events are considered in this study.

The main objective of this research is to apply machine learning and statistical methods to detect the hidden temporal patterns in the sludge volume index (SVI) data and related water-quality parameters occurring before high SVI values (sludge bulking) occur, and then the hidden temporal patterns can be used to forecast high SVI values in the future. Three methods are applied in this research, the improved Time Series Data Mining (TSDM) method, the Hidden Markov Models (HMMs) method, and the combined method of Hidden Markov Models and multinomial logistic regression (MLR).

The results and analysis show that the improved TSDM method and the HMMs method are capable to detect and predict sludge bulking events. The improved TSDM method can have a sludge bulking event prediction accuracy between 60% and 100%. The HMMs method could provide warning information to the WWTP operators, even if the HMMs method only detects the first state of the pattern leading to sludge bulking. Once the first pattern state was detected, there was high probability (>80% in all cases, mostly > 90%) that sludge bulking would occur. However, both of these methods have limitations because they are new methods applied to the sludge bulking problem. For the combined method, although the results are not useful for the detection of sludge bulking, some wastewater quality parameters are found to have significant impact on the sludge bulking, i.e., sludge retention time (SRT) and effluent pH for all three batteries.

ACKNOWLEDGMENTS

Yuanhao Zhao, B.E.

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I want to thank the Department of Civil and Environmental Engineering to provide funding for my graduate study at Marquette University.

I am deeply indebted to my advisor Dr. Melching, whose help, stimulating suggestions and encouragement helped me in all the time of writing of this thesis. I also want to thank Dr. Feng who helped me during my research. I would like to address my thanks to Dr. Crandall for his help during my graduate study and valuable suggestions to this thesis.

I have furthermore to thank Dr. Bansal who gave me lots of advice and encourage me go ahead with my research and thesis. I also want to thank Dr. Heng Zhang who helped me and gave me the data for this research.

Especially, I want to give my special thanks to my parents whose love enabled me to complete this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INDTRODUCTION

1.1. Sludge Bulking and Sludge Volume Index (SVI)

The activated sludge process is the most commonly used process in the treatment of municipal and industrial wastewater. In the process, air (or pure oxygen) is passed through a mixture of sewage and recycled sludge (known as activated sludge) to allow micro-organisms to break down the organic components of the sewage in an aeration tank. The effluent from the aeration tank is continually drawn off as new sewage enters the tank. This effluent is known as mixed liquor because it is a mixture of wastewater and activated sludge that has grown in the aeration tank during the consumption of organic waste. The activated sludge in the mixed liquor must then be settled in a sedimentation tanks so that the supernatant clear water can be separated from the sludge to pass on to further stages of treatment. The general activated sludge process in a wastewater treatment plant is shown in Figure 1.1.

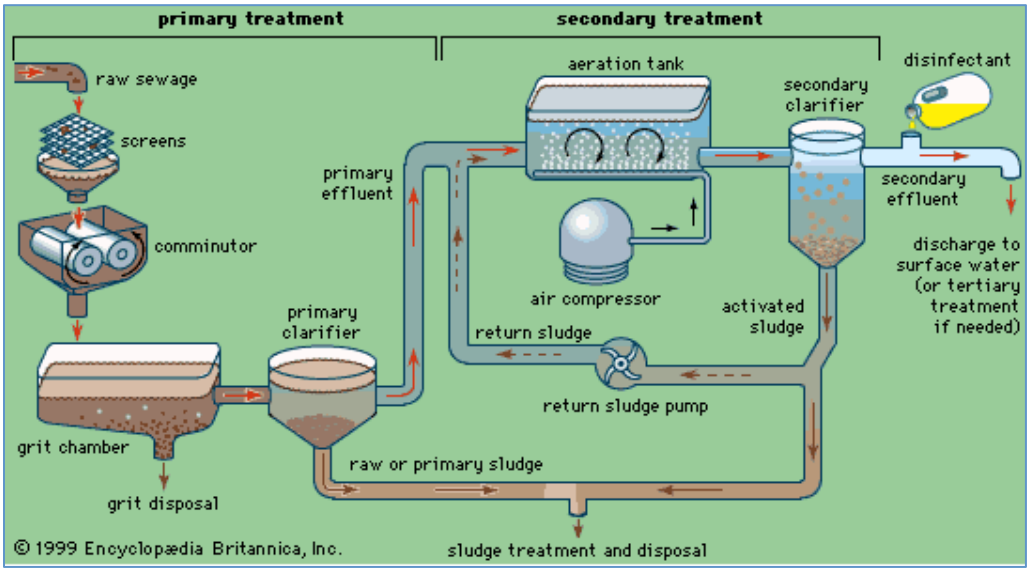

Figure 1.1: General activated sludge process (after Encyclopadia Britannica, Inc., 2012.)

Sludge bulking (Sezgin et al., 1978) occurs when the sludge fails to separate out in the sedimentation tanks (secondary clarifier in Fig. 1.1), i.e. the sludge has poor settling characteristics. Bulking is the term used to describe activated sludge that settles slowly and compacts poorly. The sludge bulking problem was discovered more than seventy years ago, and it is the most notable cause of activated sludge plant failure (i.e. exceeding discharge permit quality limits) worldwide (Madoni et al., 2000). Not only does the sludge bulking incur heavy penalties due to noncompliance with discharge permits, but it also results in severe, poor quality of discharged treated wastewater effluent, as well as the expensive cost of methods to remedy the bulking problem, e.g., addition of chemicals like inorganic coagulants and flocculants such as ferric chloride and alum, and installing additional aeration capacity. Meanwhile, it also compacts poorly; after thickening, a unit weight of bulking sludge occupies a larger volume than an equivalent weight of normal sludge (Pipes, 1979).

There are two types of bulking problem. One is nonfilamentous bulking which is caused by excess production of exopolysaccharides by bacteria. However, this type of bulking is rare and is corrected by chlorination (Bitton, 2005). The other one is filamentous bulking which is the most common form of sludge bulking. The main cause of sludge bulking is the growth of filamentous bacteria. Activated sludge flocs are made up of biological and nonbiological components. The biological component consists of a wide variety of bacteria, fungi, and some metazoans. The nonbiological component is made up of inorganic and organic particulates (Jenkins et al., 2004). Filamentous microorganisms grow in long strands that have much greater volume and surface area than conventional floc and are very slow to settle. As filamentous bacteria grow in the

sludge, the sludge settles less and less because the filamentous bacteria do not compress well.

The Sludge Volume Index (SVI) (Forster, 1971) is an empirical measurement used to characterize the sludge bulking problem. If sludge bulking occurs, the wastewater treatment process can generate a high SVI value and very turbid supernatant (i.e. effluent from the sedimentation tank with high suspended solids). However, the definition of "High SVI" is different for different wastewater treatment plants (WWTPs) and different research works. Some WWTPs claim that sludge bulking occurs when SVI is larger than 100 mL/g (Soyupak, 1989). Some different values for SVI are 150, 180, even 200 mL/g (Rensink, 1974). In this thesis, the SVI value representing sludge bulking is set to 120 mL/g (lower value) and 150 mL/g (higher value) depending on different situations and analysis methods applied.

1.2. Objectives

Sludge bulking is an unusually complex process caused by a variety of variables, including wastewater characteristics, design limitations, and operational issues. There is no scientifically robust evidence to reveal the detailed causes of sludge bulking problems, or a reliable method to forecast the occurrence of sludge bulking.

Numerous methods have been applied to detect and prevent sludge bulking. The most widely and reliable used method to detect filamentous bacteria which leads to sludge bulking is the Microscopic Examination Methods (Jenkins et al., 2004), which use a microscope, to observe the quantity and categories of the filamentous organisms. However, this method is costly and it cannot prevent sludge bulking effectively due to the long time needed for the identification process for the different kinds of filamentous

bacteria. Some researchers have recommended some operational regulations for the WWTPs, i.e. restrain the organic loading and maintain the dissolved oxygen concentration, to try to reduce the likelihood of sludge bulking. Although such recommendations could be useful for the operation of WWTPs, they cannot detect or predict the occurrence of sludge bulking problems.

Over the years, researchers have applied numerous mathematical modeling approaches, such as various biological models, time series analysis, and artificial neural networks (Capodaglio et al., 1991), trying to analyze and model the SVI data to detect sludge bulking problems. However, these models often fail to reliably forecast sludge bulking events. The reason for the poor forecasting performance of these methods is the central focus of these methods is always on the general statistical characteristics of the entire data set, e.g., the point-by-point "curve-fitting" strategy,

The main objective of this thesis is to apply machine learning and statistical methods to detect the hidden temporal patterns in the SVI data and related water-quality parameters before high SVI values (sludge bulking) occur, and then use the hidden temporal patterns to forecast high SVI values in the future.

Three machine learning methods are applied in this thesis, the improved Time Series Data Mining (TSDM) method, the Hidden Markov Models (HMMs) method, and the combined method of Hidden Markov Models and Multinomial Logistic Regression Model. For the TSDM method, an 8 year ammonia time series data is tested first, then the SVI data are tested to detect the temporal patterns and sludge bulking events. For HMMs method, only the SVI data are tested to detect the possibility of temporal pattern states and the event state for each SVI point. For the combined method, a multinomial

logistic regression model is applied to model the pattern states and event state from the HMM method with the SVI and other physical and chemical variables data.

The significance of this research is all three methods focus on the detection of temporal patterns before the sludge bulking event instead of on the point-to-point time series prediction. Once a predictive pattern is detected, no matter of the depth of our understanding and the validity of the definition of sludge bulking, future events could be predicted faster than by the previous methods. Also, in the analyses of the results of the combined method, some variables in the wastewater treatment process are revealed to have a significant correlation with sludge bulking problems.

1.3. Data Collection (Period I-Ammonia Test, Period II-SVI Test)

All data were collected from the North Side Water Reclamation Plant (NSWRP) of the Metropolitan Water Reclamation District of Greater Chicago (MWRDGC). There are four treatment batteries at the NSWRP, batteries A, B, C, and D. Wastewater treatment plant data were collected daily from influent and effluent for each battery.

1.3.1. Period I – Ammonia Test

Prior to October 2010, the MWRDGC provided the outflow (effluent) data of the NSWRP from 2001 to 2008, which includes flow, temperature, Biochemical Oxygen Demand (BOD), Dissolved Oxygen (DO), and ammonia concentrations (NH_3). During this time, the ammonia concentration was made the object of testing and research. According to the permit limit for the effluent of the NSWRP, the ammonia concentration should not be higher than 2.5 mg/L. Similar to the definition of the high SVI value for the sludge bulking problem, the event value (high ammonia concentration value) was set

to 2 mg/L in the analysis here.

1.3.2. Period II – Sludge Volume Index (SVI) Test

In October 2010, more detailed data on the effluent from the NSWRP, including SVI values, were obtained. The detailed data included values for the 4 different treatment batteries (A, B, C, and D) from 2002 to 2009. The different SVI values of those 4 treatment batteries then were used as the detection data set. Also, in order to discover and detect the relationships and hidden patterns between other variables and the SVI data, some variables were selected for analysis as listed in Table 1.

Table 1.1: Variables selected from NSWRP battery data for evaluation of the relation to sludge bulking (i.e. high SVI values)

Abbreviation	Description
DO	Dissolved Oxygen
Temperature	Water Temperature
Flow	Wastewater Influent Flow Rate
F/M	Food to Microorganisms Ratio
RSSS	Returned Sludge Suspended Solids
MLSS	Mixed Liquor Suspended Solids
MLVSS	Mixed Liquor Volatile Suspended Solids
NH3	Ammonia
BOD	Biochemical Oxygen Demand
COD	Chemical Oxygen Demand
pH	pH value for effluent
SRT	Sludge Retention Time

1.4. Scope of Work

Three methods are applied in this thesis, the improved Time Series Data Mining

(TSDM) method, the improved Hidden Markov Models (HMMs) method, and the combined method of Hidden Markov Models and Multinomial Logistic Regression Model. For the TSDM method, the ammonia data and the SVI data were studied to detect the hidden temporal patterns. The HMMs method was applied to the SVI data alone. The combined method used the SVI data and the data on other water quality and WWTP operation variables.

1.4.1. Improved Time Series Data Mining (TSDM) Method

The improved Time Series Data Mining (TSDM) method was originally introduced by Mr. Hai Huang and Dr. Xin Feng at Marquette University (Feng and Huang, 2005). The TSDM method focuses on predicting events by looking for the temporal patterns before the events happen in the time series. The core of the TSDM method is identification of data clusters and optimization of the temporal patterns. The method uses two-step optimization algorithms to find the temporal pattern clusters. Once the temporal pattern clusters are found, the training step is complete. Then the computer program will embed the test data into the phase space. Once the data points in the reconstructed phase space fall into the clusters which contain the temporal patterns, the computer program will consider those data points as the patterns that can be used to forecast events.

1.4.2. Improved Hidden Markov Models Method (HMMs)

Hidden Markov Models (Rabiner, 1989) are statistical Markov models in which the system being modeled is assumed to be a Markov process with unobserved (hidden)

states. HMMs are especially known for their application in temporal pattern recognition. The improved method was generated by Dr. Bansal and Mr. Wei at Marquette University and uses a Mixture of Gaussian function and an Expectation-Maximization (EM) algorithm to normalize the training data to obtain the threshold and transition probabilities for normal state (normal data) and abnormal states (pattern and event data), then the probabilities of each state are calculated by a Mixture-Gaussian probability model. Once the threshold and the state probabilities for the training data are calculated by the HMM program, the program uses the Viterbi Algorithm to predict the probabilities of the hidden states for the test data.

1.4.3. The Combined Method of Hidden Markov Models and Multinomial Logistic Regression Model (MLRM)

A multinomial logistic regression (Combs-Orme, 2009) model is a regression model, which generalizes logistic regression by allowing more than two discrete outcomes. It is a model used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables. The combined method is an attempt to apply a multinomial logistic regression model to use other wastewater parameters to predict the probability of the states of the SVI data in the HMMs method. In the training part, the SVI states data in the HMM method and the selected wastewater parameters are used to build the multinomial logistic regression model. In the testing part, the multinomial logistic regression model is used to perform the SVI state forecasting.

CHAPTER 2 LETERATURE REVIEW ON SLUDGE BULKING STUDY

2.1. The Problem of Sludge Bulking

The sludge bulking problem has been noticed since the application of the activated sludge process in wastewater treatment in the 1920's. Heukelekian (1941) proposed the definition of "sludge bulking," which is a "disease" of the sludge developed in the course of purification of sewage under unfavorable environmental conditions. Heukelekian (1941) also concluded some factors which could cause sludge bulking problems, including inadequate supply of oxygen, organisms involved in the activated sludge, and high concentration of food material.

2.2. Causes of Sludge Bulking

A large number of studies have been done on the relationship between sludge bulking and other variables in the activated sludge process. Filamentous organisms have been known to cause sludge bulking in the activated sludge process, but there are many different kinds of organisms and each is sensitive to different environmental conditions (Jenkins et al., 2004). Oxygen deficiency has been proposed as primarily responsible for sludge bulking (Bhatla, 1967). The influence of pH and organic loading on filamentous bulking (sludge bulking) was investigated in 1970's (Yasuda, 1976), and it had been found that a pH rage of 6-9 stimulates the growth of filamentous organism and cause the sludge bulking. Furthermore, Kappeler and Gujer (1994) suggested that operating conditions and reactor design should be optimized in order to obtain better performance to avoid the possibility of sludge bulking. Some key wastewater and process parameters that can be monitored have been considered to be related to sludge bulking, including

flow rate, pH, temperature, nutrient content, dissolved oxygen (DO), food to microorganism (F/M) ratio, and soluble biochemical oxygen demand (BOD) (Metcalf & Eddy, 2003). Although the sludge bulking problem has been studied and investigated for many years, there is no robust scientific evidence and theory to explain the process, principles, and causes of sludge bulking.

2.3. Non-mathematical Approaches

Due to the hazards of sludge bulking to the operations of wastewater treatment plants, the methods of detection and prevention of sludge bulking are important for WWTPs. Rensink (1974) recommended that WWTPs could restrain the organic loading (≤ 300 g BOD5/day/kg MLSS) to avoid high SVI values which are the sign of possible sludge bulking. As previously mentioned, low DO concentrations could cause sludge bulking more easily. It has been established in European WWTPs operation that the aeration tanks should be designed for and operated with a minimum DO concentration of 2 mg/L (Chudoba, 1985). The microscopic examination methods (Jenkins et al., 2004) are the most widely used techniques for identification of filamentous bulking organisms, including the Total Extended Filament Length (TEFL) Measurement Method, the Simplified Filament Counting Technique, the Nocardioform Organism Filament Counting Technique, etc. Microscopic examination methods require a microscope to observe the quantity and categories of the filamentous organisms that could lead to sludge bulking problems. Due to the large amount of different species of filamentous organisms (nearly thousands), such methods cannot exactly and immediately detect the filamentous organisms. Meanwhile, most of the filamentous organisms are still very poorly characterized, mainly due to the problems of cultivation and maintenance of

cultures (Martins et al., 2004). Plus, the microscopic examination methods need to be done day by day which is costly and time-consuming for the wastewater treatment plant operators.

2.4. Mathematical Approaches

Scientists, engineers, and researchers have tried to apply numerous mathematical methods and computer tools to help improve wastewater treatment operations to prevent sludge bulking problems. A prototype of computer-based design was developed by Kao et al (1983) to facilitate wastewater treatment plant operations. But the design obtained from the method of Kao et al. (1983) needs to be more robust and complete. The Activated Sludge Model No. 1 (ASM1) (Henze et al., 1987), is a major reference for design and operation of wastewater treatment plants. ASM1 was improved as the Activated Sludge Model No. 2 (ASM2) in 1995 (Henze et al., 1995), and the Activated Sludge Model No. 3 (ASM3) in 1999 (Gujer et al., 1999). All three activated sludge models were more focused on the biochemical parameters rather than the sludge related variables. Besides, the ASMs do not cover all aspects of activated sludge systems, particularly impacts of different operational scenarios on the activated sludge microbial community, activated sludge settling problem (sludge bulking), etc. (Sin et al., 2006)

Some systems were proposed to help wastewater treatment plant operators to diagnose sludge bulking. Hiraoka et al. (1988) developed a computer-based filamentous microorganisms identification support system. It is an expert system, which was assumed to be applied by a field operator or expert who already has expert knowledge on filamentous organisms. If people who do not have expert knowledge on filamentous organisms use this support system, it can only provide 50% chance of getting the correct

answer. Meanwhile, it is still unclear how many kinds of filamentous organisms could cause sludge bulking, and only twenty-four types of filamentous organisms were considered by Hiraoka et al. (1988).

Martinez et al. (2006) developed a Decision Support System (DSS) that used chemical oxygen demand (COD) and phosphorus (P) measurements to develop control strategies for sludge bulking. It only focused on the non-filamentous bulking which is caused by phosphorus deficiency (NFBPD), so it cannot be applied to filamentous bulking (sludge bulking). Figure 2.1 demonstrates the control strategy.

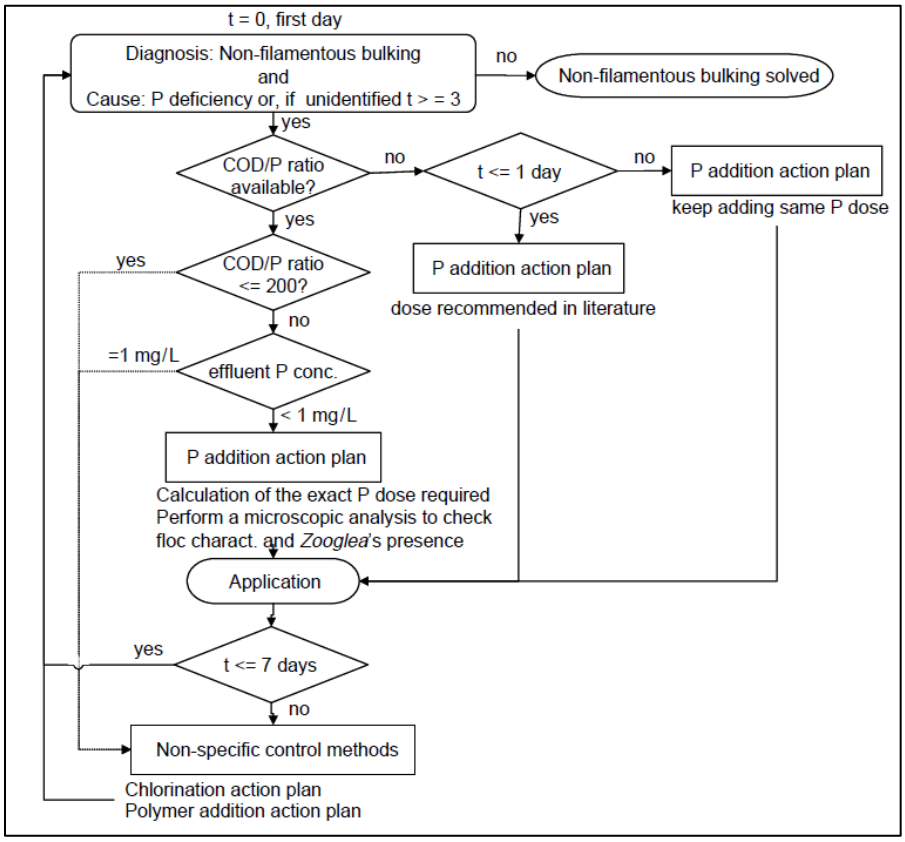

Figure 2.1: Flow diagram followed by the dynamic DSS to design the control strategy for NFBPD (after Martinez et al., 2006)

In the case analyzed by Martinez et al. (2006), the DSS provided advice for solution of non-filamentous bulking by adding P chemicals to the treatment process.

However, after three days of continuous sludge bulking problems, the DSS made a conclusion that it faced an unidentified problem, which showed the DSS has a limitation when applied to actual operation.

Chan and Koe (1991) developed a prototype expert system for diagnosing the sludge bulking problem, but their paper focused more on the expert system architecture than the sludge bulking problem. Besides, this expert system proposed by Chan and Koe (1991) contains 80 diagnosis rules.

Chen and Beck (1993) described the development of a multi-species model of the activated sludge process, its application to the assessment of various operational strategies for the control of bulking, and its simplification for incorporation into an on-line estimation scheme using a Kalman filter. This on-line estimation scheme would be the first step in the development of an expert system.

Ng et al. (2000) proposed the development of an expert system for sludge bulking control. They generated a non-linear regression model, which used COD, MLVSS, pH, and the F/M ratio as the main factors. A 15-day prediction example of the SVI data is shown in Figure 2.2 with Pearson's correlation coefficient (R^2) equal to 0.96. However, the average error of the example was 31 mL/g comprising 15% of the average SVI value of 208 mL/g. From the figure showing prediction results, it can be seen that the results failed to reflect the sudden arise of SVI (sludge bulking) in the first 4 days. Also, they noted that unfortunately, for bulking control, the conventional approach to knowledge-based expert system design is not easy.

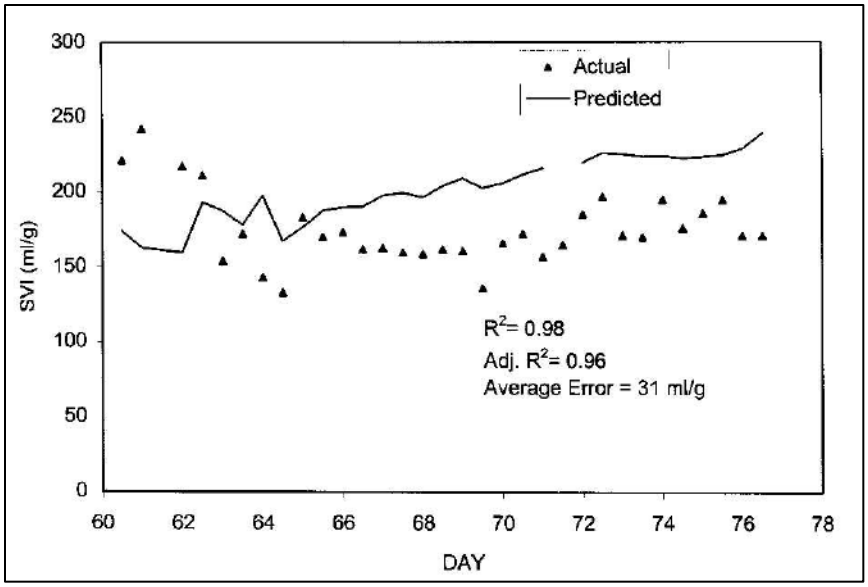

Figure 2.2: Prediction result of the expert system of Ng et al. (2000) over 15 days

Bayo et al. (2006) developed logistic regression models for the occurrence of bulking (defined by an SVI > 150 mL/g) and identified two statistically significant variables that appeared to be important to the occurrence of a higher SVI: season (a surrogate for temperature) and pH (under 7.5). In the logistic regression, all the considered independent variables (pH, conductivity, temperature, season, settleable solids, total solids, COD, and BOD) were subdivided into groups (ranges).

Belanche et al. (2000) developed a soft-computing time-delay method to predict sludge bulking, which applied heterogeneous neural networks (HNN), classical neural networks, probabilistic networks (PNN), and the k-nearest neighbors (KNN) algorithm. They found that a two-day delay was better than a one-day delay. The method had the classification ability of 70% - 73% and a prediction ability of 73%. They made the conclusion that the poor performance of the method can be attributed almost entirely to the chaotic data.

Capodaglio et al. (1991) applied autoregressive, moving average (ARMA) models to the SVI data, autoregressive transfer function (ARTF) models to relate SVI as a

function of the F/M ratio, and artificial neural network (ANN) models to relate SVI to time series of a number of parameters—BOD/N ratio, N/P ratio, mixed liquor temperature, mixed liquor DO, and F/M ratio. From the 20-day prediction results of the paper shown in Figure 2.3, it can be seen that the results only showed the ability of following the trend of the SVI data line instead of sludge bulking prediction ability. From Figure 2.3, for the sludge bulking event, which is defined as SVI > 150 mL/g, there were 10 events of sludge bulking: 2 events were predicted by ARTF, 2 events were predicted by ARMA, and 4 events were predicted by ANN. The best prediction ability model was obtained by ANN at only 40%.

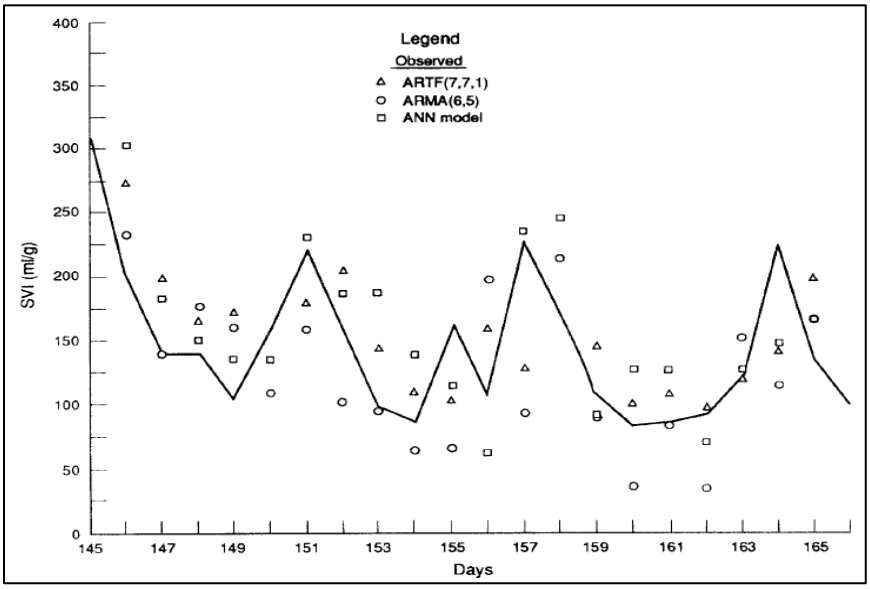

Figure 2.3: Observed and predicted SVI values during a 20-day test period (after Capodaglio et al., 1991)

From this review of previous research, non-mathematical methods provide the strategies to prevent sludge bulking and methods to diagnose different kinds of filamentous bulking bacteria. But non-mathematical methods cannot detect and prevent the occurrence of sludge bulking. The foregoing mathematical approaches tried to focus on the prediction of the sludge bulking problem. However, these approaches are based

on point-by-point prediction. Although they demonstrate the ability to follow the trend of the SVI data, the forecasting ability of sludge bulking events of these approaches is poor. Further, the expert system approach needs expert knowledge of the sludge bulking problem, and it is restricted to the actual application to a specific wastewater treatment plant. In this thesis, three machine learning approaches are applied to detect and predict sludge bulking problems. These machine learning approaches are 'Black Box Methods,' which only consider the input and output. So they do not require expert knowledge of the sludge bulking problem. Furthermore, these machine learning approaches focus on the detection of the patterns before the sludge bulking events, and they can save time and cost for the WWTP operation.

CHAPTER 3 MACHINE-LEARNING APPROACHES APPLIED IN THESIS

3.1. Machine-Learning Approach

The machine learning approach (Bishop, 2006), a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. The machine learning approach is a combination of mathematics and computer science. It studies and proposes algorithms that allow computer systems analyzing a data-related problem to improve automatically through experience, i.e. from training data (Grangier, 2008). In this thesis, three machine learning approaches are applied to the study of sludge bulking, the improved Time Series Data Mining, Hidden Markov Models, and a Multinomial Logistic Regression Model. These approaches learn from the training data set, and then are applied to detect the patterns in testing data and forecast possible future events. For example, the Time Series Data Mining method is trained by the SVI data from 2002 to 2006, and the method learns the information on the causes of high SVI values from the training data set. Using the information gained from the learning process, the method can detect and predict the future high SVI values in testing data set of 2007.

3.2. Improved Time Series Data Mining (TSDM) Model

3.2.1. Introduction to Time Series Data Mining

A time series is a sequence of data points. Time series analysis is widely used in signal processing, econometrics, and mathematical finance. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistical and other characteristics of the data. Some models are developed for time series

forecasting to predict future values based on previously observed values.

The framework of Time Series Data Mining (TSDM) (Povinelli and Feng, 2003) overcomes limitations (i.e. stationary and linearity requirements) of traditional time series analysis (i.e. Autoregressive Integrated Moving Average Model) techniques by adapting data mining concepts for analyzing time series. The TSDM framework focuses on predicting events, which are important occurrences within the time series (e.g., the high SVI values characteristic of sludge bulking). Consider a time series:

$$\{x(t), t = 0,1, \dots, n\} \tag{3.1}$$

where t is the time index and n is the total number of observations. An event is defined by an event characteristic function:

$$g(t) = g\left(x_t, x_{t-1}, x_{t-2}, \dots, x_1\right) > C \tag{3.2}$$

where C > 0 is a given constant, and g(t) is the event characteristic function. For the sludge bulking problem, the event is considered as a SVI value larger than 150 mL/g, defined as:

$$g(t) = x(t) - 150.0 > 0 \tag{3.3}$$

which means that an SVI value larger than 150 mL/g is defined as an event. A temporal pattern is a time-ordered, fixed structure in the sequence data. It occurs repeatedly and is closely correlated with the occurrences of critical events on the observed date, as shown in Figure 3.1. Figure 3.1 shows the example of the conception of hidden temporal patterns and events. Figure 3.1 illustrates the close correlation between the temporal patterns and the event. The left graph displays a section of an individual time series with a 5-dimensional (5 data points) temporal pattern that is repeated three times. The figure shows the occurrences of the event following the temporal pattern. The right portion shows the similarity of the three 5D temporal patterns occurring before the events.

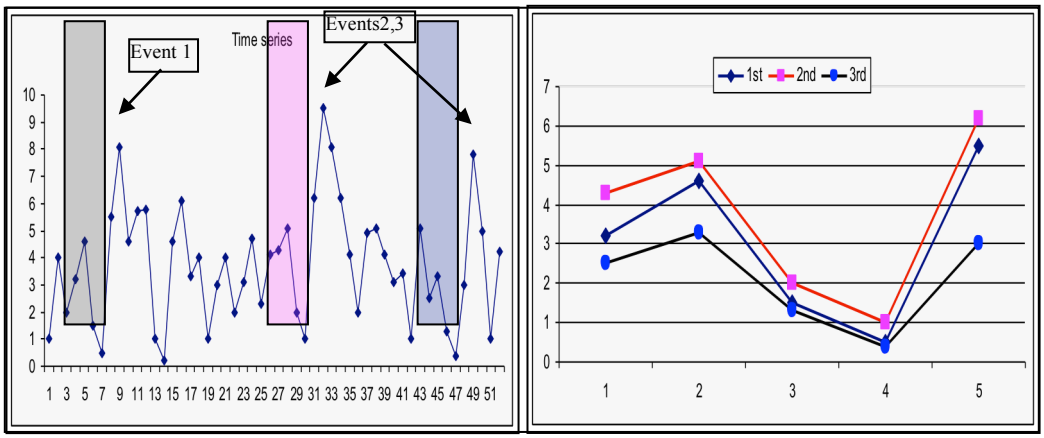

Figure 3.1: Example of Temporal Patterns and Events

The improved Time Series Data Mining method was developed by Huang (2001) base on the work of Povinelli (1999) also reported in Povinelli and Feng (2003). The improved TSDM method proposed a definition of a fuzzy set cluster which applied a Gaussian membership function to prevent the noisy data points from being included in the cluster. Also, the improved TSDM method proposed a two-step optimization algorithm for data mining, which is discussed in section 3.2.2.3. The two-step optimization algorithm can increase the efficiency of computing compared with the genetic algorithm used in the work of Povinelli and Feng. (2003).

3.2.2. Main Components of the Improved TSDM method

The improved Time Series Data Mining method has four main components : definition of the event threshold value (sludge bulking), determination of the phase space time-delay embedding dimension - Q, data mining and optimization (finding temporal pattern clusters in the training data set), and determination of the enlarge ratio of the temporal pattern cluster radius.

I. Sludge Bulking Event Threshold Value

As previously mentioned, the sludge bulking event is set by the case of the SVI larger than 150 mL/g. However, if the computer program fails to detect the hidden temporal patterns in the training data set, reduction of the event value to less than 150 mL/g (e.g., 120 mL/g) should be considered.

II. Time-Delay Embedding

The time series data are transformed into a multi-dimensional Reconstructed Phase Space (RPS) (Montgomery et al., 2008) denoted by R^Q, to represent the underlying dynamics, according to:

$$Y_j = (x_j, x_{j+\tau}, x_{j+2\tau}, \dots, x_{j+(Q-1)\tau}) \quad (3.4)$$

where j = 1, 2, … , n-(Q-1)τ, Q is the dimension of the vector Y_j, called the embedding dimension, and τ is a delay time (or time delay). The embedding dimension, Q, can be calculated by the false nearest neighbor method (Kantz and Schreiber, 2004). For the time delay, τ, common sense should be used to choose τ, such as 1, 2, and 3.

The false nearest neighbor procedure is a method to obtain the optimum embedding dimension for phase space reconstruction. By checking the neighborhood of points embedded in projection manifolds of increasing dimension, the algorithm eliminates 'false neighbors': This means that points apparently lying close together due to projection are separated in higher embedding dimensions. A natural criterion for catching embedding errors is that the increase in distance between two neighbored points is large when going from dimension d to d(Q)+1. This criterion is stated by designating as a false nearest neighbor any neighbor for which the following is valid:

$$\left[\frac{R_{d+1}^2(t,r) - R_d^2(t,r)}{R_d^2(t,r)}\right]^{1/2} = \frac{|x(t+\tau) - x(t_r+\tau)|}{R_d(t,r)} > R_{tol} \quad (3.5)$$

Where t and t_r are the times corresponding to the neighbor and the reference point, respectively; R_d denotes the distance in phase space with embedding dimension *d (Q)*, and R_{tol} is the tolerance threshold. These thresholds can be determined by the false nearest neighbor algorithm.

In some cases, the calculated Q may not the best Q for prediction. In such a case, the analyst should try other embedding dimensions. This transformation will make it possible to apply clustering and optimization algorithms to detect the significant temporal pattern vectors. Takens (1981) showed that if *Q* is large enough, the phase space is homeomorphic to the state space that generated the time series.

III. Data Mining and Optimization

The core of the TSDM method is identification of data clusters and optimization of the temporal patterns. The method uses two-step optimization algorithms: pre-searching and gradient-based searching. The first step, the pre-searching step uses the subtractive clustering method (Chiu, 1994), in which data points in the phase space are considered as the candidates for cluster centers. Then the second step, the gradient-based searching algorithm uses gradient-based searching algorithms (Snyman, 2005) to further optimize the temporal pattern clusters obtained from the first step. A clustering example is shown in Figure 3.2. The circles are the clusters found by the improved TSDM method, and the blue points are the temporal patterns.

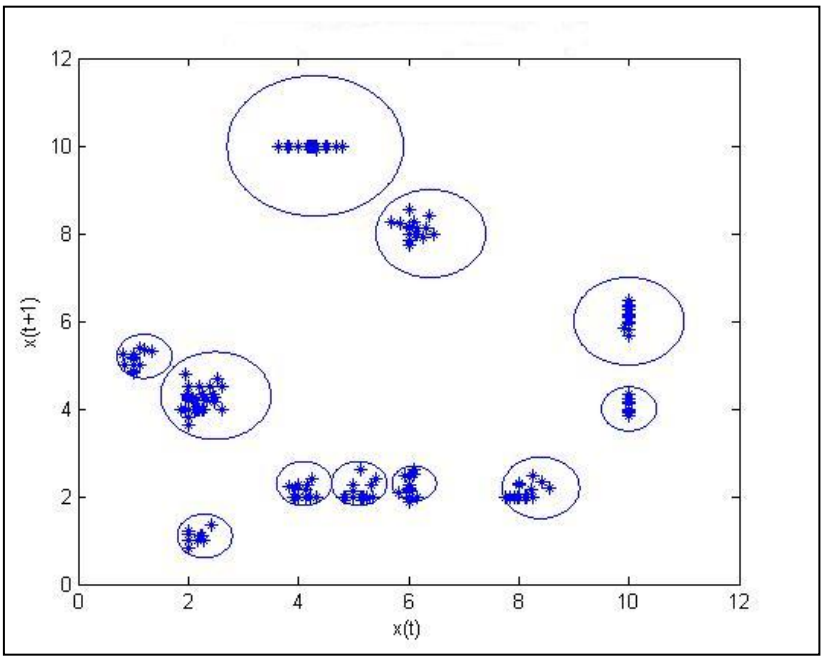

Figure 3.2: Example of clustering on a phase space with τ=1 and Q=2

IV. Radius Enlarge Ratio for Temporal Pattern Clusters

Once the temporal pattern clusters are found, the training step is complete. Then the computer program will embed the test data into the reconstructed phase space. Once the data points in the reconstructed phase space fall into the clusters which contain the temporal patterns, the computer program will consider those data points as the temporal patterns which can be used to predict events. The cluster radius enlarge ratio is used to magnify the radius of the temporal pattern cluster because, in the reconstructed phase space, sometimes the temporal pattern points may not be in the cluster but near to the cluster. In such cases, the radius needs to be enlarged to contain those points. However, magnifying the radius will lead the temporal pattern clusters to contain points which are not part of the temporal patterns. So the radius enlarge ratio should be chosen carefully. Normally, it is set to between 1 and 2.

3.2.3. Process of the Improved Time Series Data Mining Model

Figure 3.3 summarizes the process of the improved TSDM method.

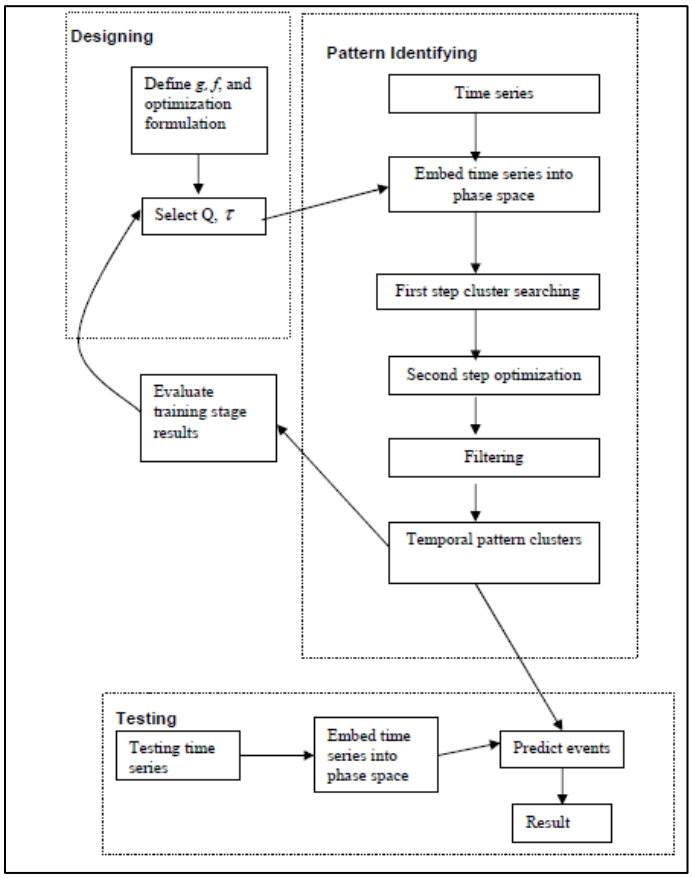

Figure 3.3: Process of the improved TSDM method (after Huang, 2001)

3.3. Improved Hidden Markov Models (HMMs)

3.3.1. Introduction to Hidden Markov Models (HMMs)

Hidden Markov Models were first introduced by Baum and Petrie (1966). One of the first and most widely used applications of HMMs is in speech recognition (Huang et al., 1990). HMMs are finite models that describe a probability distribution over an infinite number of possible sequences. The improved HMMs applied in this thesis were developed by Dr. Bansal and his students at Marquette University, focusing on modeling

temporal patterns and event detection.

It is assumed that the probability distribution depends on a hidden sequence of states. Suppose the hidden sequence is $\{Z_t, t = 1,2, \dots, m\}$, where $Z_t \in S$ $= \{s_0, s_1, \dots, s_\partial, s_{\partial+1}\}$, a set of all possible states. State s_0 is set as the normal state, states $\{s_1, s_2, \dots, s_\partial\}$ are set as pattern states, and $s_{\partial+1}$ is set as the event state. Assume that $\{Z_t, t = 1,2, \dots.\}$ follows a Markov model with the transition probability matrix:

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} & \text{K} & p_{0\partial+1} \\ p_{10} & p_{12} & \text{K} & p_{1\partial+1} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ p_{\partial+10} & p_{\partial+11} & \text{K} & p_{\partial+1\partial+1} \end{bmatrix} \tag{3.6}$$

where $p_{ij} = P\{Z_t = s_j \mid Z_{t-1} = s_i\}$

To structure the temporal pattern and the event occurrence, the state space set S is partitioned as $S = \{s_0, \boldsymbol{s}_1, s_{\partial+1}\}$, where $\boldsymbol{s}_1 = (s_1, s_2, \dots, s_\partial\}$ is a sequence of the pattern states. In order to reflect the transition of a process from the normal state to the development of a pattern state and then to the event state, the transition matrix $\boldsymbol{P}$ must take a special form:

$$\boldsymbol{P} = \begin{bmatrix} 1-p & p & 0 & \cdots & 0 \\ 1-q_1 & 0 & q_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1-q_\partial & 0 & \cdots & 0 & q_\partial \\ 1-r & 0 & \cdots & 0 & r \end{bmatrix} \tag{3.7}$$

Note that $p = P(Z_t = s_1 | Z_{t-1} = s_0)$ is the probability that at a given time t the state would jump from the normal state to the beginning of the pattern state, but the probability is 0 that it would jump to the second or a higher pattern state. Once the state is in a pattern state, it moves to the next pattern state in a stepwise manner or it goes back to the normal state without completing the pattern. At the last pattern state, when the

process completes the pattern, it reaches the event state with probability $q = P\ (Z_t = s_{\partial+1} | Z_{t-1} = s_{\partial})$. It is also assumed that once the process reaches the event state, it remains in the event state with probability r or it goes back to the normal state. The main point here is that the event occurs only when the pattern is completed, and if the pattern breaks down, the process goes back to the normal state before the event occurs.

There are three main functions and algorithms in the improved HMMs. For the temporal pattern detection process, the Mixture of Gaussian function and Expectation-Maximization (EM) algorithm are performed. For the prediction process, the Viterbi Algorithm is applied to predict the possibilities of the states.

3.3.2. Main Functions in Hidden Markov Models (HMMs) and HMMs process

It is assumed that the time series data follow the Gaussian model distribution. For the training process, the Mixture of Gaussian function and EM algorithm are used to estimate the initial value of the training data, i.e. threshold for each state (normal, pattern, and event states) and transition probability matrix. The computer software can obtain and learn the hidden patterns possibilities in training data set. For the testing process, the Viterbi Algorithm is used to detect and predict the most probable state through the HMMs method. The general process is shown in Figure 3.4.

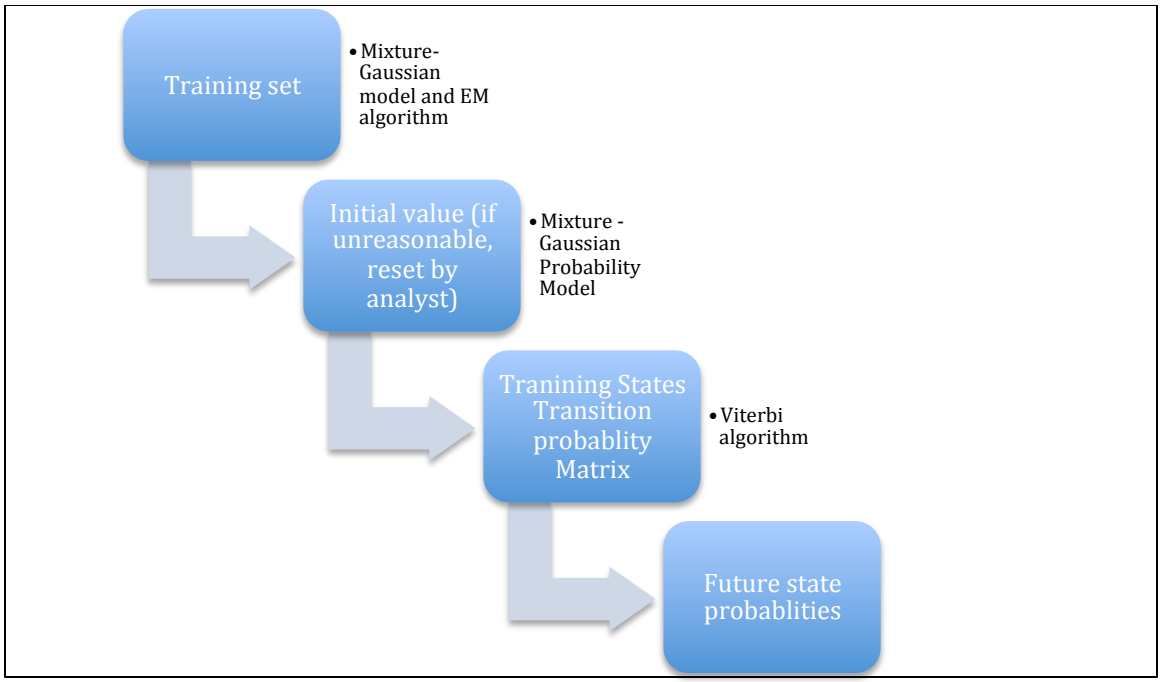

Figure 3.4: General Process of Improved HMMs

3.3.3. Notable Parameters in the Improved Hidden Markov Models (HMMs) Method

I. Length of Temporal Pattern – L

As previously mentioned, the pattern dimension Q was introduced in the improved Time Series Data Mining method. For the improved HMMs, in most cases, the dimension of the pattern states may not be known. It is suggested to apply the false nearest neighbor method (Kantz and Schreiber, 2004) to find the length of the pattern states. It can be considered that the length of temporal pattern L is equal to the pattern dimension Q. In this thesis, the L value is calculated as 3, it means there are 3 states for the pattern state. Also, the normal state and the event state should be considered for the total length of the data. This means there are 5 states for the data. State 1 is the normal state (normal data), states 2, 3, and 4 are the pattern states, and state 5 is the event state.

II. The Initial States Value for the SVI in Transition Matrix

During the testing procedure for the HMMs method, one problem that was noted is that sometimes the thresholds found by the Mixture of Gaussians function were not reasonable. For instance, sometimes the computer program found the threshold for the normal state of the SVI data was 120 mL/g, and the threshold for the event state of the SVI data was 80 mL/g. For the sludge bulking problem, such conditions are not reasonable because the event state threshold value should be higher than the value of the normal state threshold. Also, once such a condition happened, the program failed to predict the future probabilities of the states. So threshold values for different states were set by the analyst after the Mixture of Gaussians function to prevent such an unstable condition. In the improved HMMs computer programs, threshold values were set to [80, 120, 120, 120, 200] which means [normal state value, pattern state point 1 value, pattern state point 2 value, pattern state point 3 value, event state value].

3.3. The Combined Method of Hidden Markov Models (HMMs) and Multinomial Logistic Regression (MLR) Model

3.4.1. Introduction

The improved TSDM and HMMs methods were applied to detect the temporal patterns and predict future sludge bulking events considering the SVI data alone. It may be useful that if sludge bulking events could be detected by measured values of other wastewater parameters. A Multinomial Logistic Regression (MLR) Model can predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables. In the improved HMMs method, the

hidden states (normal, patterns, and event) can be obtained by the Viterbi Algorithm. This combination method tries to use other chemical or physical variables to predict the sludge bulking event state.

Suppose a dependent variable has M categories. One value (typically the first, the last, or the value with the highest frequency) of the dependent variable is designated as the reference category (Menard, 2001). In our case, the hidden states is the dependent variable, and it has five categories, normal state, pattern state 1, pattern state 2, pattern state 3, and event state. And normal state is chosen as the reference category.

For m = 2, 3, 4, and 5, the probability is calculated as:

$$P(Y_i = m) = \frac{\exp\,(X_i * \beta_m)}{1 + \sum_{j=2}^{5} \exp\,(X_i * \beta_j)} \tag{3.8}$$

For the reference category (normal state), the probability is:

$$P(Y_i = 1) = \frac{1}{1 + \sum_{j=2}^{5} \exp\,(X_i * \beta_j)} \tag{3.9}$$

where Y_i is the observed outcome for the ith observation on the dependent variable, X_i is a vector of the *i*th observations of all the explanatory variables, and β_j is a vector of all the regression coefficients in the *j*th regression. From the foregoing equations, it can be seen that the multinomial logistic regression model measures the possibility of the hidden state for each SVI data value as a function of other wastewater quality parameters.

3.4.2. The Combined Method Process

There are many wastewater quality variables available from the data collected from the North Side Water Reclamation Plant (NSWRP) of the Metropolitan Water Reclamation District of Greater Chicago. The first step is to choose some important variables that are considered to be highly correlated to the SVI data. Some statistical

analysis is chosen, like crosstab analysis, correlation function analysis, etc., to detect the important variables. Several wastewater treatment quality parameters used in the combined method are listed in Table 1.1.

The second step is the same as the training step of the improved HMMs method, to get the hidden states for each data point in the training data set. Then the hidden states and selected parameters are applied to build the multinomial logistic regression model.

Finally, with the multinomial logistic regression model built in the second step, the method performs the sludge bulking event state prediction process by applying the data for the selected variables. Unlike the improved TSDM method and HMMs method, the prediction process is done considering other wastewater quality variables instead of the SVI data alone. Figure 3.5 demonstrates the general process of the combined method.

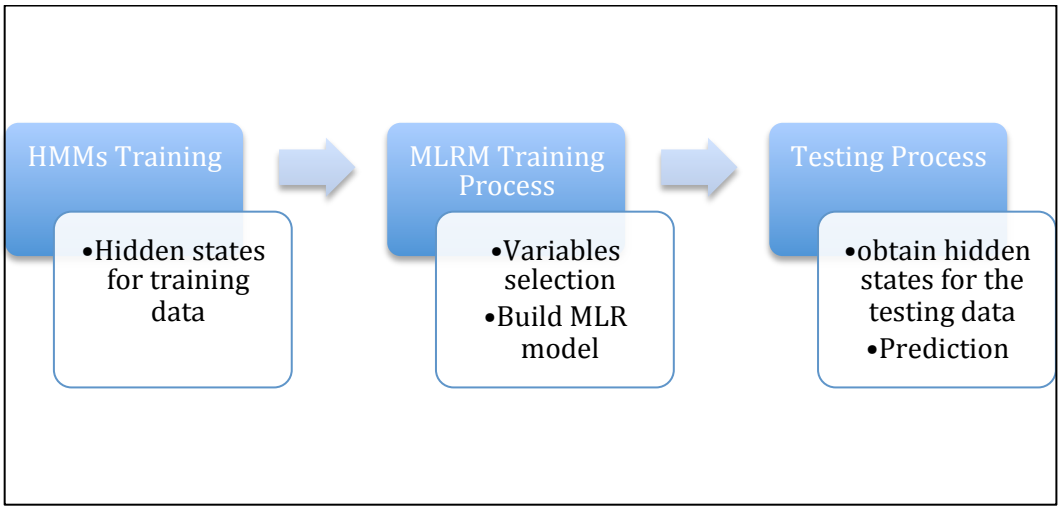

Figure 3.5: General Process of Combined Method of HMMs and MLR Model

CHAPTER 4 ANALYSIS ON DETECTION OF HIGH AMMONIA CONCENTRATION AND SLUDGE BULKING PROBLEMS USING THE IMPROVED TIME SERIES DATA MINING (TSDM) METHOD

The general process of the improved TSDM method is described in Chapter 3. The application of the improved TSDM method to high ammonia concentrations and sludge bulking are described in this chapter. The improved TSDM method computer programs were written by Mr. Hai Huang in the MATLAB at Marquette University. Later these programs were modified to make them more suitable for application to data from WWTPs.

The application of the improved TSDM method to problems with WWTP operations includes three sections. First, synthetic data created by the author is applied to the improved TSDM method, and the steps of the improved TSDM method are demonstrated. After that, the ammonia data and the Sludge Volume Index (SVI) data for the North Side Water Reclamation Plant (NSWRP) are analyzed by the improved TSDM method. The results are presented and discussed in the following sections.

4.1. Synthetic Data Test and Discussion

Seventeen hundred synthetic data points were generated by author. The normal value points were generated randomly between 0 and 11. Then the event value is defined as a time series x value is greater than 10, which means the event function is:

$$g(t) = x(t) - 10.0 > 0 \tag{4.1}$$

There are three values before each event value, and these pattern values were generated to follow a trend of down to up, e.g., 4.6, 4.0, and 4.3. After that, the noise was generated which follows a standard normal distribution with a mean value of 0 and a

standard deviation value of 1, then the noise was added into the whole synthetic time series data. For the original synthetic data without noise, the embedding dimension Q is three because there are three points before each event value. However, after adding the noise into the synthetic data, the embedding dimension Q perhaps has another value.

The SVI data, which is analyzed later in Section 4.4 of this chapter, has a large range of values. For better similarity to the SVI data, the synthetic data are generated with a high standard deviation value. Table 4.1 lists the parameters of the synthetic data, and a plot of the synthetic data and the event value line are shown in Figure 4.1.

Table4.1: Parameters of the Synthetic Data

Total number	1700
Mean	1.5373
Standard deviation	2.6375
Maximum value	13.3481
Minimum value	-2.7999
Event value	10
Number of events (≥10)	53

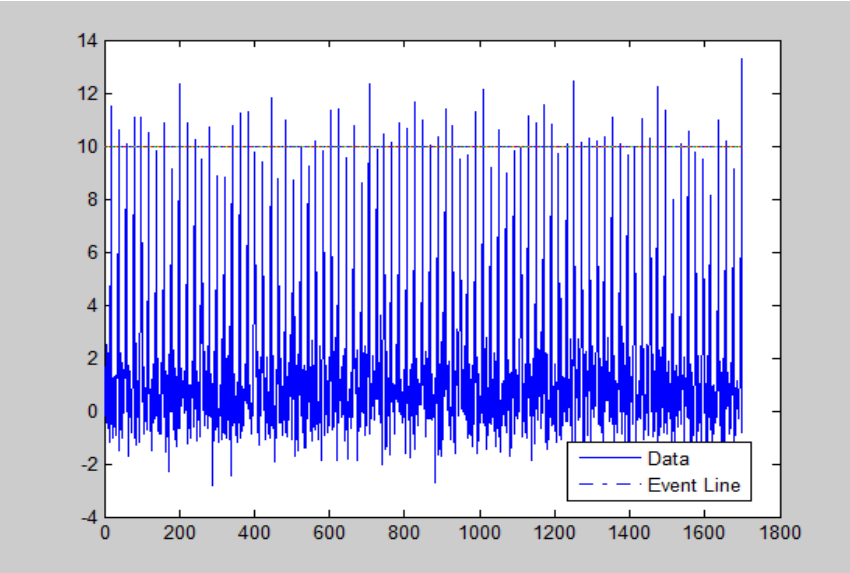

Figure 4.1: Plot of the Synthetic Data

Two parameters of time series embedding, the time-delay, τ, and the embedding dimension, Q, need to be determined before the training process. Normally, the time-delay, τ, is set to 1, which is used to detect the temporal patterns before the next value. The embedding dimension Q can be calculated by the false nearest neighbor algorithm. In this synthetic data test, the Q is calculated as seven. However, it should be noted that the time-delay and embedding dimension can be changed by the analyst to get the best conditions for the test.

Another import parameter of the TSDM method is the radius enlarge ratio of the temporal pattern clusters. It is used to magnify the radius of the temporal pattern cluster because, in the phase space, under some conditions, the temporal pattern points may not be in the cluster but near to the cluster. In such cases, the radius needs to be enlarged to include those points. However, magnifying the radius will lead the temporal pattern clusters to contain points which are not the temporal patterns. So the radius enlarge ratio should be chosen carefully by several attempts. Normally, it is set to between 1 and 2. The process of choosing the radius enlarge ratio is shown in the Section 4.1.2.

4.1.1. Training Process

From the total of 1700 data points in the synthetic time series, the first 1400 data points are used as the training time series to find the temporal pattern clusters applying the improved TSDM method. The embedding dimension, Q, is calculated as seven, time-delay, τ, is chosen as 1, and enlarge ratio of temporal pattern clusters radius is set to 1 (the best value for the enlarge ratio will be discussed later). The two-step optimization algorithm is used in the training process. The first step takes all data points as the temporal pattern cluster center candidates. The clusters are sorted according to their

objective function values, and the best clusters that found in the first optimization step are preserved for the second optimization step. After that, the temporal pattern clusters in the training time series are found. Tables 4.2 and 4.3 demonstrate the results of the first optimization step and second optimization step, respectively.

Table 4.2: Cluster results of the first optimization step search for synthetic data

Cluster No.	Cluster Center							Radius	Cluster Size
1	4.0558	4.2886	5.7595	-0.6967	1.1004	2.1821	0.4128	0.2000	1
2	3.2776	4.6463	5.7994	1.4756	1.1135	0.4651	-0.0769	1.4151	2
3	6.2625	3.5319	5.0101	-0.1189	0.2264	0.3374	-0.9423	0.2000	1
4	6.4022	4.6818	2.9991	1.4035	0.7253	1.3274	-1.0846	0.2000	1
5	2.9993	4.9595	5.2653	1.0672	1.1239	0.4290	0.1128	1.2744	2
6	3.1020	3.5794	5.1845	0.7671	-0.7704	1.6204	0.0719	0.2000	1
7	6.2637	3.0657	5.6251	2.5457	-0.4570	-0.1129	2.3735	0.2000	1
8	3.2410	5.3776	4.1207	0.9234	1.0761	3.6834	0.0177	0.2000	1
9	3.6634	6.0120	4.1233	1.2173	0.1461	1.4522	0.9289	0.2000	1
10	3.2421	3.0379	4.0585	1.5284	1.1841	-1.3327	1.2251	0.2000	1
11	4.9326	2.8724	4.4449	0.8712	-0.3952	1.5285	-0.5066	0.1930	1
12	5.6213	3.5916	3.7823	0.1162	-0.1668	-0.7331	0.1944	0.2000	1
13	5.3808	3.1553	6.0224	2.0271	-0.1939	0.2269	1.7574	0.2000	1
14	3.3251	3.7732	4.9593	0.4437	0.2856	0.4157	0.3875	0.2000	1
15	3.2915	3.0439	5.7947	0.1383	0.4993	-0.5338	0.5900	0.2000	1

Table 4.3: Cluster results of the second optimization step search for synthetic data

Cluster No.	Cluster Center							Radius	Cluster Size
1	3.3265	3.7669	5.7852	0.4437	0.2867	1.6204	0.4076	2.5554	16
2	4.0556	3.5446	5.7943	0.1641	0.0891	0.3549	0.4127	2.5657	27

From Table 4.2, it can be seen that 15 small temporal pattern clusters are detected by the first optimization step. After the second optimization step, these 15 small clusters are combined into the two bigger clusters listed in Table 4.3.

4.1.2. Testing Process

The final 300 data points are applied as the testing time series data. After

embedding the testing time series data into the same dimension phase space, the final cluster identification found five pattern points inside the temporal pattern cluster from the testing time series data meaning the improved TSDM method made five predictions based on these five temporal pattern points. Figure 4.2 shows the results of the testing process.

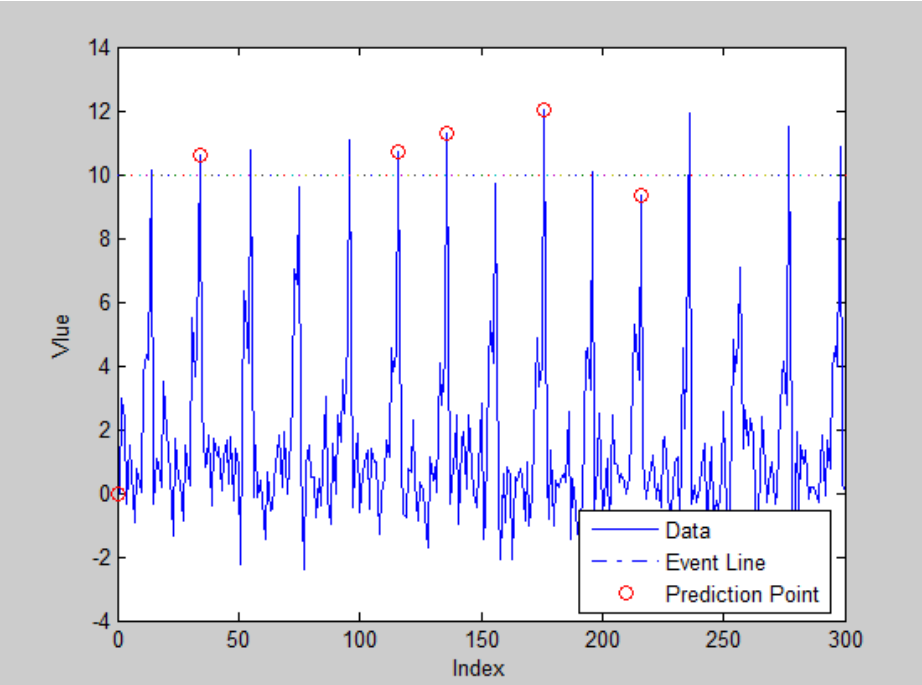

Figure 4.2: Testing Process Results for the Synthetic Data

From Figure 4.2, it can be seen that four predictions are correct predictions. The correct percentage is 80%. However, there are 11 events in the testing time series data, so the accuracy percentage is 36.36%. The reason for the low accuracy percentage is the radius of the temporal pattern clusters has not been magnified, so that not all temporal pattern points fell into the temporal pattern clusters. So, the radius enlarge ratio needs to be set larger than 1. But the radius enlarge ratio should be chosen carefully because some points that are not temporal pattern points may fall into the enlarged temporal pattern cluster and it will reduce the correct percentage. Table 4.4 demonstrates the relationship

of the radius enlarge ratio with the correct percentage and accuracy percentage. From Table 4.4, it can be seen that with the increase of the radius enlarge ratio, the accuracy percentage is increased. However, the correct percentage initially increased and then decreased with the increase of the radius enlarge ratio. So the analyst needs to find the balance of the correct percentage and accuracy percentage to determine a suitable radius enlarge ratio. For the synthetic data set, the best radius enlarge ratio is 1.6 because the improved TSDM method can obtain 100% accuracy percentage and an acceptable correct percentage of 73.33% which is better than that of 1.5 and 1.7. For the ammonia and the SVI data tests this process is repeated to determine the appropriate radius enlarge ratio for the first data combination test.

Table 4.4: Selection of Enlarge Ratio for Temporal Pattern Cluster Radius

Enlarge ratio	Total number of events	No. of predictions	Correct prediction	Correct Percentage	Accuracy Percentage
1	11	5	4	80.00%	36.36%
1.1	11	8	7	87.85%	63.64%
1.2	11	10	9	90.00%	81.82%
1.3	11	10	9	90.00%	81.82%
1.4	11	11	9	81.82%	81.82%
1.5	11	14	11	71.43%	90.91%
1.6	11	15	11	73.33%	100.00%
1.7	11	16	11	68.75%	100.00%

4.2. Period I – Ammonia Test and Discussion

Prior to obtaining the SVI data, ammonia data were used as the experiment subject. Effluent ammonia data from the NSWRP from 2001 to 2008 were considered. The WWTP's effluent ammonia limitation of 2.5 mg/L only applies from April to October, and, thus, only data from these months are considered. The event was set as an ammonia concentration larger than 2.0 mg/L, which is less than 2.5 mg/L, as it was

known from discussions with the MWRDGC that the WWTP operators would like to prevent effluent concentrations from even getting near the permit limit.

4.2.1. Ammonia Data Basic Analysis

Eight years of ammonia data were obtained in period I. Before May 1, 2001, no ammonia data were available from the NSWRP. The missing data from April 2001are set to 0 in the test. Table 4.5 shows some basic analysis of the ammonia data, including the number of events, the mean value, and the standard deviation for each test year of ammonia data.

Table 4.5: Basic Analysis of Ammonia Data in each Test Year

Year	Number of data in the year	Mean	Standard deviation (STD)	Number of Events (≥2 mg/L)
2001	214	0.3642	0.4468	2
2002	214	0.5897	0.3803	1
2003	214	0.8074	0.4864	6
2004	214	0.8602	0.6321	13
2005	214	0.8521	0.7703	7
2006	214	0.5268	0.5105	6
2007	214	0.3356	0.2788	0
2008	214	0.1508	0.1387	0

From Table 4.5, it can be seen that in the test year 2001, the standard deviation is larger than the mean value. The reason is the ammonia data in April were set to 0 because of the lack of data from the NSWRP in test year 2001. The mean values for the test years of 2003 to 2005 are higher than the other years. Meanwhile, there are more events in the test years of 2003, 2004, and 2005. Finally, no events occurred in the test years of 2007 and 2008, so there is no need to do the test on these two years.

4.2.2. Training and Testing Process

According to the previous basic ammonia data analysis, it is necessary to select the training data set and testing data set. It is better to have more events in the training data set in order to let the TSDM programs better learn and generate the temporal pattern clusters. So the data of test year 2004 is selected to be included in the training data set. Two combination sets of training and testing sets are applied: (a) training data set of 2001 to 2004, testing data set of 2005; and (b) training data set of 2001 to 2005, testing data set of 2006.

I. First Data Combination: Training data of 2001 to 2004, Testing data of 2005

According to the previous description of parameters of the improved TSDM method, the initial parameters need to be determined before the training and testing process. The embedding dimension, Q, is calculated by the false nearest neighbor method as 4. Time delay, τ, is set to 1 as to detect the temporal pattern 1 day before the event. The cluster radius enlarge ratio is originally selected as 1, the best radius enlarge ratio for the ammonia test will be analyzed later. These parameters are listed in Table 4.6.

Table 4.6: Initial Parameters

Parameter	Value
Embedding dimension- Q	4
Time delay τ	1 day
Training Procedure	Two step optimization algorithm
Cluster radius enlarge ratio	1

The resulting temporal pattern clusters found in the training process by two-step optimization are listed in Table 4.7. From Table 4.7, it should be noted that some temporal pattern clusters have a small radius and few pattern points, and even have only

one point in them. The radii of these temporal pattern clusters are very small, and from the phase space view, they are several points in the phase space. These clusters are not effective because there will be a very small chance that the data points from the test time series data will fall into these clusters. Clusters No. 6, No. 8, and No. 13 are more effective than the others, because of their larger size and inclusion of more pattern points. Especially cluster No. 13 has the largest radius and most temporal pattern points. So clusters No. 6, No. 8, and No. 13 may be effective.

Table 4.7: Temporal Pattern Clusters of Training Data in the First Data Combination

Cluster No.	Cluster Center				Radius	Cluster Size
1	1.5500	0.6600	0.8400	1.1700	0.0377	1
2	1.5100	1.6800	2.3200	1.2400	0.0103	1
3	2.4300	2.6000	2.4700	2.1700	0.1274	1
4	1.0400	0.7700	1.0000	1.0300	0.0295	1
5	0.7700	0.6000	0.5600	0.7100	0.0128	1
6	2.1358	1.5125	1.7500	1.7498	0.4050	4
7	1.2100	0.3000	3.1500	2.4300	0.1874	1
8	2.0823	1.4086	1.7504	1.7500	0.4084	4
9	1.3500	1.0900	0.4500	1.2700	0.0500	1
10	2.4700	2.1700	1.3200	1.7100	0.0173	1
11	0.3300	0.2800	0.3700	0.2100	0.0106	1
12	1.2600	1.0800	0.5700	0.3900	0.0394	1
13	2.4456	2.6000	2.3009	1.4488	0.8620	15
14	0.9300	1.0900	1.6200	0.9200	0.0297	1
15	1.2400	1.1500	1.1200	0.9600	0.0379	1
16	1.2400	0.8900	1.1800	1.5600	0.0223	1
17	0.5000	0.2700	1.4300	1.4400	0.0143	1
18	0.7700	0.8500	0.7400	1.3500	0.0113	1
19	1.3200	1.7100	1.6500	2.0900	0.0070	1
20	0.8000	0.1900	1.5800	1.0100	0.0575	1

After the training process, the test time series data are embedded into the phase space with same embedding dimension. If any data point falls into one of the three

temporal pattern clusters, the improved TSDM method will make a prediction. The result is shown in Figure 4.3 and Table 4.8.

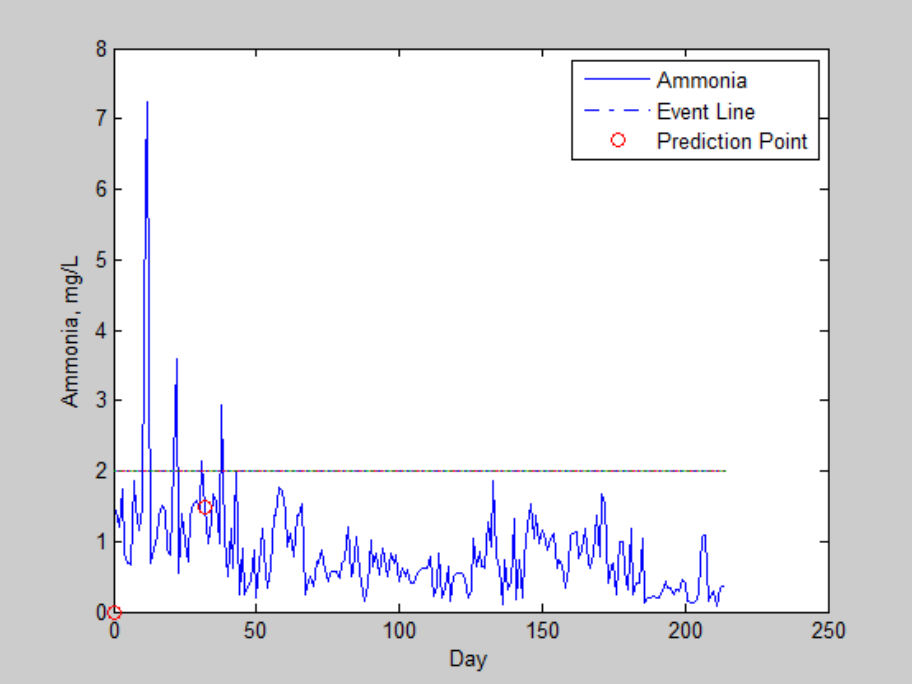

Figure 4.3: Testing Result of Ammonia in 2005 with Radius Enlarge Ratio as 1

Table 4.8: Testing Result of Ammonia in 2005

Testing Year	Total number of events	Number of detected patterns	Correct predictions	Correct Percentage	Accuracy percentage
2005	7	1	0	0%	0%

From Figure 4.3 and Table 4.8, it can be seen that one event was incorrectly predicted by the improved TSDM method. Also, the total number of events is 7, and the accuracy percentage is only 0%. However, the radius enlarge ratio is 1which means the temporal pattern clusters may not include those pattern points located slightly outside of the clusters in the phase space. So the selection process for a better radius enlarge ratio was performed. Table 4.9 lists the results for the selection of the radius enlarge ratio. The radius enlarge ratio is chosen as 1.9 according to the results in Table 4.9, because it has the highest correct percentage and highest accuracy percentage.

Table 4.9: Selection of Radius Enlarge Ratio for Ammonia Test

Enlarge ratio	Total number of events	No. of predictions	Correct prediction	Correct Percentage	Accuracy Percentage
1	7	1	0	0%	0%
1.1	7	1	0	0%	0%
1.2	7	1	0	0%	0%
1.3	7	1	0	0%	0%
1.4	7	1	0	0%	0%
1.5	7	1	0	0%	0%
1.6	7	2	0	0%	0%
1.7	7	4	0	0%	0%
1.8	7	5	1	20%	14.29%
1.9	7	6	2	33.33%	28.57%
2.0	7	7	2	28.57%	28.57%
2.1	7	2	2	25%	28.57%

Figure 4.4 shows the test result of 2005 with the radius enlarge ratio as 1.9. From Figure 4.4, it can be seen that only two high ammonia concentration events were detected and predicted by the improved TSDM method. From the prediction point of view, the accuracy percentage is not acceptable for the WWTP. However, from the prevention point of view, it can be found that some false positive detection points occurred prior to an actual event, and those false positive points are still useful for the prevention of the high ammonia concentrations.

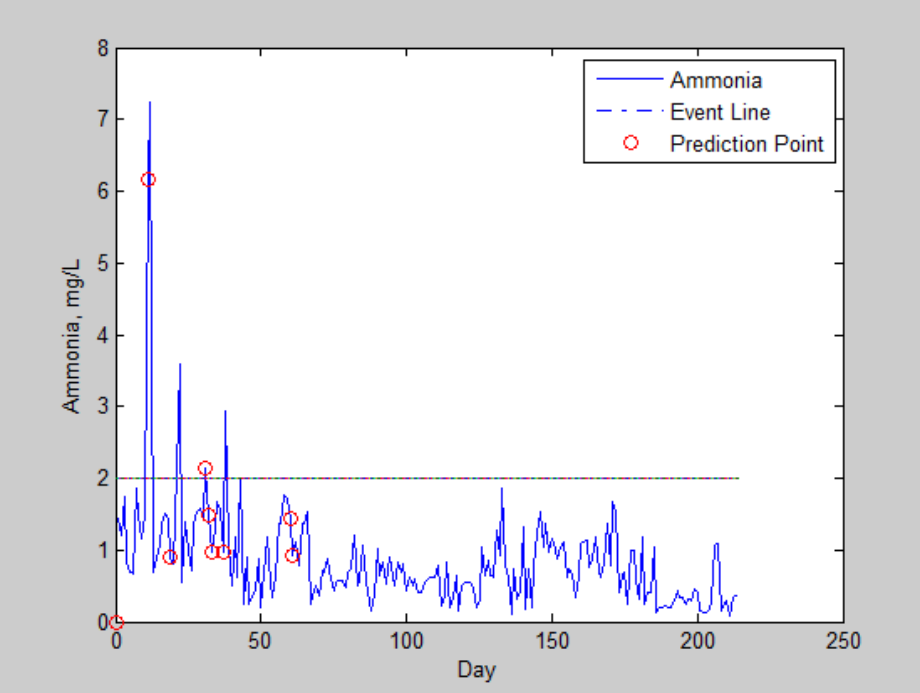

Figure 4.4: Testing Result of Ammonia in 2005 with Radius Enlarge Ratio as 1.9

II. Second Data Combination: Training data of 2001 to 2005, Test data of 2006

After the first data combination test, the second data combination test was performed. Also, the initial parameters need to be determined before the training and testing process. The embedding dimension, Q, is calculated by false nearest neighbor method as four. Time delay, τ, is set to one to detect the temporal pattern 1 day before the event. The cluster radius enlarge ratio is set to1.9 as per the previous section.

The resulting temporal pattern clusters found in the training process are listed in Table 4.10. From Table 4.10, comparing with the test of the first data combination, the radii of the second data combination temporal pattern clusters are smaller. Also, no cluster has more than three temporal pattern points. This result means that it is highly unlikely that the test time series data points will fall into these temporal pattern clusters.

Table 4.10: Temporal Pattern Clusters of the Training Data in the Second Data Combination

Cluster No.	Cluster Center				Radius	Cluster Size
1	6.16	1.49	1.16	1.49	0.2	1
2	1.49	1.16	1.49	1.85	0.0632	1
3	2.1	0.81	0.91	1.44	0.0466	1
4	1.55	0.66	0.84	1.17	0.0379	1
5	1.51	1.68	2.32	1.24	0.0148	1
6	2.43	2.6	2.47	2.17	0.1274	1
7	1.04	0.77	1	1.03	0.0297	1
8	0.77	0.6	0.56	0.71	0.0129	1
9	0.97	1.56	1.67	1.29	0.0376	1
10	1.94	1.58	1.75	1.75	0.0261	1
11	1.21	0.3	3.15	2.43	0.1001	1
12	1.35	1.09	0.45	1.27	0.0501	1
13	2.47	2.17	1.32	1.71	0.0016	1
14	0.33	0.28	0.37	0.21	0.0076	1
15	1.26	1.08	0.57	0.39	0.0286	1
16	2.4330	2.5771	2.4640	2.17	0.4754	2
17	0.93	1.09	1.62	0.92	0.0298	1
18	2.1712	1.3200	1.7423	1.75	0.2591	3
19	1.24	1.15	1.12	0.96	0.0389	1
20	1.24	0.89	1.18	1.56	0.0298	1

Figure 4.5 and Table 4.11 demonstrate the testing result of second data combination. From Figure 4.5 and Table 4.11, it can be seen no events can be predicted because no temporal pattern was detected by the TSDM method. As previously discussed the reason for this result is that the temporal pattern clusters in the training step are too small to let the test time series points fall into them.

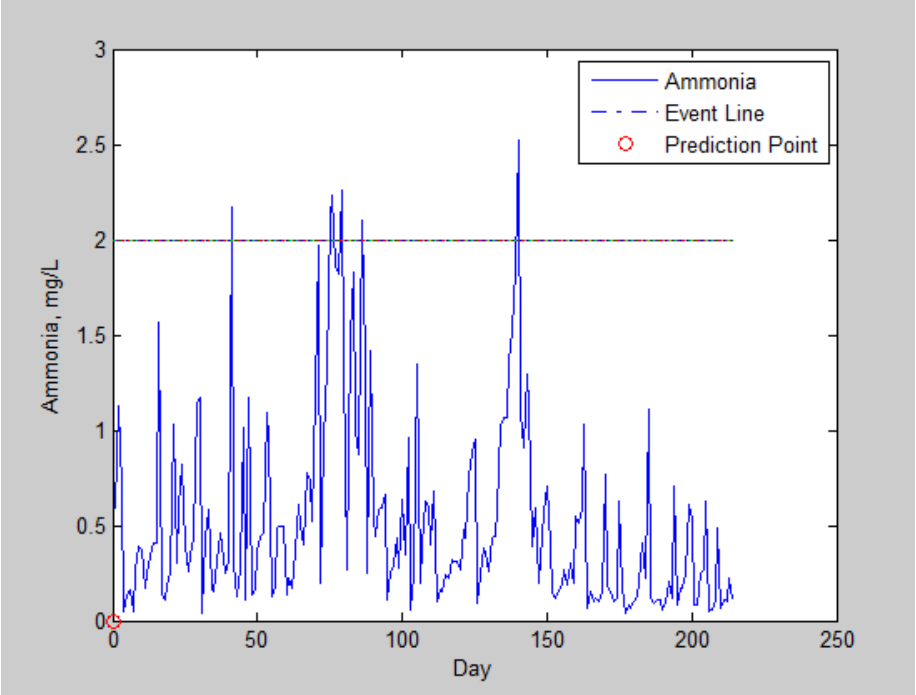

Figure 4.5: Testing Result of Ammonia in 2006

Table 4.11: Testing Result of Ammonia in 2006

Testing Year	Total number of events	Number of detected patterns	Correct predictions	Correct Percentage	Accuracy percentage
2006	6	0	0	0%	0%

4.3. Improvement of the TSDM Process by Modifying the Initial Parameters

From the previous ammonia test, the test results are not acceptable. Two aspects are considered as possible causes for the poor results.

First, the ammonia data might be chaotic, and there are not enough events in the training data set for the improved TSDM method to learn and find the temporal pattern clusters. Although the event value can be reduced to 1.5 or 1.2 mg/L to get more events, such values are far below the permit limits of the WWTP, and so operators do not try to avoid these values.

Second, the initial parameters of the improved TSDM method can be modified

before the training and testing process. The radius enlarge ratio is 1.9, which is suitable for the improved TSDM method. Another parameter that can be modified is the embedding dimension, Q. Although the embedding dimension, Q, is calculated by the false nearest neighbor method in the training process, the analyst can change the Q value artificially. According to Huang (2001), if the radius of the temporal pattern clusters is too small after the training process, this is a sign that the embedding dimension, Q, is too high.

Since the Q in the second data combination test is four, and the cluster size of all the clusters is small. The Q in the second data combination test is changed to three and two to check whether the results will be better.

Table 4.12 and Figure 4.6 show the training and testing results when Q is changed to three in the second data combination test. From Table 4.12 and Figure 4.6, it can be seen that the temporal pattern clusters are still small, and the testing result is still not acceptable. Only one point was predicted, and this prediction point was incorrect. So Q is changed to 2 and the second data combination test was redone. The results are shown in Table 4.13 and Figure 4.7. The training and testing results with Q = 2 for the second data combination test are still not acceptable. The temporal pattern clusters are very small, and it is hard for the testing time series data to fall into these temporal pattern clusters. Besides, only one incorrect point is predicted. From the foregoing discussion, it can be concluded that the ammonia data are chaotic and lack sufficient events for the improved TSDM method to learn and detect the temporal patterns so that useful warning information on high ammonia concentrations can be made.

Table 4.12: Temporal Pattern Clusters in the Second Data Combination Test with Q = 3

Cluster NO	Cluster Center			Radius	Cluster Size
1	6.16	1.49	1.16	0.2	1
2	1.49	1.16	1.49	0.0632	1
3	2.1	0.81	0.91	0.0466	1
4	1.55	0.66	0.84	0.0379	1
5	2.43	2.6	2.2886	0.0148	1
6	1.51	1.68	2.32	0.1274	1
7	1.04	0.77	1	0.0297	1
8	0.77	0.6	0.56	0.0129	1
9	0.97	1.56	1.67	0.0376	1
10	2.4309	2.6	2.2856	0.0261	1
11	1.94	1.58	1.75	0.1001	1
12	1.21	0.3	3.15	0.0501	1
13	1.35	1.09	0.45	0.0015	1
14	2.47	2.17	1.32	0.0076	1
15	0.33	0.28	0.37	0.0286	1
16	1.26	1.08	0.57	0.4754	2
17	2.17	1.32	1.71	0.0298	1
18	0.93	1.09	1.62	0.2591	3
19	1.24	1.15	1.12	0.0389	1
20	1.24	0.89	1.18	0.0298	1

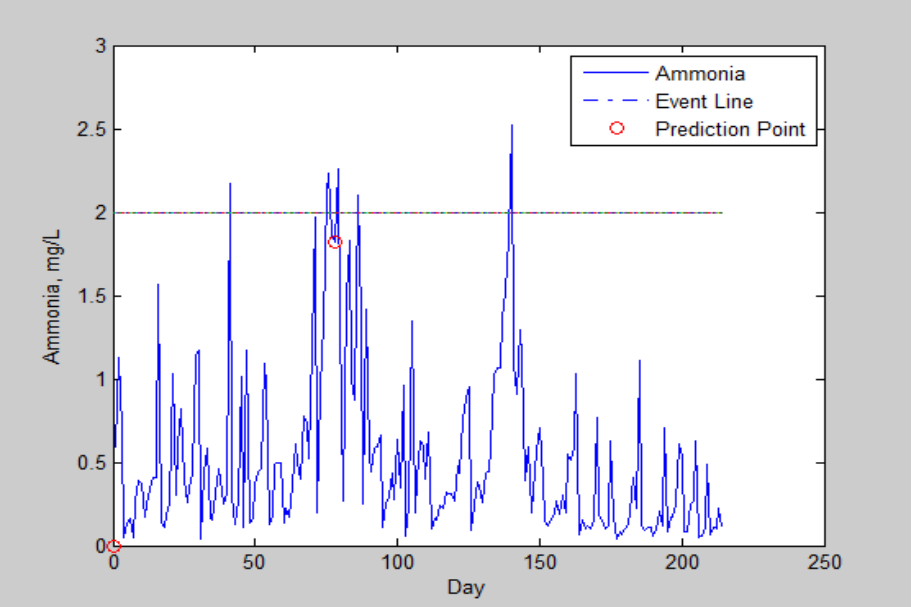

Figure 4.6: Testing Result of Ammonia in 2006 for the Second Data Combination Test with Q = 3

Table 4. 13: Temporal Pattern Clusters in the Second Data Combination Test with Q = 2

Cluster NO	Cluster Center		Radius	Cluster Size
1	6.16	1.49	0.2	1
2	1.49	1.16	0.0141	1
3	2.1	0.81	0.0022	1
4	2.47	2.5328	0.3064	3
5	1.55	0.66	0.0018	1
6	2.47	2.5364	0.3074	3
7	1.51	1.68	0.0006	1
8	1.04	0.77	0.0052	1
9	0.77	0.6	0.0050	1
10	0.97	1.56	0.0002	1
11	1.94	1.58	0.0336	1
12	2.47	2.5363	0.3073	3
13	1.21	0.3	0.0022	1
14	1.35	1.09	0.0076	1
15	1.26	1.08	0.0020	1
16	2.17	1.32	0.0222	1
17	0.93	1.09	0.0038	1
18	1.24	1.15	0.0037	1
19	1.24	0.89	0.0021	1
20	0.77	0.85	0.0006	1

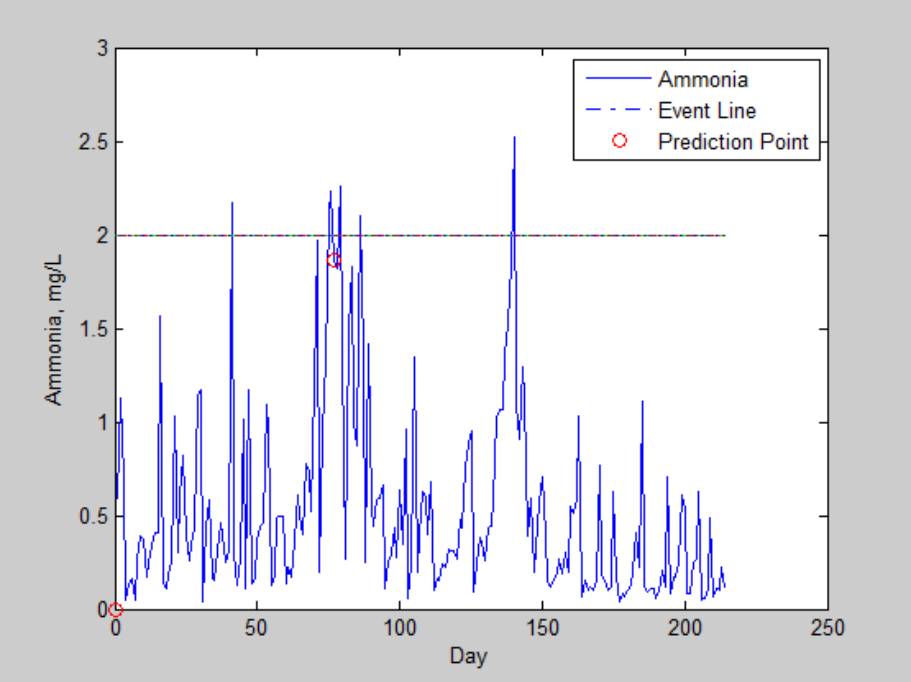

Figure 4.7: Testing Result of Ammonia in 2006 for the Second Data Combination Test with Q = 2

4.4. Period II – Sludge Volume Index Test

The SVI data were provided by the MWRDGC for the NSWRP in October 2010. Also, some wastewater treatment chemical and physical parameters were included in the database, i.e. preliminary and solids data (total wastewater flow, air flow, total solids, etc.), treatment operational data for each battery (return flow, etc.), nitrogen analysis data, and some lab analysis data. There are four treatment batteries at the NSWRP, and data for each battery are available from 2002 to 2009. As previously mentioned, the event value of sludge bulking is set by a SVI value greater than 150 mL/g.

Table 4.14: Initial Analysis of the SVI data for each treatment battery

	Battery B			Battery C			Battery D		
er of (>150 g)	Mean, ml/g	STD, ml/g	Number of Events (>150 ml/g)	Mean, ml/g	STD, ml/g	Number of Events (>150 ml/g)	Mean, ml/g	STD, ml/g	Number of Events (>150 ml/g)
	105.8274	21.0963	13	93.8329	17.5902	7	92.9260	26.8806	16
	93.4904	17.2024	3	85.8877	22.4328	8	71.3616	12.1319	0
	99.4699	25.5545	18	86.4590	21.1241	9	75.4344	14.6658	0
	113.5534	34.2264	34	102.5452	25.2072	15	81.3260	19.4970	7
	94.1859	14.8562	4	86.2466	21.3960	3	72.3342	12.4941	0
	99.4740	20.6648	10	87.6493	16.1452	0	71.0959	12.5354	0
	98.8306	18.9977	8	89.0464	20.2651	8	66.8634	8.1448	0
	99.2904	19.4732	5	90.5479	21.9277	0	70.3507	11.9667	0

Table 4.14 demonstrates an initial analysis of the SVI data for each treatment battery. From Table 4.14 some information can be found. 2005 has more sludge bulking events than other years, and it has the highest mean SVI value. Battery D is least affected by sludge bulking problems, compared to the other three batteries, especially, no sludge bulking events happened from 2006 to 2009 in Battery D. Battery B has more sludge bulking events than the other batteries. For this reason, the SVI data in Battery B was first studied in the sludge volume index test. Then the tests for batteries A and C are performed. Battery D will not be analyzed.

4.4.1. Results of SVI Test for Battery B

Because test year 2005 has most sludge bulking events, it is better to include the SVI data of 2005 in the training set, so the improved TSDM method can analyze more events to detect and learn the temporal patterns. Four different data combination tests were performed: (A) Training set: 2002 to 2005, Testing set: 2006; (B) Training set: 2002 to 2006, Testing set: 2007; (C) Training set: 2002 to 2007, Testing set: 2008; (D) Training set: 2002 to 2008, Testing set: 2009.

I. SVI Test of Data Combination A for Battery B

Table 4.15 lists the initial parameters of the first data combination test. Embedding dimension, Q, is calculated by false nearest neighbor method as three. Time delay is chosen as 1 to look for the 1 day ahead prediction. Cluster radius enlarge ratio is firstly set as 1. The selection process for the radius enlarge ratio will be performed later like the selection process in the ammonia test.

Table 4.15: Initial Parameters

Parameter	Value
Embedding dimension- Q	3
Time delay τ	1 day
Training Procedure	Two step optimization algorithm
Cluster radius enlarge ratio	1

Table 4.16 shows the result of temporal pattern clusters found in the training process. The testing results are shown in Table 4.17 and Figure 4.8. In Table 4.16, it can be seen that all clusters only have one temporal pattern, and the radii are all very small. According to the discussion in Section 4.3, such a condition means the embedding dimension Q is large. Also, from Table 4.17 and Figure 4.8, no temporal pattern is found from the training results.

Table 4.16: Temporal Pattern Clusters of Training Data in the First Data Combination

Cluster No.	Cluster Center			Radius	Cluster Size
1	301	243	222	0.2	1
2	243	222	154	0.2	1
3	203	163	137	0.2	1
4	227	212	181	0.2	1
5	174	190	166	0.2	1
6	221	233	188	0.2	1
7	222	154	159	0.2	1
8	188	152	95	0.2	1
9	204	345	301	0.2	1
10	280	203	163	0.2	1
11	212	181	187	0.2	1
12	154	159	141	0.2	1
13	233	188	152	0.2	1
14	162	201	162	0.2	1
15	198	124	146	0.2	1

Table 4.17: Testing Result of SVI in 2006 for Battery B with Q = 3

Testing Year	Total number of events	Number of detected patterns	Correct predictions	Correct Percentage	Accuracy percentage
2006	4	0	0	0	0

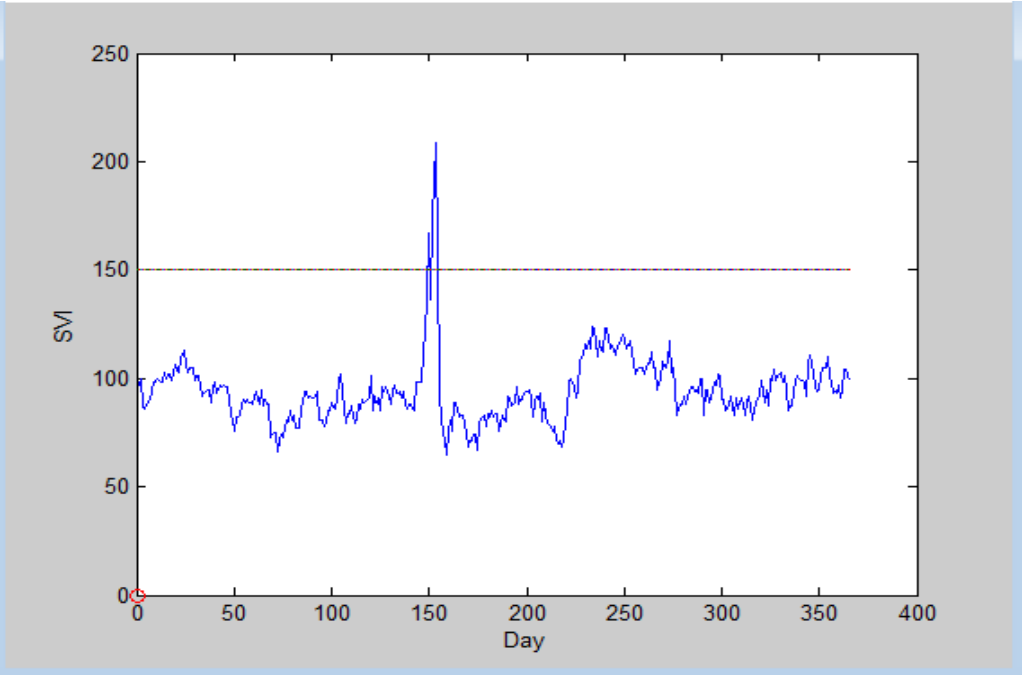

Figure 4.8: Testing Result of SVI in 2006 for Battery B with Radius Enlarge Ratio equal to 1 and Q = 3

For method improvement purposes, the embedding dimension is reset to two. Table 4.18 shows the training result of temporal pattern clusters by changing Q to 2. Four large temporal pattern clusters are found by the improved TSDM method, clusters Nos. 1, 2, 5, and 6. Each of them has more than 80 temporal pattern points. From the phase space view, the centers of these clusters are really close and these clusters have almost the same radius. So the analyst can consider them as one big cluster.

Table 4.18: Temporal Pattern Clusters of the Training Data for Battery B in First Data Combination with Q =2

Cluster No.	Cluster Center		Radius	Cluster Size
1	222.0000	190.0000	53.1174	86
2	222.0000	190.1546	53.1526	86
3	301.0000	243.0000	0.2000	1
4	243.0000	222.0000	0.2000	1
5	222.0000	189.5308	52.7910	84
6	222.0040	190.0000	53.1918	87
7	203.0000	163.0000	0.2000	1
8	227.0000	212.0000	0.2000	1
9	174.0000	190.0000	0.2000	1
10	221.0000	233.0000	0.2000	1
11	222.0000	154.0000	0.2000	1
12	188.0000	152.0000	0.2000	1
13	204.0000	345.0000	0.2000	1
14	280.0000	203.0000	0.2000	1
15	212.0000	181.0000	0.2000	1

Testing results are shown in Table 4.19 and Figure 4.9. It can be seen that the results are better by changing the embedding dimension to 2. The method could detect one event, and the accuracy percentage is 25%, which is not very high. However, from Figure 4.9, it can be seen the only predicted sludge bulking event in 2006 is detected after the occurrence of the sludge bulking problem. This means the sludge bulking problem was not detected efficiently at the first event point.

Table 4.19: Testing Result of SVI in 2006 for Battery B with Q = 2

Testing Year	Total number of events	Number of detected patterns	Correct predictions	Correct Percentage	Accuracy percentage
2006	4	1	1	100%	25%

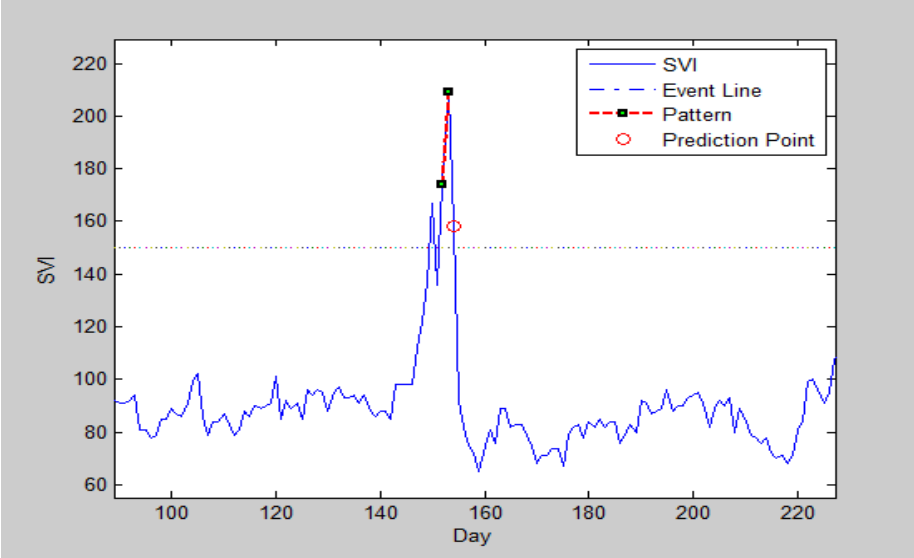

Figure 4.9: Testing Result of SVI in 2006 for Battery B with Q = 2

It should be noted that the radius enlarge ratio was set to 1. As previously mentioned, the radius enlarge ratio needs to be carefully selected like in the ammonia test. Table 4.20 shows the selection for the radius enlarge ratio for the SVI tests. It can be seen that 2.1 is the best radius enlarge ratio for the SVI test, because the accuracy percentage is 100% and correct percentage is the highest. For the rest of the SVI test, the radius enlarge ratio is chosen as 2.1.

Table 4.20: Selection of Radius Enlarge Ratio for SVI Test

Enlarge ratio	Total number of events	No. of predictions	Correct prediction	Correct Percentage	Accuracy Percentage
1	4	1	1	100%	25%
1.1	4	1	1	100%	25%
1.2	4	1	1	100%	25%
1.3	4	2	1	50%	25%
1.4	4	3	2	66.67%	50%
1.5	4	4	2	50%	50%
1.6	4	4	2	50%	50%
1.7	4	5	3	60%	75%
1.8	4	5	3	60%	75%
1.9	4	5	3	60%	75%
2.0	4	5	3	60%	75%
2.1	4	6	4	66.67%	100%

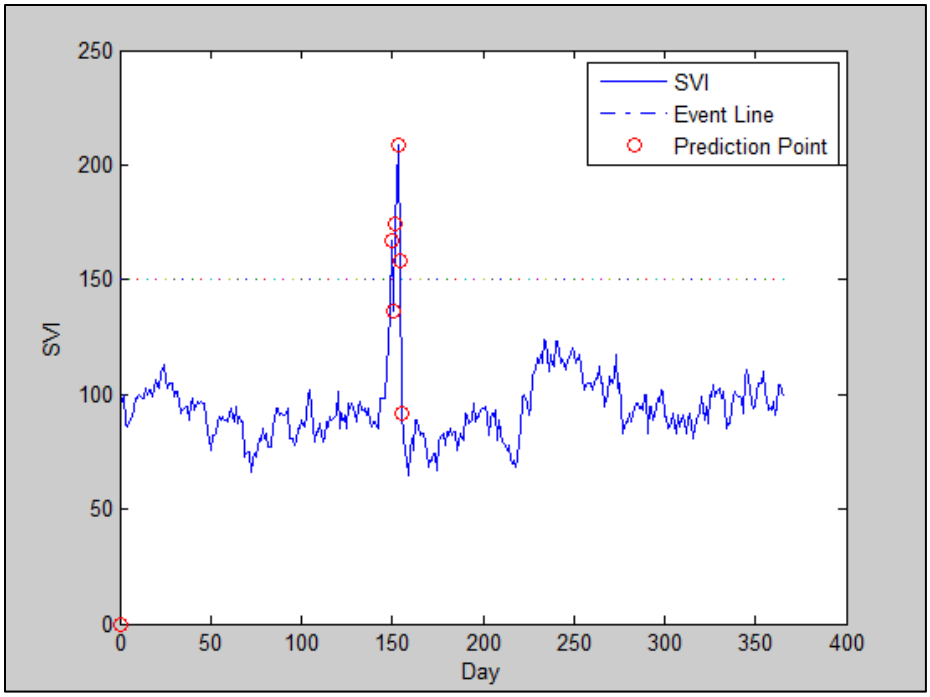

Figure 4.10: Testing Result of SVI in 2006 for Battery B with Q = 2

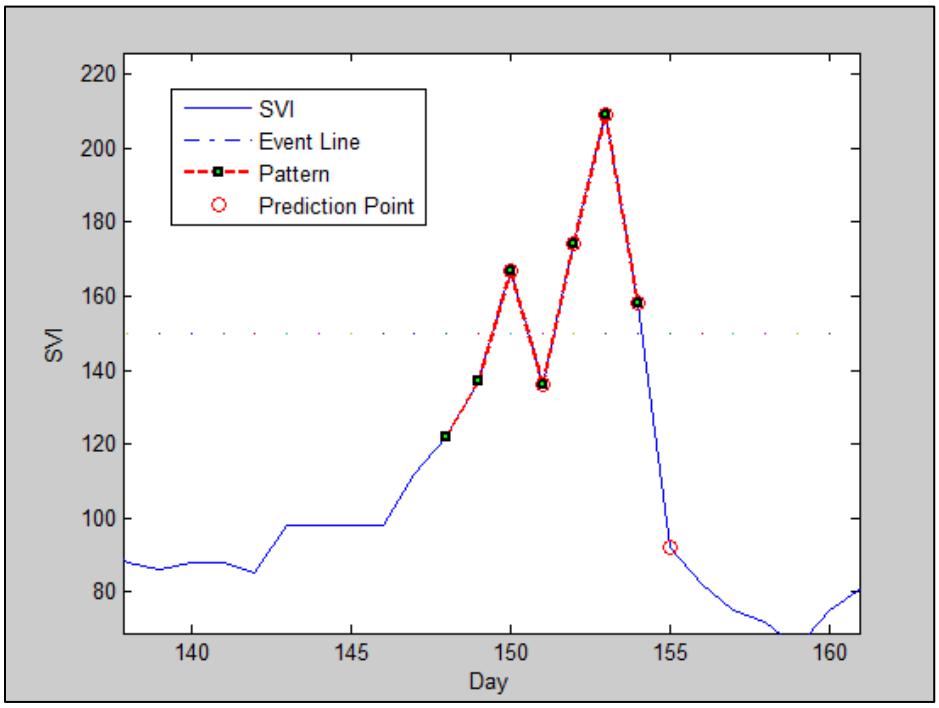

Figure 4.11: Pattern Plot for Testing Result of SVI in 2006 for Battery B with Q = 2

Figure 4.11 shows the enlargement of the detected patterns and prediction period of time in Figure 4.10. From Figure 4.11, it can be seen that once the pattern was

detected, the prediction point was found by the improved TSDM method. However, some results of interest should be noted. First, the patterns and the prediction events are overlap. Second, the first prediction point is a effective prediction point, meaning the improved TSDM method can provide warning information for sludge bulking problems.

II. SVI Test of the Data Combination B for Battery B

Like the test of data combination A, the embedding dimension is calculated as three in combination B. Still, no large temporal pattern cluster was found by the improved TSDM method. So Q is set as 2. Table 4.21 lists the results of the training process for Q = 2. The improved TSDM method with Q = 2 found four large temporal pattern clusters that can effectively be considered a single large pattern cluster.

Table 4.21: Temporal Pattern Clusters of the Training Data for Battery B for Data Combination B with Q =2

Cluster No.	Cluster Center		Radius	Cluster Size
1	222.0011	190.0000	52.4163	86
2	222.0000	189.9844	52.0768	86
3	301.0000	243.0000	0.2000	1
4	243.0000	222.0000	0.2000	1
5	222.0000	189.0897	52.3449	86
6	222.0000	189.2123	52.3090	86
7	203.0000	163.0000	0.2000	1
8	227.0000	212.0000	0.2000	1
9	174.0000	190.0000	0.2000	1
10	221.0000	233.0000	0.2000	1
11	222.0000	154.0000	0.2000	1
12	188.0000	152.0000	0.2000	1
13	204.0000	345.0000	0.2000	1
14	280.0000	203.0000	0.2000	1
15	212.0000	181.0000	0.2000	1

Testing results are shown in Table 4.22 and Figure 4.12. Figure 4.13 shows an enlargement of the pattern and prediction period of time in Figure 4.12. From Figure

4.13, it can be seen there are two sludge bulking periods, day 262 - 270 and day 320 - 330. Both sludge bulking periods were detected by the improved TSDM method, so the accuracy percentage is 100% which is a very high value. However, it should be noted that several prediction events in both periods are false positive predictions. From the prediction point of view, these false positive prediction points are not useful, and the correct percentage is 35.71%. But from the WWTP operator point of view, these false positive points can provide warning information for the impending sludge bulking, considering they also have high values near to the event line. This warning information may allow both sludge bulking periods to be prevented by the improved TSDM method.

Table 4.22: Testing Result of Battery B in 2007 with Q = 2

Testing Year	Total number of events	Number of detected patterns	Correct predictions	Correct Percentage	Accuracy percentage
2007	10	28	10	35.71%	100%

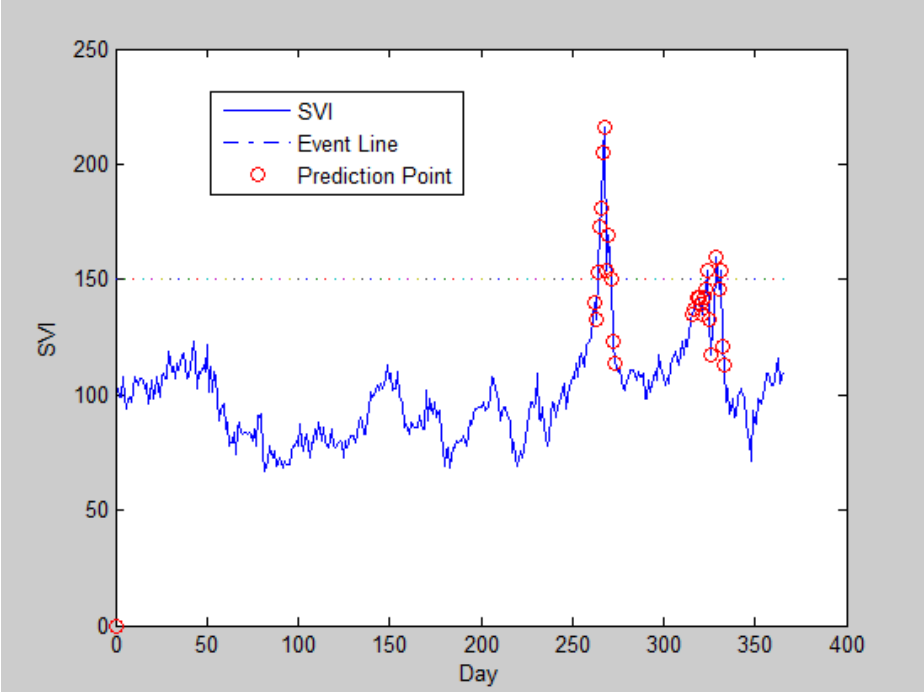

Figure 4.12: Testing Result of SVI in 2007 for Battery B with Q = 2

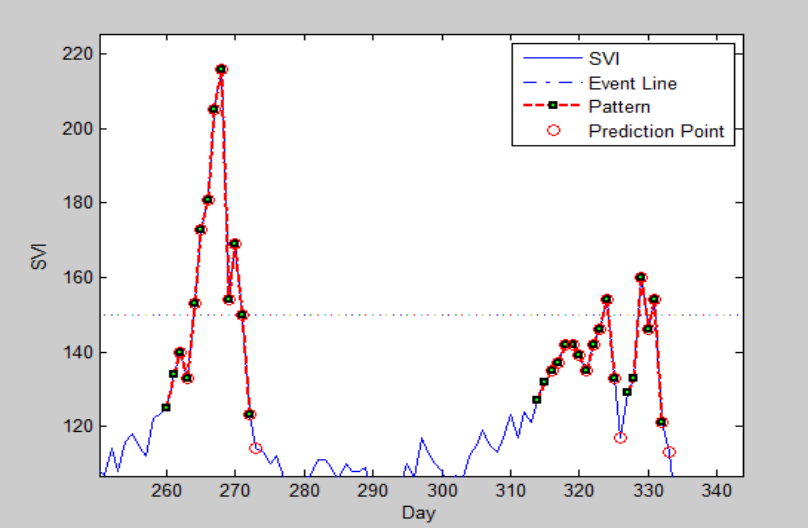

Figure 4.13: Pattern Plot for Testing Result of SVI in 2007 for Battery B with Q = 2

III. SVI Test of Data Combination C for Battery B

Again, the improved TSDM method cannot yield a good result with Q = 3, so Q is set to 2. Table 4.23 lists the results of the training process. Again the four large temporal pattern clusters can effectively be considered a single large pattern cluster.

Table 4.23: Temporal Pattern Clusters of the Training Data for Battery B In Data Combination C with Q = 2

Cluster No.	Cluster Center		Radius	Cluster Size
1	221.2982	188	51.0512	99
2	222	187.9883	51.3325	100
3	301	243	0.2	1
4	243	222	0.2	1
5	203	163	0.2	1
6	227	212	0.2	1
7	221.9990	188	51.3031	100
8	221.6069	188	51.2184	99
9	174	190	0.2	1
10	221	233	0.2	1
11	222	154	0.2	1
12	188	152	0.2	1
13	204	345	0.2	1
14	280	203	0.2	1
15	212	181	0.2	1

Table 4.24 and Figure 4.14 show the test result of SVI in 2008. Figure 4.15 shows an enlargement of pattern and prediction period of time in Figure 4.14. It should be noted that there are several prediction points before the sludge bulking event. Although they are false positive predictions, they do provide warning information that could prevent the sludge bulking before it happens.

Table 4. 24: Testing Result of Battery B in 2008 with Q = 2

Testing Year	Total number of events	Number of detected patterns	Correct predictions	Correct Percentage	Accuracy percentage
2008	8	19	8	42.21%	100%

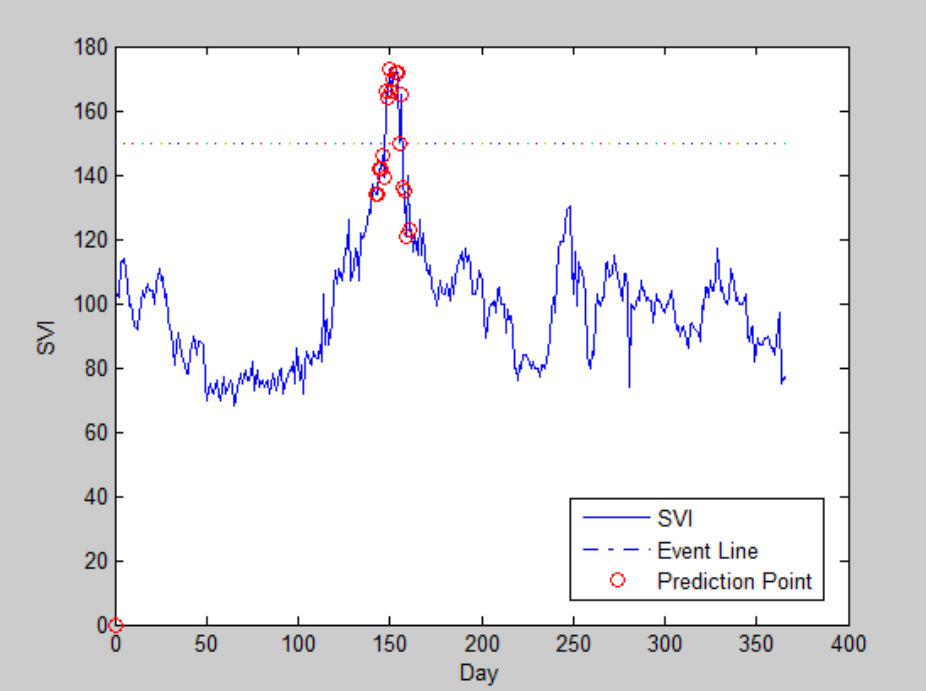

Figure 4.14: Testing Result of SVI in 2008 for Battery B with Q = 2

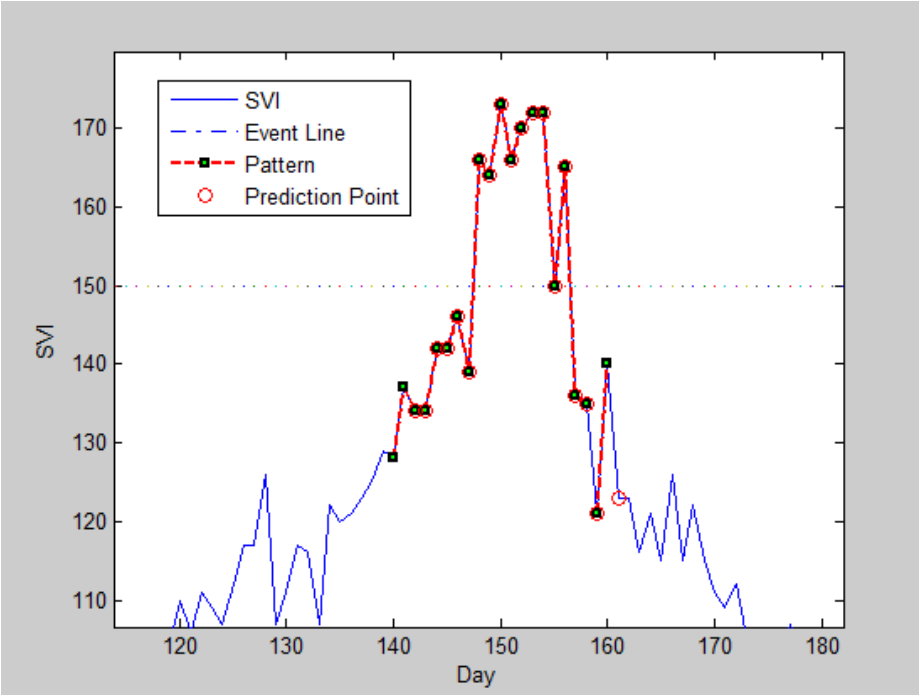

Figure 4.15: Pattern Plot for Testing Result of SVI in 2008 for Battery B with Q = 2

IV. SVI Test of Data Combination D for Battery B

Table 4.25 lists the results of the training process. Again the four large temporal pattern clusters can be effectively considered as a single large pattern cluster. Table 4.26 and Figure 4.16 show the testing results of SVI in 2009. Figure 4.17 shows an enlargement of the pattern and two prediction periods of time in Figure 4.16. The first sludge bulking period from day 100 to day 110 can be effectively prevented by the warning information from Figure 4.17. But the second sludge bulking period cannot be prevented because the second point of the first pattern is already higher than 150 mL/g.

Table 4.25: Temporal Pattern Clusters of Training Data in Data Combination D for Battery B with Q =2

Cluster No.	Cluster Center		Radius	Cluster Size
1	221.0000	187.3783	50.4337	105
2	221.0000	187.0016	50.4361	107
3	301.0000	243.0000	0.2000	1
4	243.0000	222.0000	0.2000	1
5	203.0000	163.0000	0.2000	1
6	227.0000	212.0000	0.2000	1
7	221.2012	187.0000	50.4990	106
8	221.0000	186.9829	50.6703	111
9	174.0000	190.0000	0.2000	1
10	221.0000	233.0000	0.2000	1
11	222.0000	154.0000	0.2000	1
12	188.0000	152.0000	0.2000	1
13	204.0000	345.0000	0.2000	1
14	280.0000	203.0000	0.2000	1
15	212.0000	181.0000	0.2000	1

Table 4.26: Testing Result of Battery B in 2009 with Q = 2

Testing Year	Total number of events	Number of detected patterns	Correct predictions	Correct Percentage	Accuracy percentage
2009	5	23	4	17.39%	80%

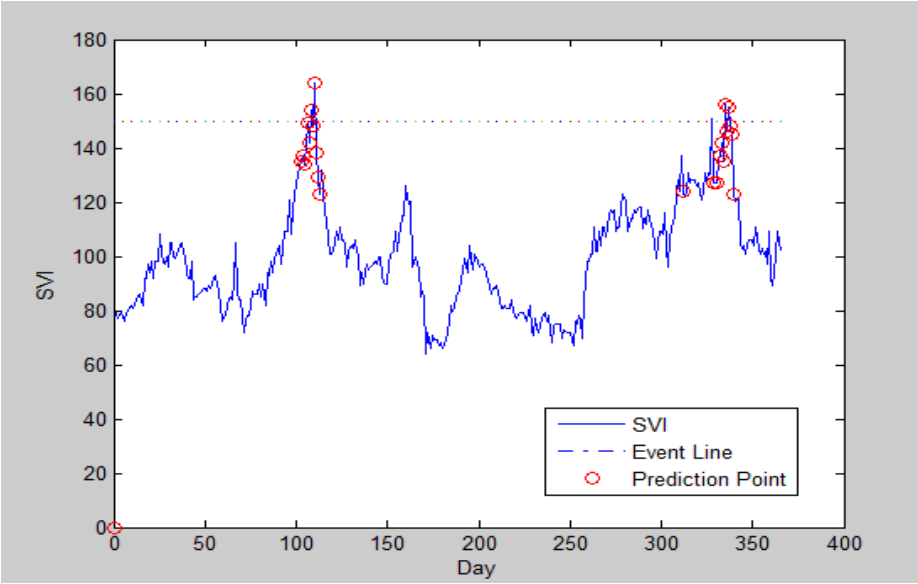

Figure 4.16: Testing Result of SVI in 2009 for Battery B with Q = 2

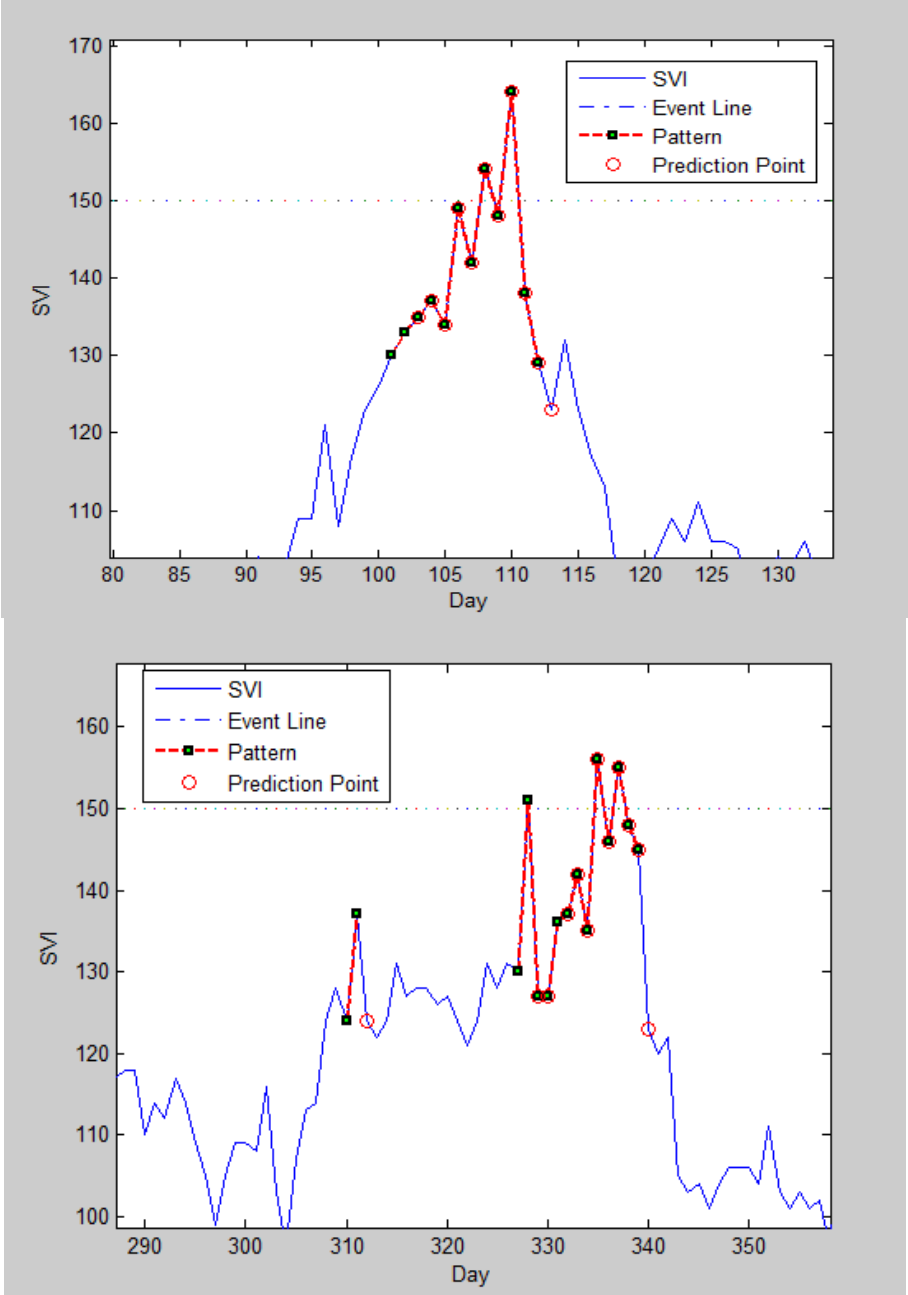

Figure 4.17: Pattern Plot for Testing Result of SVI in 2009 for Battery B with Q = 2

4.4.2. Results of SVI Test for Battery A

Table 4.27 shows the initial parameters for the improved TSDM method applied to data for Battery A. The initial parameters are same as the test of Battery B except the embedding dimension was set to Q = 2 which yields better results than Q = 3.

Table 4.27: Initial Parameters

Parameter	Value
Embedding dimension- Q	2
Time delay τ	1 day
Training Procedure	Two step optimization algorithm
Cluster radius enlarge ratio	2.1

Three different data combination tests are performed: (A) Training set: 2002 to 2005, Testing set: 2006; (B) Training set: 2002 to 2006, Testing set: 2007; (C) Training set: 2002 to 2007, Testing set: 2008. No sludge bulking events occurred in 2009 (Table 4.13) and no testing process can be performed for 2009. Figures 4.18, 4.20, and 4.22 and Table 4.28 show the testing results for three data combination tests. It can be seen that the sludge bulking events can be effectively detected by the improved TSDM method. Also, from Figures 4.19, 4.2,1 and 4.23, it can be seen that warning information can be provided by the improved TSDM method in 2007 and 2008, but not in 2006. From the prevention point of view, the sludge bulking periods in 2007 and 2008 can be effectively prevented using the results of the improved TSDM method.

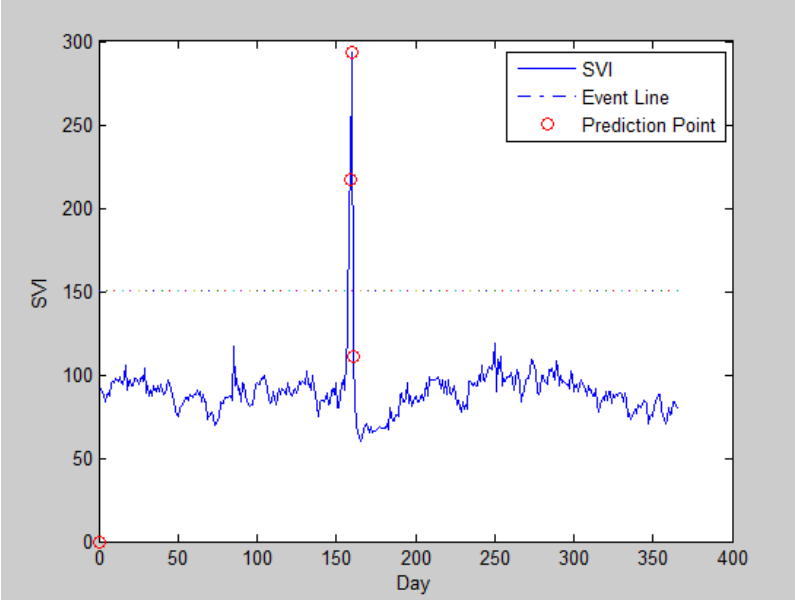

Figure 4.18: Testing Result of Battery A in 2006 with Q = 2

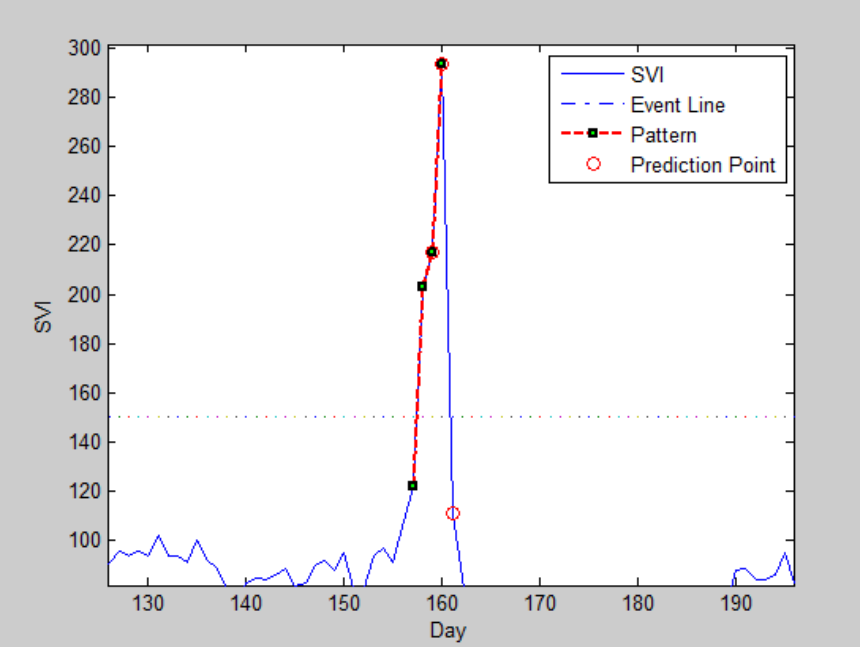

Figure 4.19: Pattern Plot for Testing Result of SVI in 2006 for Battery A with Q = 2

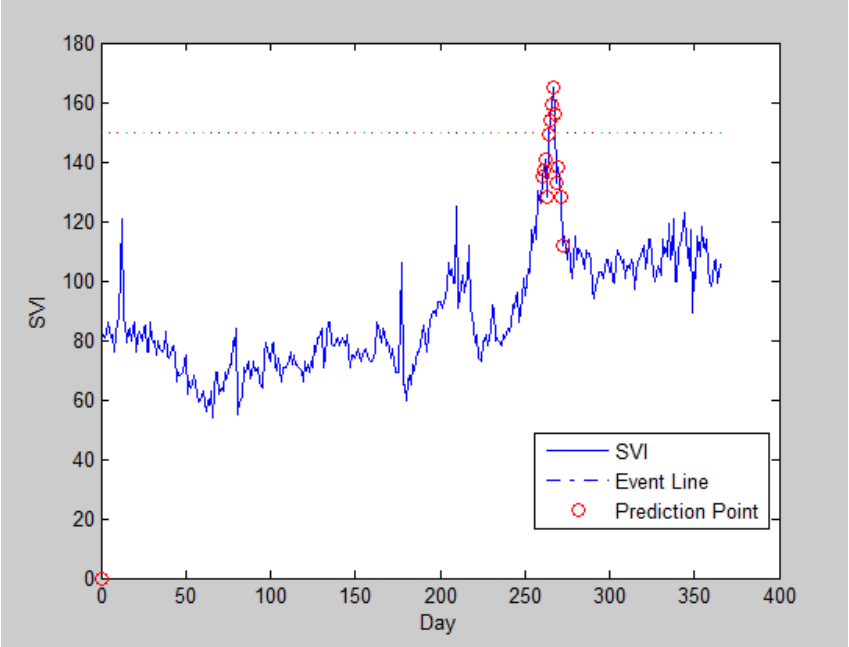

Figure 4.20: Testing Result of Battery A in 2007 with Q = 2

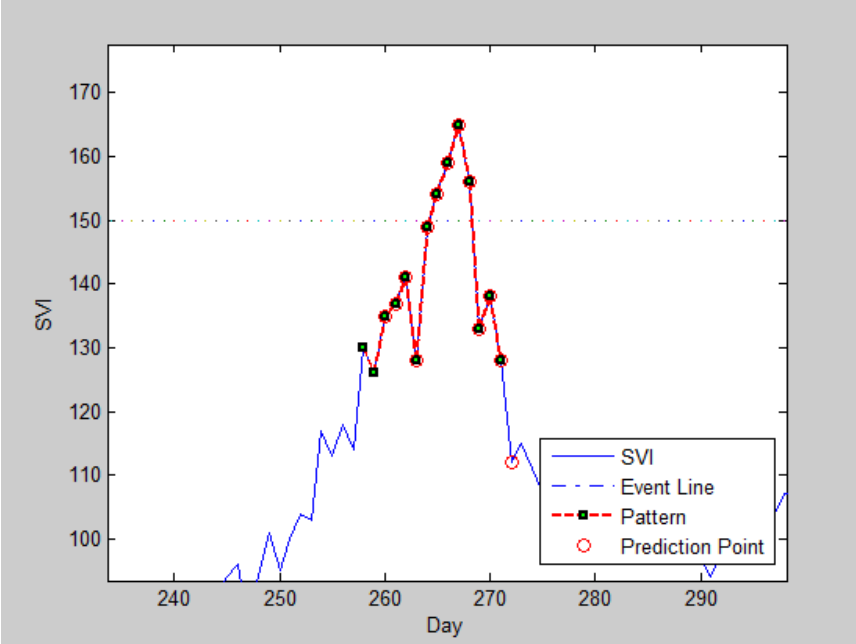

Figure 4.21: Pattern Plot for Testing Result of SVI in 2007 for Battery A with Q = 2

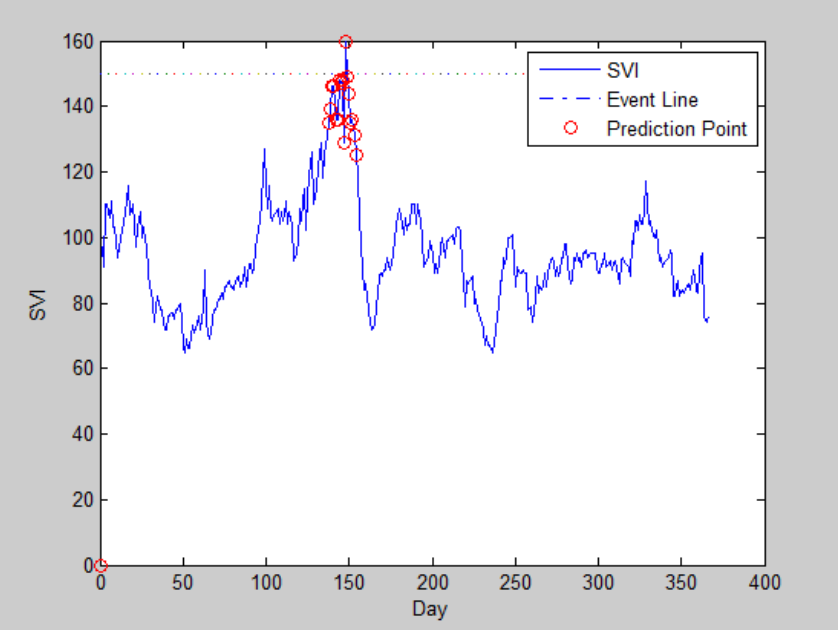

Figure 4.22: Testing Result of Battery A in 2008 with Q = 2

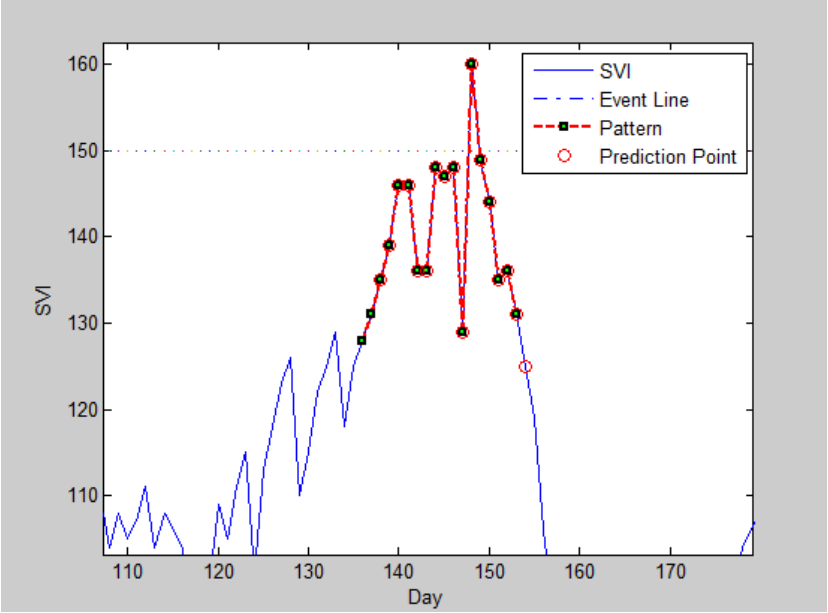

Figure 4.23: Pattern Plot for Testing Result of SVI in 2008 for Battery A with Q = 2

Table 4.28: Testing Result of SVI for Battery A with Q = 2

Training set data	Testing Year	Total number of events	Number of detected patterns	Correct predictions	Correct Percentage	Accuracy percentage
2002-2005	2006	3	3	2	66.67%	66.67%
2002-2006	2007	4	13	4	30.77%	100%
2002-2007	2008	1	17	1	5.88%	100%

4.4.3. Results of SVI Test for Battery C

The initial parameters of the improved TSDM method are same as in Table 4.27. Two different data combination tests are performed: (A) Training set: 2002 to 2005, Testing set: 2006; (B) Training set: 2002 to 2007, Testing set: 2008. No sludge bulking event occurred in 2007 and 2009 (Table 4.13). Figures 4.24 and 4.25 and Table 4.29 show the testing results for two data combination tests. Unfortunately, no prediction point is found by the improved TSDM method for Battery C as shown in Figures 4.24 and 4.25.

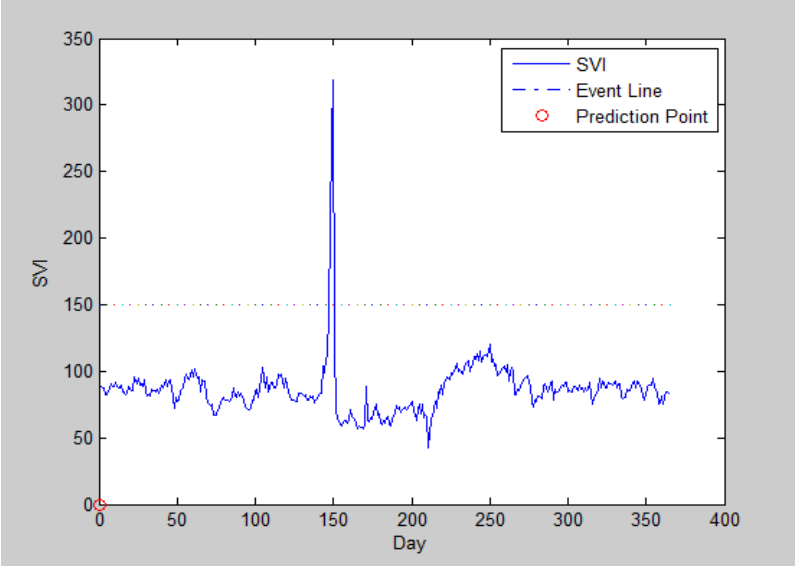

Figure 4.24: Testing Result of Battery C in 2006 with Q = 2

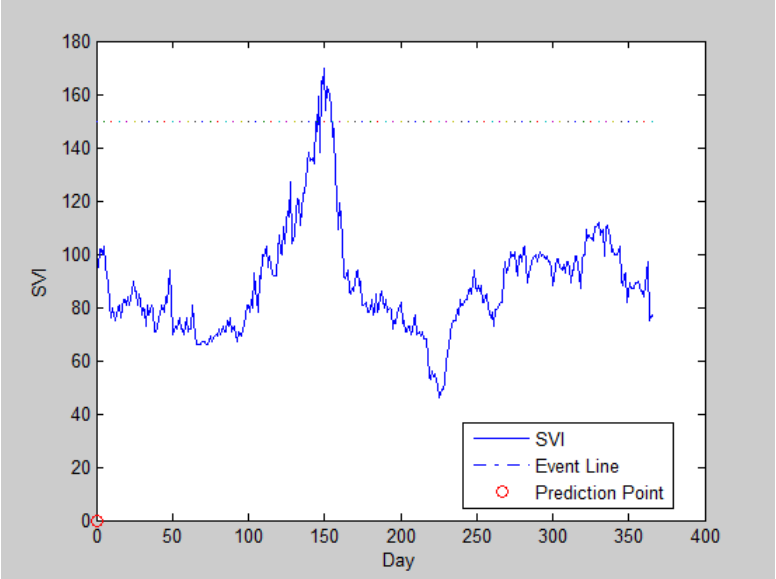

Figure 4.25: Testing Result of Battery C in 2008, Q = 2

Table 4. 29: Testing Result of SVI for Battery C, Q = 2

Training set data	Testing Year	Total number of events	Number of detected patterns	Correct predictions	Correct Percentage	Accuracy percentage
2002-2005	2006	3	0	0	0%	0%
2002-2007	2008	8	0	0	0%	0%

I. Method Performance Improvement for Testing of Battery C

According to Section 4.3, if the testing results are not acceptable, the analyst should consider reducing the embedding dimension, Q, to get better testing results. In the previous testing for Battery C, Q is 2, which cannot be reduced any further.

Another approach is to reduce the event value to enlarge the number of temporal pattern points in the training process. So the event value is set to 120 mL/g, and then the improved TSDM method was rerun. Figures 4.26 to 4.29 and Table 4.28 show the results of the improvement approach. From Figures 4.26 to 4.29, it can be seen that the prediction results are better after reducing the event value to 120 mL/g. Meanwhile, the accuracy percentage gets higher for both testing years of 2006 and 2008. Even when the event value is set to 120 mL/g instead of 150 mL/g, it does have a possible useful warning effect for sludge bulking problems for the WWTP. Also, the sludge bulking period in test year 2006 and 2008 can be prevented because the TSDM method can provide warning information because of the high false positive prediction points before the occurrence of sludge bulking.

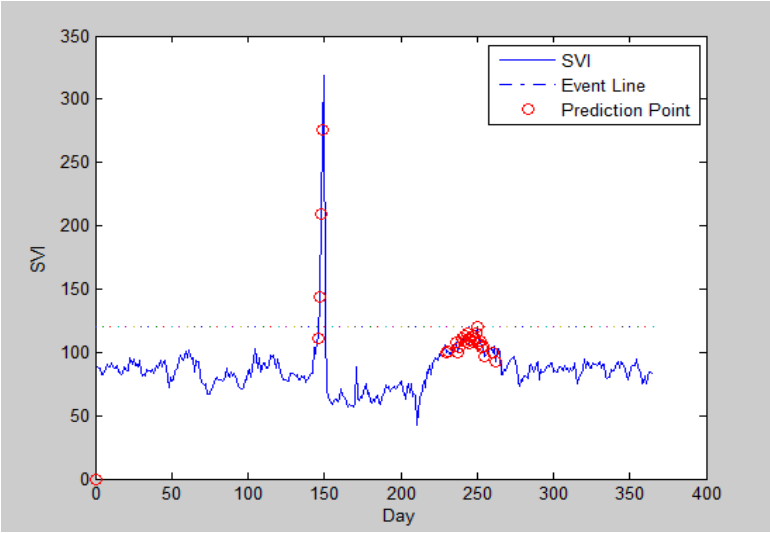

Figure 4.26: Testing Result of Battery C in 2006 with Q = 2 and event value = 120 mL/g

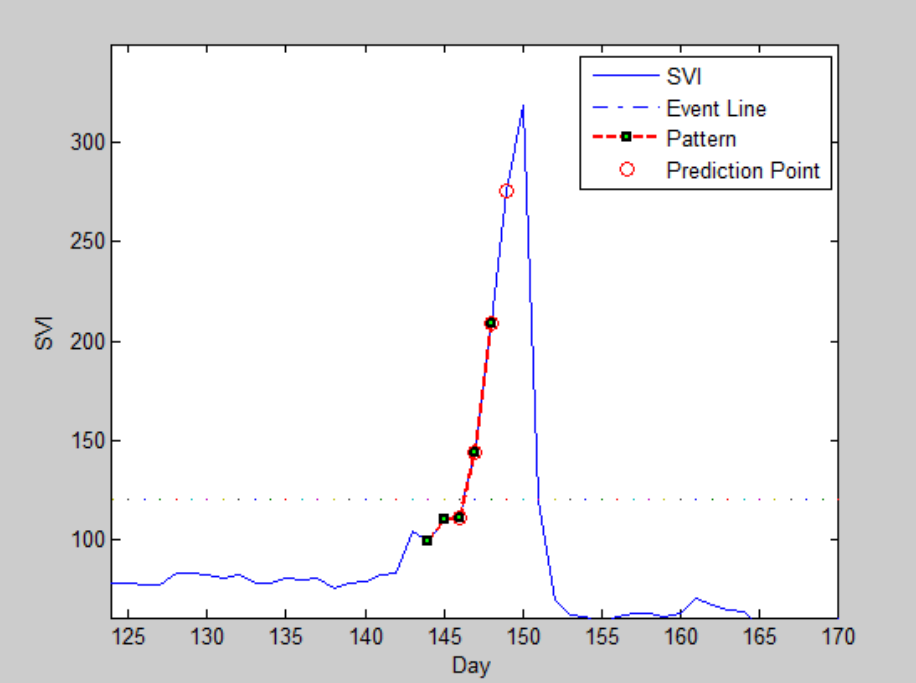

Figure 4.27: Pattern Plot for Testing Result of SVI in 2006 for Battery C with Q = 2 and event value = 120 mL/g

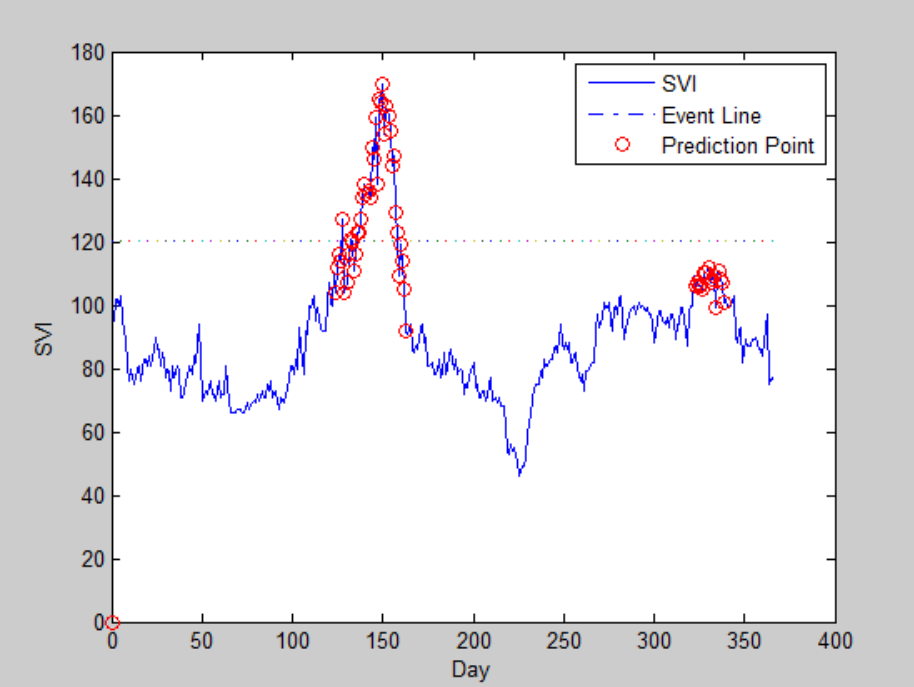

Figure 4.28: Testing Result of Battery C in 2008 with Q = 2 and event value = 120 mL/g

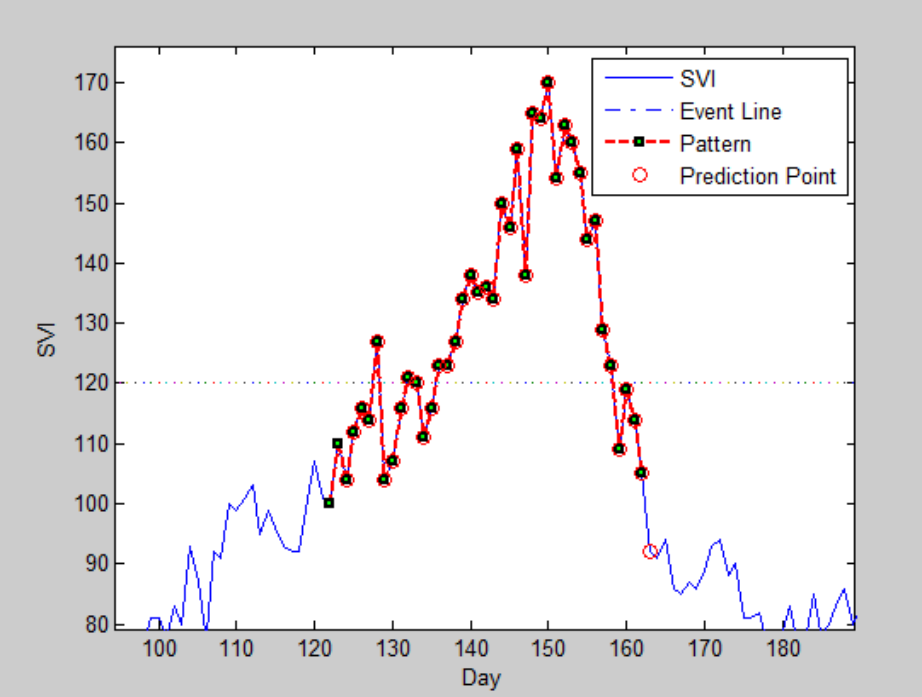

Figure 4.29: Pattern Plot for Testing Result of SVI in 2008 for Battery C with Q = 2 and event value = 120 mL/g

Table 4.30: Improved Testing Results of Battery C with event value = 120 mL/g

Training set data	Testing Year	Total number of events	Number of detected patterns	Correct predictions	Correct Percentage	Accuracy percentage
2002-2005	2006	4	27	3	11.11%	75%
2002-2007	2008	25	56	25	46.64%	100%

4.5. Discussion and Conclusions

From the results presented in this chapter, it can be concluded that the improved Time Series Data Mining method can be applied to the sludge bulking problem for WWTPs. However, the results do not mean that the improved TSDM method can be applied to all data directly. Several aspects of the TSDM method should be considered and modified during the process of the actual application of the TSDM method to get the best results.

First, the embedding dimension, Q, must be carefully determined. From the

synthetic data test, the Q value can be calculated by the false nearest neighbor method (Q = 7), and the TSDM method can determine training temporal pattern clusters and prediction results using the calculated Q. But in the SVI data test, the calculated Q = 3 did not yield good results. So a modification for Q was required. From the training and testing results after reducing Q, better results could be obtained. The accuracy percentage was as high as 100%, which is a very high quality prediction accuracy percentage.

Second, the radius enlarge ratio for the temporal pattern clusters must also be carefully selected. From the analysis of the training process for the synthetic data, it can be seen that the enlarge ratio is crucial for the accuracy of the testing process. A large enlarge ratio will lead to a lower correct percentage, and a small enlarge ratio will lead to inadequate cluster sizes. So the enlarge ratio should be determined by several tests to generate a balanced situation.

Third, the data applied in the method must include a sufficient number of events for training purposes. From results presented in this chapter, it can be seen that the test of the synthetic and the SVI data could be acceptable. However, for the ammonia data test, even when the embedding dimension, Q, and enlarge ratio are modified, acceptable result could not be obtained. Another approach to improve the results is to lower the events value, which is shown as an example in the testing process of the SVI data of Battery C. However, it should be noted that the event value for the ammonia data is set to 2 mg/L, which is already a value lower than the permit limit of the WWTP. It is not meaningful if the event value of ammonia is set smaller than 2 mg/L. So the only explanation is the ammonia data are too chaotic for the improved TSDM method to yield useful results

when the event value is set to 2 mg/L.

In the meantime, warning information for the detection and prevention of sludge bulking periods is also a notable result of the improved TSDM method. As previously discussed, detecting the temporal pattern before the first point of the sludge bulking period is very important, because the pattern can provide warning information to the WWTP operator. It should be noted that not all sludge bulking periods for 3 Batteries are effectively predicted or indicated by pattern identification. In test year 2006 for Batteries A and C, the improved TSDM method failed to provide warning information for the sludge bulking event, because the SVI value has a sudden jump to the event line. For instance, for Battery A in test year 2006, the sludge bulking event happened with a sudden jump in the SVI value from 130 to 200 mL/g. Such a sudden jump in the SVI value cannot be detected by the improved TSDM method because no such jump happened previously in the training data for the improved TSDM method to learn.

Overall, the improved TSDM method can be applied to the real-world WWTP data. Because of the dynamic features of the real-world data, the components of the improved TSDM method should be modified as necessary to get good results. The short coming of the improved TSDM method is that an event can only be predicted by a completed pattern. From previous analysis, if the final point of a pattern is higher than the event value, the TSDM method fails to provide warning information to the WWTP operator. The HMMs method is introduced in next chapter, and it does not need a completed pattern to predict an event.

CHAPTER 5 ANALYSIS AND DETECTION OF SLUDGE BULKING PROBLEMS USING THE HIDDEN MARKOV MODELS (HMMs) METHOD

The principles and the general process of Hidden Markov Models (HMMs) method were described in Chapter 3. This chapter shows the results of the application of the HMMs method to the SVI data, and a discussion of these results also is made. As previously mentioned, some important parameters of the HMMs method should be set and modified before and during the process of training the HMMs to the SVI data. The determination of these parameters is given in each section.

5.1. Introduction

As previously mentioned in Table 4.13, Battery D has the fewest sludge bulking events, and it does not have any sludge bulking problems from 2006 to 2009. Thus, the SVI data for Battery D will not be tested for the HMMs method. So the HMMs method is applied to the SVI data of Batteries A, B, and C.

Like the improved TSDM method, the HMMs method also needs to learn the patterns and events in the training data set. So the SVI data of 2005 is selected to be included into the training data set for each battery. Also, if no sludge bulking event occurred in a certain year, the SVI data of this year will not be tested for sludge bulking, but the SVI data of this year will still be added into the training data set.

Before the testing process, several important components of the HMMs method should be considered: i.e. the length of pattern and the initial value for each state.

The length of pattern can be determined by the false nearest neighbor method as in the improved TSDM method. For example, the length of pattern is calculated as three for the SVI data in Battery A. At the same time, the normal state and event state should

be considered. So there are five states for the HMMs method. State 1 is the normal state, which means normal values for the SVI data. States 2, 3, and 4 are the pattern states; they demonstrate the pattern values for the SVI data. State 5 is the event state data which represents an event value.

The initial values for each state can be determined by the analyst. As previously mentioned, the reason why these values need to be set is sometimes the thresholds found by the Mixture of Gaussian function are not reasonable, e.g., maybe the SVI value of the event state is less than that for the normal state. So the analyst needs to check the values and reset them if they are unreasonable. In most cases, the initial values need to be reset, and they have been set to [80, 120, 120, 120, 200] which means [normal state value, pattern state point 1 value, pattern state point 2 value, pattern state point 3 value, event state]. These values were chosen carefully by the author based on the different tests for many times. Besides, the values for the normal and pattern states are reasonable, under normal operations most SVI values are around 80 mL/g and in the lead in to sludge bulking (pattern states) SVI value are about 120 mL/g. Although the initial value of the event state is higher than 150 mL/g, the HMMs method will adjust the event state value in the testing process.

5.2. Analysis of Test Results for Battery A

From Table 4.13, it can be seen that no sludge bulking event happened in 2009. Three tests will be performed for Battery A: first, training set 2002 to 2005, testing set 2006; second, training set 2002 to 2006, testing set 2007; and third, training set 2002 to 2007, testing set 2008.

5.2.1. Training set: 2002 to 2005; Testing set 2006

First, the length of pattern and initial value for each state need to be set. The length of the pattern state is calculated as three, so the total number of states is five. The initial value needs to be reset, so they are [80, 120, 120, 120, 200].

The transition probability matrix estimated in the training process is shown in Eq. 5.1. Note that the probability of a jump from normal state to the first pattern state is 0.0179 which is not a high probability. However, if the first pattern state is detected, the probabilities for continuing through the pattern states are 0.9998 and 0.9379, these are high values. Last, if the final pattern state occurs, the probability at a jump from pattern states to the event state is 0.9599. Combining the probabilities there is a 90.01% chance that once the first pattern state is detected the process will continue on to the event state. This means that the occurrence of the pattern state might comprise very valuable warning information for WWTP operators.

$$\hat{\mathbf{P}} = \begin{Bmatrix} 0.9821 & 0.0179 & 0 & 0 & 0 \\ 0.0002 & 0 & 0.9998 & 0 & 0 \\ 0.0621 & 0 & 0 & 0.9379 & 0 \\ 0.0401 & 0 & 0 & 0 & 0.9599 \\ 1 & 0 & 0 & 0 & 0 \end{Bmatrix} \tag{5.1}$$

Figure 5.1 shows the testing result of 2006 for Battery A. In Figure 5.1, the red line is the prediction line. It can be seen that the red line is straight when the HMMs method considered these values as normal state points. There was a sudden jump when the HMMs method detected the pattern and event state points.

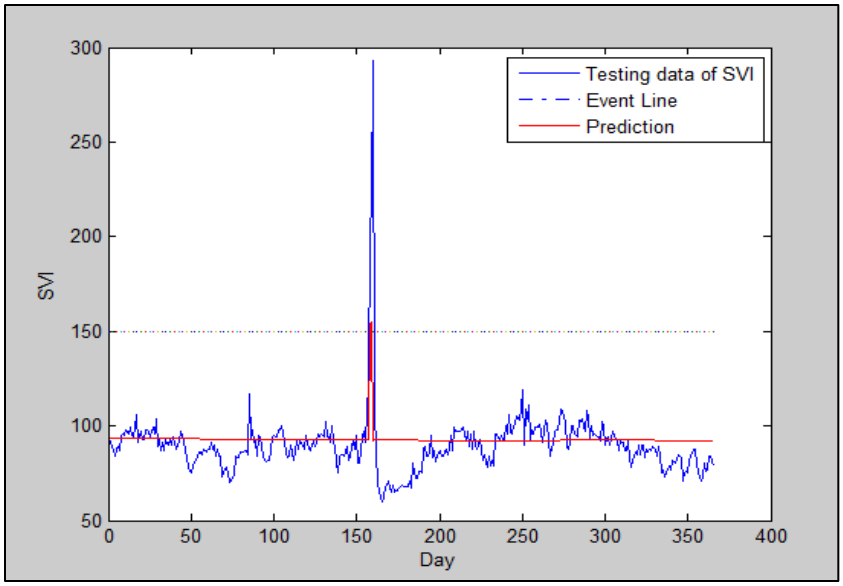

Figure 5.1: Testing result of 2006 for Battery A

Figure 5.2 demonstrates the pattern state points and event state point in testing data set of 2006. From Figure 5.2, it can be seen the sludge bulking event can be predicted, and the pattern points are detected before the event point. However, it should be noted that the values of the latter two pattern state points are higher than 150 mL/g. From the prediction point of view, the first of the pattern state points (which is less than 150 mL/g) can be detected by the HMMs method, indicating the warning ability of the HMMs method. From the WWTP operator point of view, if the HMMs method can find even first pattern state value, it can make the WWTP operator aware that there will be a high chance (90% probability for this battery) that the sludge bulking problem will happen according to the probabilities listed in the transition probability matrix. For the test year of 2006, the sludge bulking problem is detected effectively.

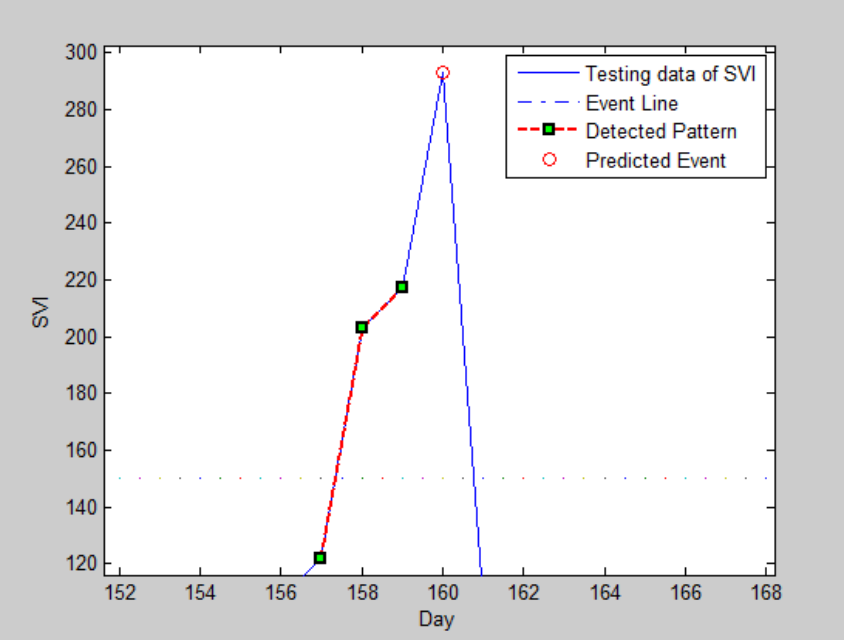

Figure 5.2: Predicted Pattern and Event Points in 2006 for Battery A

5.2.2. Training set: 2002 to 2006; Testing set 2007

In this data set, the length for the pattern states and initial values for each state are same as those in Section 5.2.1. The length of pattern state is three, and initial value for each state was set to [80, 120, 120, 120, 200].

Eq. 5.2 lists the transition probability matrix estimated in the training process. The probability of a jump from normal state to the pattern state is 0.0162, which is not a high value. But the probabilities of a jump to the next pattern states and the event state are very high. Combining the probabilities there is a 91.7% chance that once the first pattern state detected the process will continue on to the event state.

$$\hat{\mathbf{P}} = \begin{Bmatrix} 0.9838 & 0.0162 & 0 & 0 & 0 \\ 0.0006 & 0 & 0.9994 & 0 & 0 \\ 0.0492 & 0 & 0 & 0.9508 & 0 \\ 0.0352 & 0 & 0 & 0 & 0.9648 \\ 1 & 0 & 0 & 0 & 0 \end{Bmatrix} \tag{5.2}$$

Figure 5.3 shows the testing result of 2007 for Battery A. Figure 5.4

demonstrates the pattern state points and event state point in the testing data set of 2007. From Figure 5.3, it can be seen that the HMMs method made several event detections. These predicted events are concentrated on the period that the sludge bulking problem happened. Three events in Figure 5.4 were predicted by the HMMs method. The first and third ones are not true sludge bulking events. Only the second one is a true sludge bulking event. Technically, the prediction accuracy percentage is 33.33%. However, from the sludge bulking prevention point of view of the operator of the WWTP, the SVI data from day 260 to day 270 can be considered as a single long term sludge bulking event. For this reason, although the first event prediction is not correct, the beginning of the long term sludge bulking is predicted successfully by this first event.

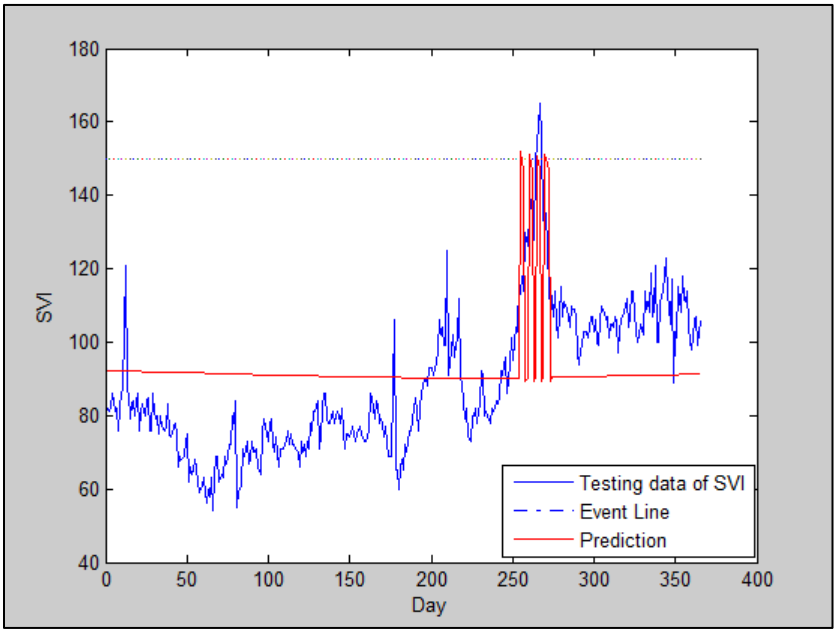

Figure 5.3: Testing result of 2007for Battery A

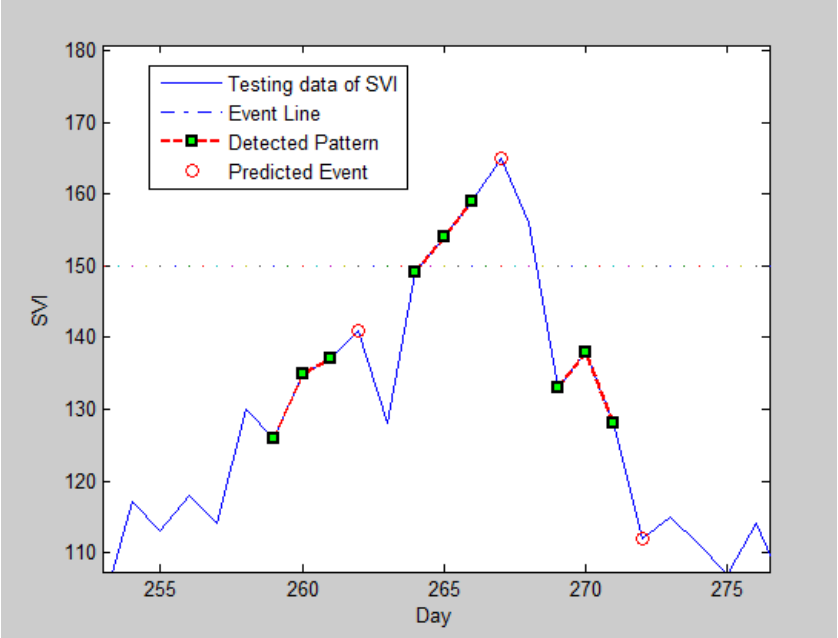

Figure 5.4: Predicted Pattern and Event Points in 2007 for Battery A

5.2.3. Training set: 2002 to 2007; Testing set 2008

The length of patterns and initial values for each state again were set to the same values as per Section 5.2.1.

The transition probability matrix estimated during the training set process is shown in Eq. 5.3. The probability of a jump from the normal state to the first pattern state is 0.0171, which is not a high value. But the probabilities of a jump to the next pattern states and the event state are very high. Combining the probabilities there is a 90.7% chance that once the first pattern state is detected the process will continue on to the event state.

$$\hat{\mathbf{P}} = \begin{Bmatrix} 0.9829 & 0.0171 & 0 & 0 & 0 \\ 0.0004 & 0 & 0.9997 & 0 & 0 \\ 0.0349 & 0 & 0 & 0.9651 & 0 \\ 0.0597 & 0 & 0 & 0 & 0.9403 \\ 1 & 0 & 0 & 0 & 0 \end{Bmatrix} \tag{5.3}$$

Figure 5.5 shows the testing result of 2008 for Battery A. Figure 5.6

demonstrates the pattern state points and event state point in the testing data set of 2008. From Figure 5.5, it can be clearly seen that the sludge bulking event around day 150 can be detected by the HMMs method. Also, from Figure 5.6, it should be noted that the sludge bulking event around day 150 is predicted as a pattern state by the HMMs method. This means these three detected sludge bulking events are false positive prediction points. However, from the sludge bulking prevention point of view, the first false positive event happened before the real sludge bulking event, which means the long term sludge bulking might be prevented because this result can warn the WWTP operator to check the operation condition of the plant and give a warning of potential sludge bulking.

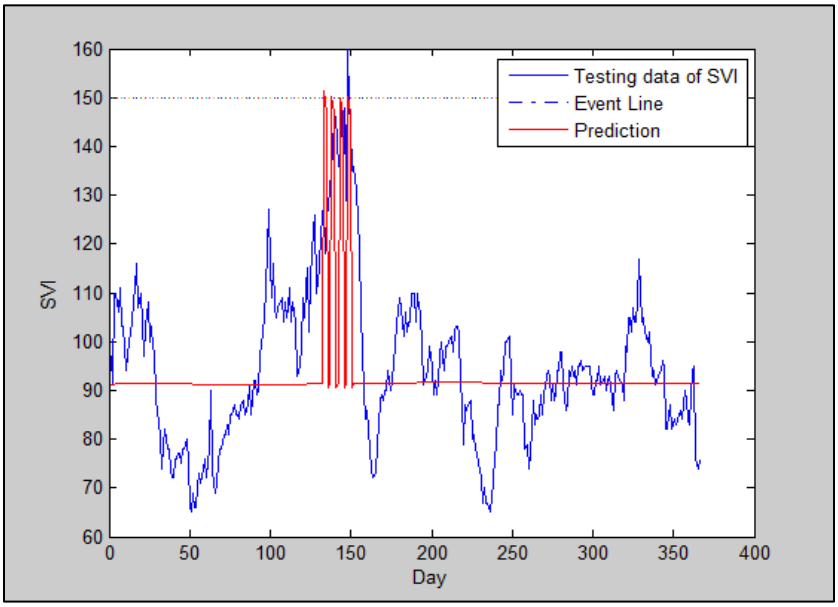

Figure 5.5: Testing result of 2008 for Battery A

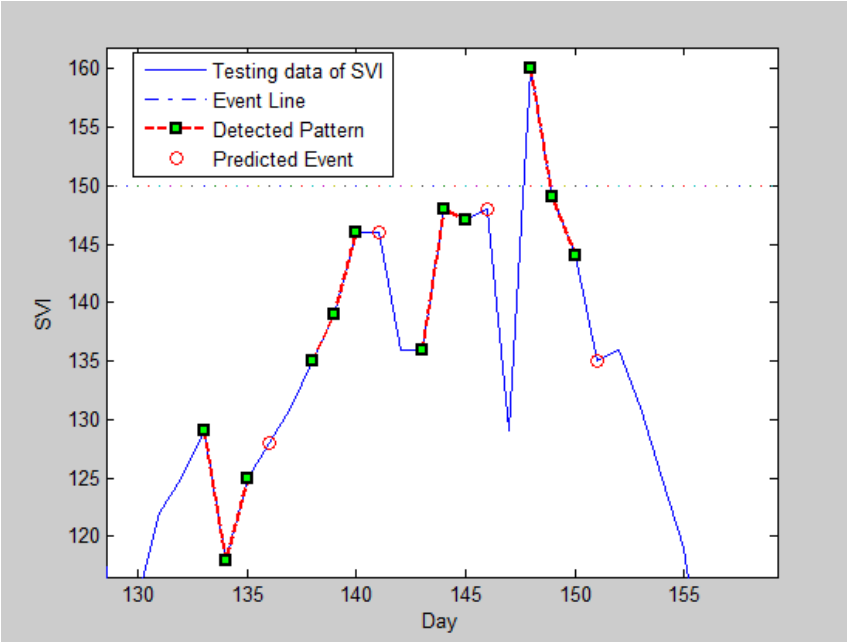

Figure 5.6: Detected Pattern and Predicted Event Points in 2008 for Battery A

5.3. Analysis of Test Results for Battery B

Four tests were performed for Battery B: first, the training set is from 2002 to 2005, the testing set is 2006; second, the training set is from 2002 to 2006, the testing set is 2007; third, the training set is from 2002 to 2007, the testing set is 2008; and last, the training set is from 2002 to 2008, the testing set is 2009.

5.3.1. Training data set: 2002 to 2005; Testing data set 2006

First, the length of pattern and initial value for each state need to be set. The length of pattern state is calculated as three, so the total number of states is five. The initial value for each state needs to be reset, so they are set to [80, 120, 120, 120, 200].

The transition probability matrix that estimated during the training process is shown in Eq. 5.4. The probability of a jump from normal state to pattern state is 0.0238, which is not a high value. It demonstrates that most of the SVI data are normal state points. But the probabilities of a jump to the next pattern states and the event state are

very high. Combining the probabilities there is an 80% chance that once the first pattern state is detected the process will continue on to the event state.

$$\hat{\mathbf{P}} = \begin{Bmatrix} 0.9762 & 0.0238 & 0 & 0 & 0 \\ 0.0028 & 0 & 0.9972 & 0 & 0 \\ 0.0106 & 0 & 0 & 0.9894 & 0 \\ 0.1890 & 0 & 0 & 0 & 0.8110 \\ 1 & 0 & 0 & 0 & 0 \end{Bmatrix} \tag{5.4}$$

Figure 5.7 shows the testing result of 2006 for Battery B, and Figure 5.8 shows the pattern state points and event state point in the testing data set of 2006. Figure 5.7 demonstrates the HMMs method is capable to detect the patterns for the sludge bulking event, and make a prediction for the highest SVI value. However, the prediction result has a problem in that the first detected pattern state point on day 149 is already a sludge bulking event prior to the event. The HMMs method failed to detect the patterns and sludge bulking event before the event occurred. But it should be noted that the cause of this problem is the same as found in the Section 5.2.1, that there is a sudden jump in the SVI value from day 148 to 149. So it can be concluded that the HMMs method lacks the capability to detect the states of a sudden jump values in SVI values to the event state.

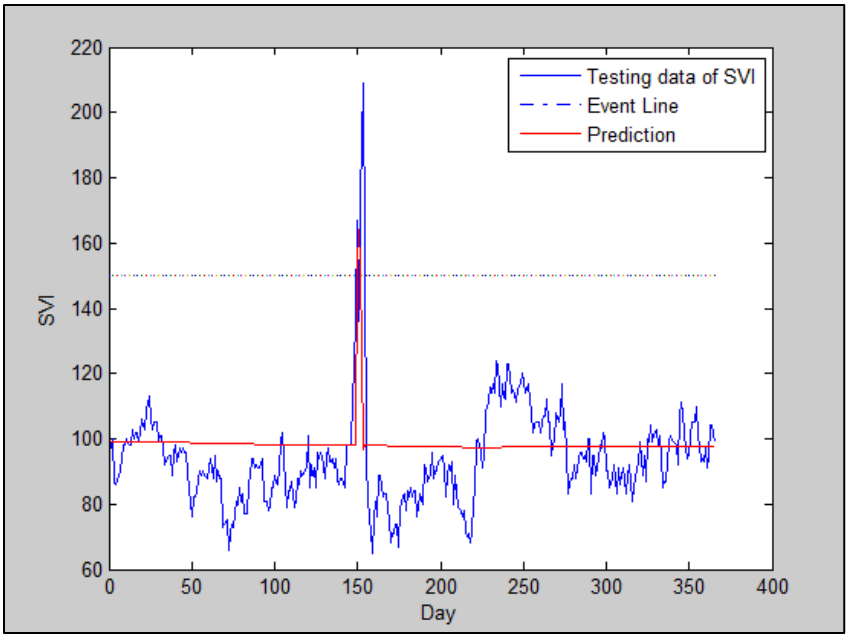

Figure 5.7: Testing result of 2006 for Battery B

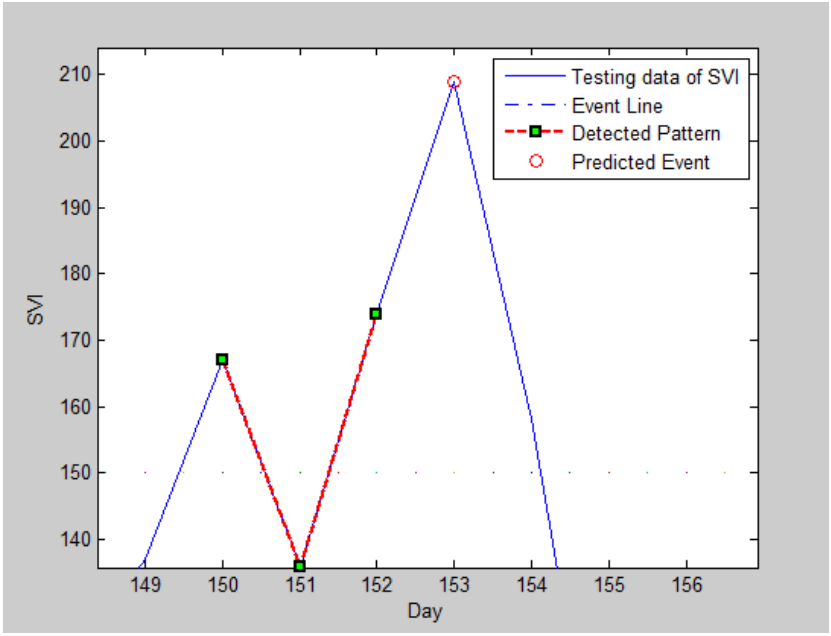

Figure 5.8: Predicted Pattern and Event Points in 2006 for Battery B

5.3.2. Training set: 2002 to 2006; Testing set 2007

First, the length of pattern and initial value for each state were set to the same values as in Section 5.3.1. The initial values for each state need to be reset, so they are [80, 120, 120, 120, 200].

The transition probability matrix estimated during the training process is shown in Eq. 5.5. The probability of a jump from normal state to pattern state is 0.0232, which is not a high value. This demonstrates that most of the SVI data are normal state points. But the probabilities of a jump to the next pattern states and the event state are very high. Combining the probabilities there is an 86.14% chance that once the first pattern state is detected the process will continue on to the event state.

$$\hat{\mathbf{P}} = \begin{Bmatrix} 0.9768 & 0.0232 & 0 & 0 & 0 \\ 0.0040 & 0 & 0.9960 & 0 & 0 \\ 0.0079 & 0 & 0 & 0.9921 & 0 \\ 0.1282 & 0 & 0 & 0 & 0.8718 \\ 1 & 0 & 0 & 0 & 0 \end{Bmatrix} \tag{5.5}$$

Figure 5.9 shows the testing result of 2007 for Battery B, and Figure 5.10 shows the pattern state points and event state points in the testing data set of 2007. From Figures 5.9 and 5.10, it can be seen that the HMMs method made several sludge bulking event predictions. Two long term sludge bulking events were predicted by the HMMs method as shown in Figure 5.10. Technically, the prediction accuracy percentage is 33.33%. However, from the sludge bulking prevention point of view of the WWTP operator, both two long term sludge bulking events can be predicted. For this reason, although the first event prediction is not correct, the beginning of the long term sludge bulking is predicted successfully by this first false positive event.

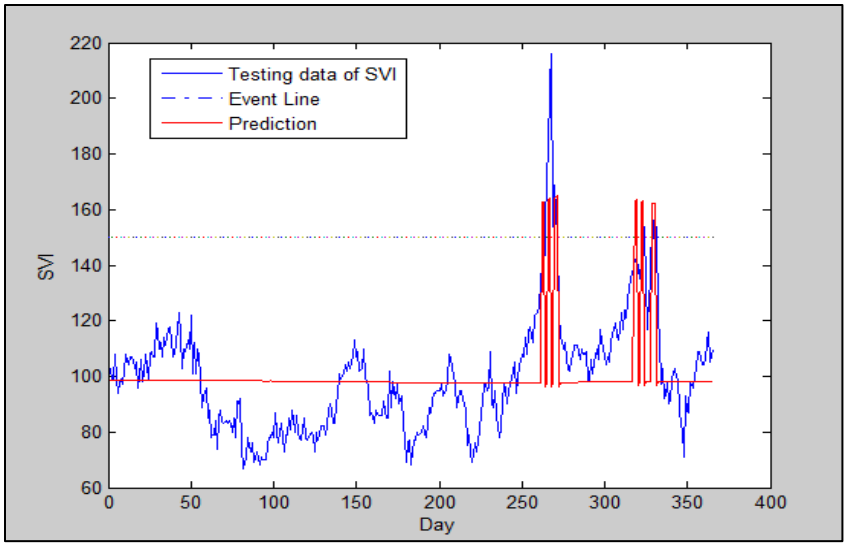

Figure 5.9: Testing result of 2007 for Battery B

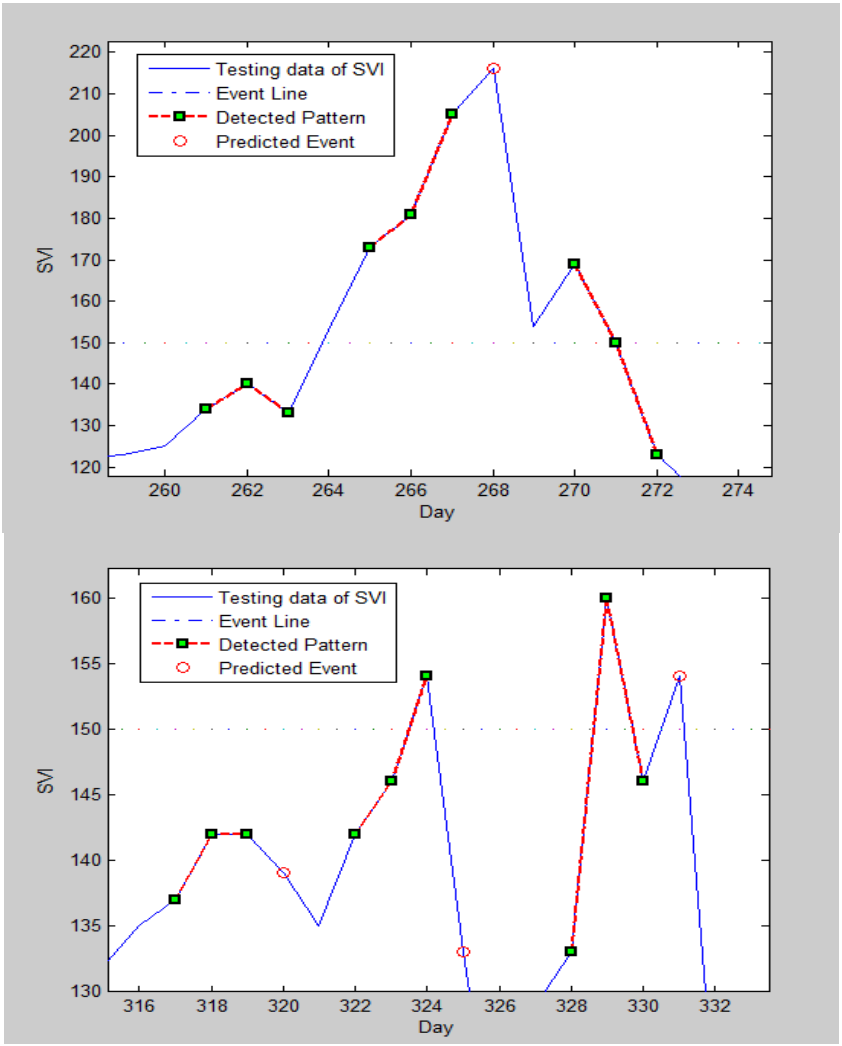

Figure 5.10: Predicted Pattern and Event Points in 2007 for Battery B

5.3.3. Training set: 2002 to 2007; Testing set 2008

First, the length of pattern and initial value for each state were set to the same values as in Section 5.3.1.

The transition probability matrix estimated during the training process is shown in Eq. 5.6. The probability of a jump from normal state to pattern states is 0.0224, which is not a high value, and this also demonstrates that most of the SVI data are normal state points. The probabilities of a jump to the next pattern states and the event state are very high. It means if the HMMs method detects the first point of the pattern states, there is a high probability (88.67%) that sludge bulking (event state) will occur.

$$\hat{\mathbf{P}} = \begin{Bmatrix} 0.9776 & 0.0224 & 0 & 0 & 0 \\ 0.0016 & 0 & 0.9984 & 0 & 0 \\ 0.0087 & 0 & 0 & 0.9913 & 0 \\ 0.1039 & 0 & 0 & 0 & 0.8961 \\ 1 & 0 & 0 & 0 & 0 \end{Bmatrix} \tag{5.6}$$

Figure 5.11 shows the testing result of 2008 for Battery B. Figure 5.12 demonstrates the pattern state points and the event state point in the testing data set of 2008. The sludge bulking events can be detected by the HMMs method. The first prediction point is a false positive result. However, in the next two days, a sludge bulking event occurred after the first false positive result. So this false positive result could send warning information to the WWTP operator to check the plant operation systems to avoid sludge bulking problems.

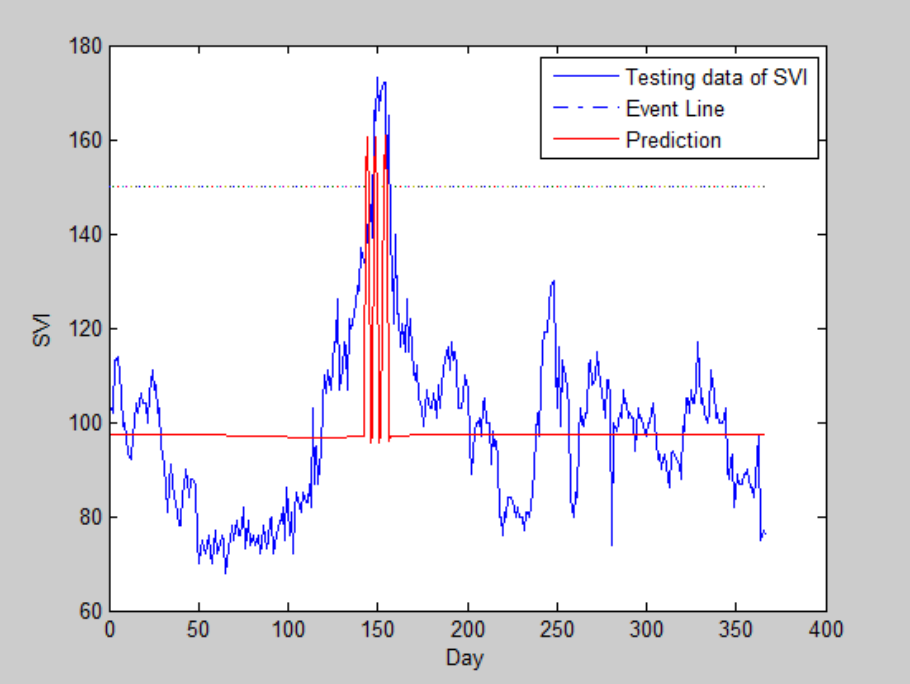

Figure 5.11: Testing result of 2008 for Battery B

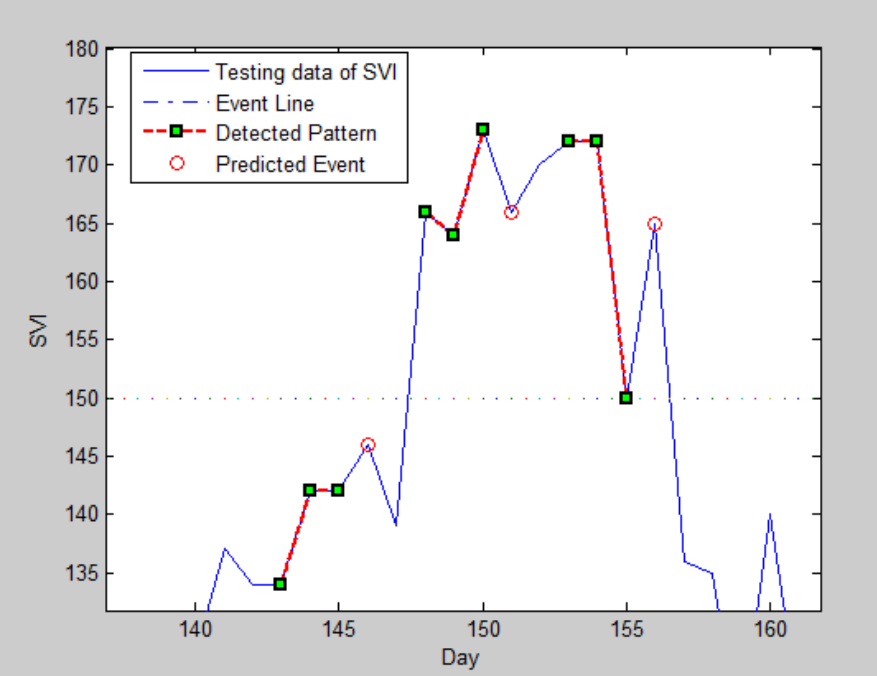

Figure 5.12: Predicted Pattern and Event Points in 2008 for Battery B

5.3.4. Training set: 2002 to 2008; Testing set 2009

First, the length of patterns and initial values for each state were set to the same values as in Section 5.3.1.

The transition probability matrix estimated during the training process is shown in Eq. 5.7. The probability of a jump from the normal state to the first pattern state is 0.0231, which is not a high value. This also reveals that most of the SVI data are normal state points. The probabilities of a jump to the next pattern states and the event state are very high. This means if the HMMs method detects the first point of the pattern state, there is a high probability (91.4%) that sludge bulking (event state) will occur.

$$\hat{\mathbf{P}} = \begin{Bmatrix} 0.9769 & 0.0231 & 0 & 0 & 0 \\ 0.0003 & 0 & 0.9997 & 0 & 0 \\ 0.0039 & 0 & 0 & 0.9961 & 0 \\ 0.0826 & 0 & 0 & 0 & 0.9174 \\ 1 & 0 & 0 & 0 & 0 \end{Bmatrix} \tag{5.7}$$

Figure 5.13 shows the testing result of 2008 for Battery B. Figure 5.14 demonstrates the pattern state points and event state point in the testing data set of 2009. From Figures 5.13 and 5.14, it can be seen that the sludge bulking events can be detected by the HMMs method. However, as previously discussed, several false positive points were detected before the sludge bulking event occurred, these false positive points still can warn the WWTP operator of an impending sludge bulking will happen considering these false positive prediction points are nearly to 150 mL/g. However, there are several false positive points after the highest SVI value, and these values are not useful for the WWTP operator.

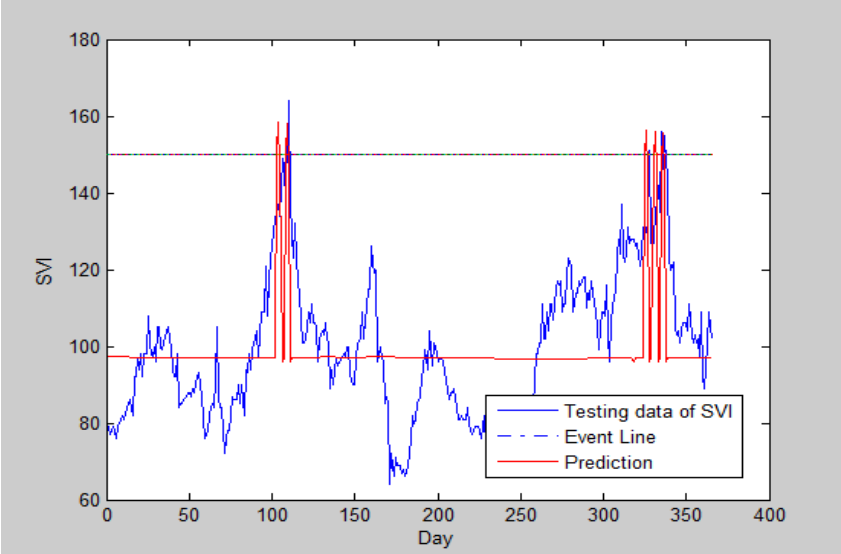

Figure 5.13: Testing result of 2009 for Battery B

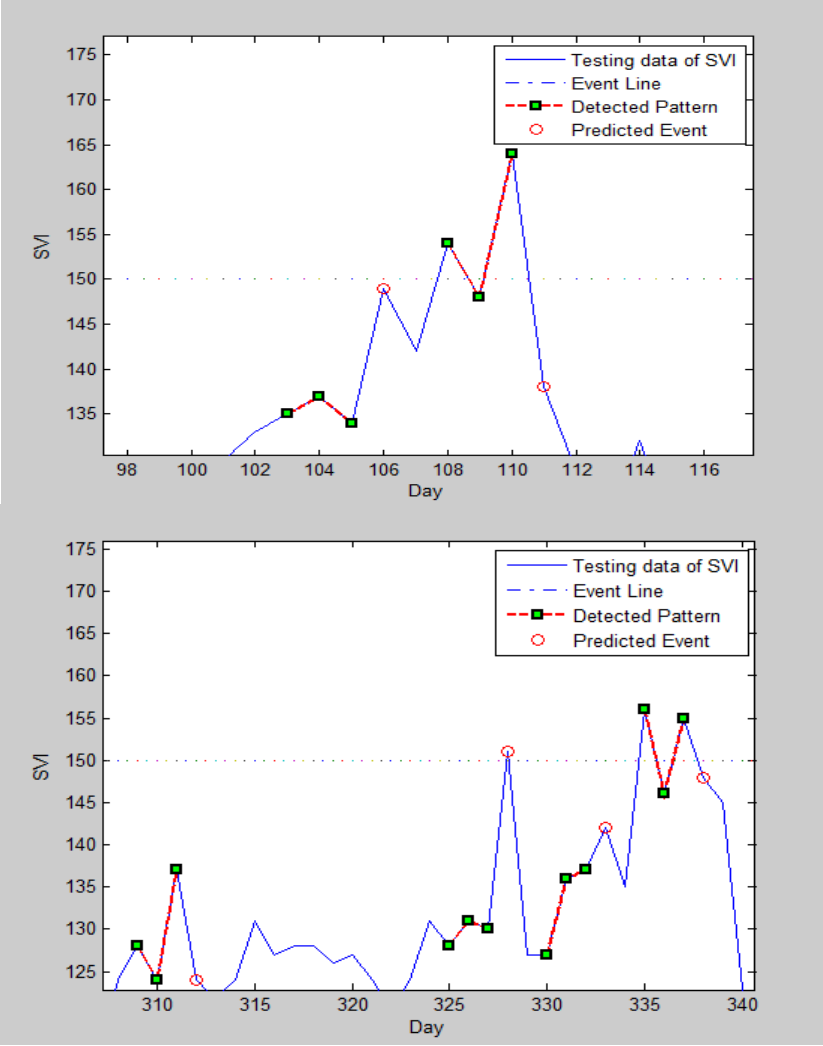

Figure 5.14: Predicted Pattern and Event Points in 2009 for Battery B

5.4. Analysis of Test Results for Battery C

For Battery C, from Table 4.13, sludge bulking problems only happened in 2006 and 2008. In this section, only the SVI data for 2006 and 2008 were tested.

5.4.1. Training set: 2002 to 2005; Testing set 2006

First, the length of pattern and initial values for each state need to be set. The length of the pattern state is calculated as 3, so the total number of states is 5. The initial values need to be reset, so they were set to [80, 120, 120, 120, 200].

The transition probability matrix estimated during the training process is shown in Eq. 5.8. The probability of a jump from normal state to the first pattern state is 0.0173 which means most of the SVI data stay in the normal state. But the probabilities of a jump to the next pattern states and the event state are very high. Combining the probabilities there is a 90.48% chance that once the first pattern state is detected the process will continue on to the event state.

$$\hat{\mathbf{P}} = \begin{Bmatrix} 0.9827 & 0.0173 & 0 & 0 & 0 \\ 0.0037 & 0 & 0.9963 & 0 & 0 \\ 0.0376 & 0 & 0 & 0.9624 & 0 \\ 0.0563 & 0 & 0 & 0 & 0.9437 \\ 1 & 0 & 0 & 0 & 0 \end{Bmatrix} \tag{5.8}$$

Figure 5.15 shows the testing result of 2006 for Battery C. Figure 5.16 demonstrates the pattern state points and event state point in the testing data set of 2006. From Figures 5.15 and 5.16, it can be seen that the HMMs method can predict the highest SVI value which is a sludge bulking event, but two event level SVI values are considered as pattern state points. The HMMs method detected the first pattern state point effectively, that reveals there is a high possibility for sludge bulking problems to occur.

So the HMMs method can warn the WWTP operator of possible impending sludge bulking.

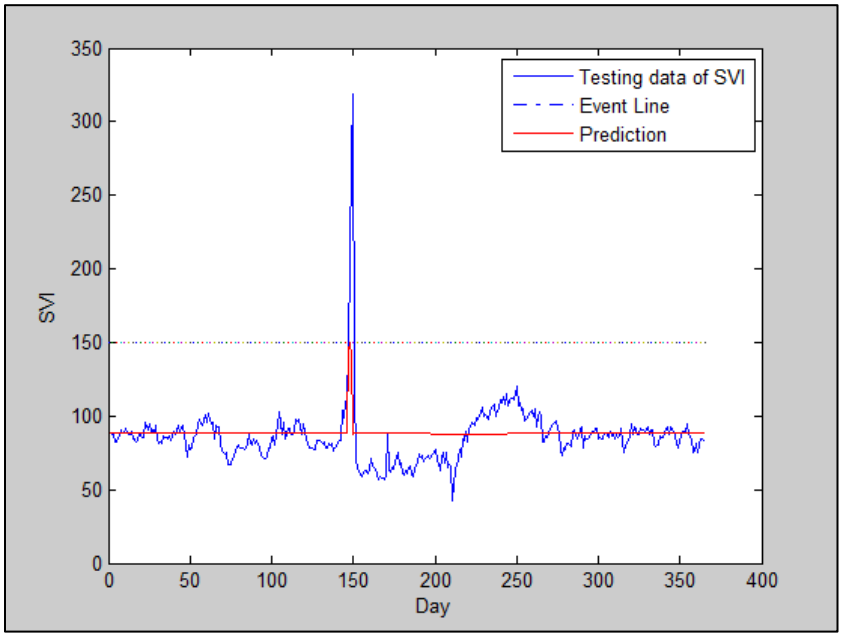

Figure 5.15: Testing result of 2006 for Battery C

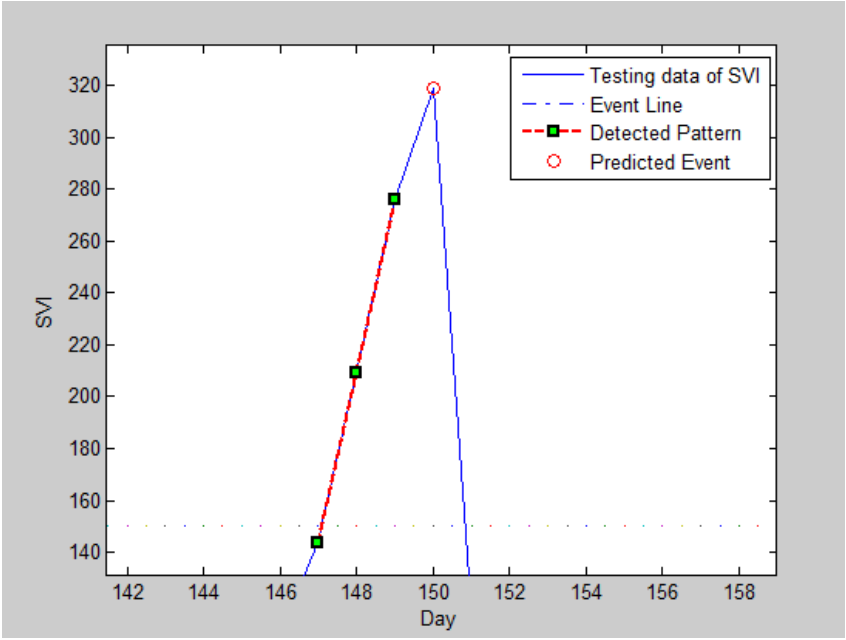

Figure 5.16: Predicted Pattern and Event Points in 2006 for Battery C

5.4.2. Training set: 2002 to 2007; Testing set 2008

First, the length of pattern and initial value for each state were set to the same

values as in Section 5.4.1.

The transition probability matrix estimated during the training set process is shown in Eq. 5.9. As previously mentioned, the probability of a jump from normal state to pattern state is 0.0194 which means most of the SVI data stay in the normal state. But the probabilities of a jump to next pattern states and the event state are very high. Combining the probabilities there is a 96.7% probability that once the first pattern state detected the process will continue on to the event state.

$$\hat{\mathbf{P}} = \begin{Bmatrix} 0.9806 & 0.0194 & 0 & 0 & 0 \\ 0.0035 & 0 & 0.9965 & 0 & 0 \\ 0.0053 & 0 & 0 & 0.9947 & 0 \\ 0.0249 & 0 & 0 & 0 & 0.9751 \\ 1 & 0 & 0 & 0 & 0 \end{Bmatrix} \tag{5.9}$$

Figure 5.17 shows the testing result of 2008 for Battery C. Figure 5.18 demonstrates the pattern state points and event state point in the testing data set of 2008. From Figure 5.17, it can be seen that the HMMs method can predict the sludge bulking event in 2008 for Battery C. In Figure 5.18, it can be seen that there are three false positive prediction points before the sludge bulking occurred. From the prediction accuracy point of view, these false positive predictions are not useful. However, from the WWTP operator point of view, since the HMMs method kept detecting pattern states and event states points and these state points are all high SVI values, the predictions should make the operator aware that sludge bulking may be impending.

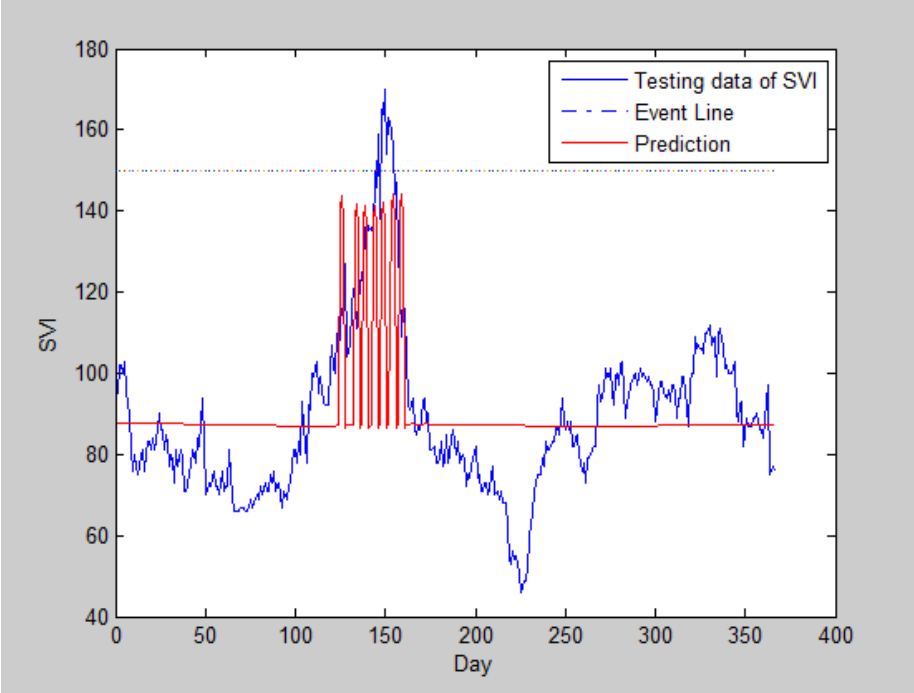

Figure 5.17: Testing result of 2008 for Battery C

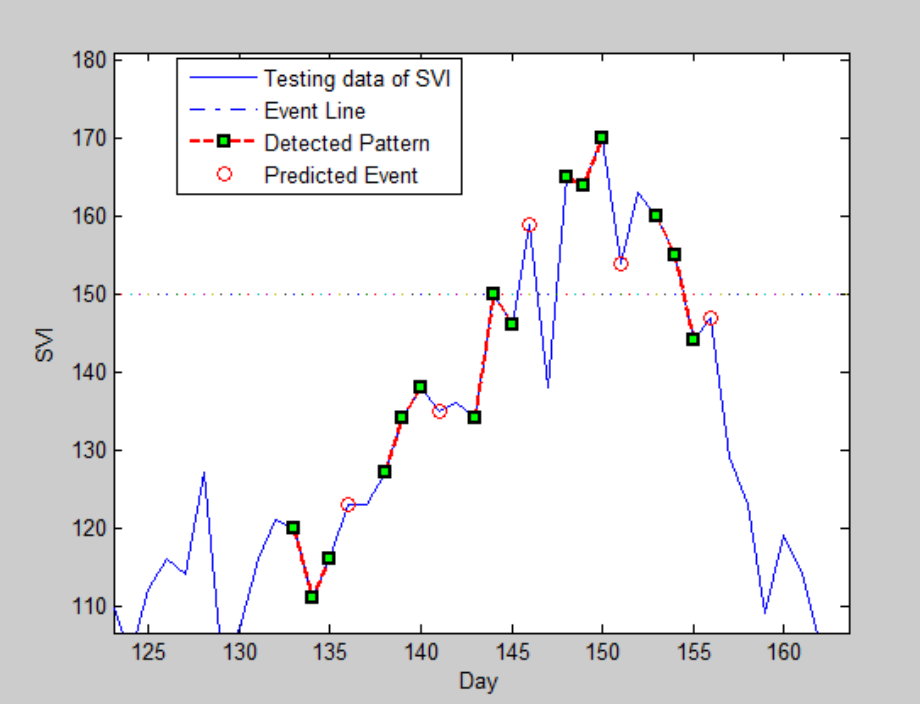

Figure 5.18: Predicted Pattern and Event Points in 2008 for Battery C

5.5. Discussion

From the results for the three batteries, it can be concluded that the HMMs

method can detect the patterns and predict the events for sludge bulking. However, from the testing results, several issues should be noticed.

First, for the SVI data of 2006 for each battery, the HMMs method only detect the highest SVI value, and considered the other events (SVI > 150 mL/g) as pattern states. All three batteries have the same result. However, by looking at the SVI data in 2006 for the three batteries, it should be noted that there was a sudden jump of the SVI data. For instance, the SVI data of day 157 in Battery A was 122 mL/g, and the SVI data of day 158 is 203 mL/g. There was as increase of 66.39% from day 157 to day 158 for the SVI data. The HMMs method only can detect the event after a complete pattern, so that's the reason the HMMs method detected the SVI data of day 158 as a pattern state. Meanwhile, it should be mentioned that such a sudden jump only happened in SVI data once in all 8 years of SVI data for each battery.

Second, sometimes the HMMs method cannot detect the event state points. The probability of the event state in the transition probability matrix is set to 0, because the transition probability matrix did not converge for the event state. Also, it was found that the results of the HMMs method are sensitive to the initial value for each state which needs to be set before the testing process. For example, for the testing process of Battery B in 2008, at first the transition probability matrix did not converge on the probability for the event state as shown in the following equation.

$$\hat{\mathbf{P}} = \begin{Bmatrix} 0.9375 & 0.0625 & 0 & 0 & 0 \\ 0.0842 & 0 & 0.9158 & 0 & 0 \\ 0.0165 & 0 & 0 & 0.9835 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{Bmatrix} \tag{5.10}$$

After modifying the initial value for the event state to 250 mL/g from 200mL/g and rerunning the HMMs method several times, the probability of the event state was found by the HMMs method as shown in Eqn. 5.6. Such results reveal that the HMMs method is still a developing method, and these limitations need to be investigated in the future.

Finally, this chapter focuses on the detection and prevention of the sludge bulking problem. From the prediction accuracy point of view, the false positive event points yielded by the HMMs method are not useful, because they are all below 150 mL/g. However, from the sludge bulking prevention point of view, these points are meaningful and useful. They are at a high SVI level, nearly to the sludge bulking event line of 150 mL/g. Those false positive points can send warning information to the WWTP operator. Also, even with the late prediction of event state, the detection of the first pattern state also can send warning information. As previously discussed, once the first pattern state was detected, there was high probability (>80% in all cases, most >90%) the event state (sludge bulking) would be occured. So the WWTP operator could apply some approaches to prevent the occurrence of sludge bulking if this warning information is received.

CHAPTER 6 ANALSIS FOR THE APPLICATION OF THE COMBINED METHOD OF HIDDEN MARKOV AND MULTINOMIAL LOGISTIC REGRESSION (MLR) MODEL

In Chapter 4 and Chapter 5, the improved TSDM method and HMMs method were applied to find hidden patterns to detect the sludge bulking problem. These two methods were applied to the SVI data itself. It may be meaningful and useful if the sludge bulking problem can be detected or predicted using data on other wastewater quality parameters. In the literature review in Chapter 2, it is discussed that some studies had discovered certain wastewater quality parameters might impact the sludge bulking problem.

In this chapter, first some wastewater quality parameters that have previously been proposed to have more impact on the sludge bulking problem than other parameters are selected for analysis. Second, the selected parameters and the SVI states data that obtained by the HMMs method are applied to build the multinomial logistic regression model. The built multinomial logistic regression model is applied to predict the future SVI data pattern and event states using other wastewater quality parameters. The performances of the combined method for different batteries are reported, and the results are presented and discussed from different aspects.

6.1. Selection of Wastewater Quality Parameters

In Chapter 2, some studies concluded that the following wastewater quality parameters have more impact on the sludge bulking problem than other parameters including dissolved oxygen (Bhatla, 1967), pH and organic loadings (Yasuda, 1976), food to microorganism (F/M) ratio, and soluble biochemical oxygen demand (BOD)

(Metcalf & Eddy, 2003). For the data available from the NSWRP, there are four types of data: preliminary and solids data (total wastewater flow, air flow, total solids, etc.), treatment operational data for each battery (return flow, MLSS, etc.), nitrogen analysis data, and lab analysis data. Each of these data includes more than ten kinds of data, so more than 40 kinds of data are available. It is better to select chemical and physical wastewater quality parameters that are related to the wastewater treatment process. The parameters selected to test in this thesis are: the F/M ratio, sludge retention time (SRT), detention time, temperature, effluent pH, RSSS, MLSS, influent NH3, effluent NH3, influent DO, effluent DO, and influent BOD.

6.2. Preliminary Analysis of the SVI Data and Other Wastewater Quality Parameters

6.2.1. Identification of Normal and Abnormal States for Other Wastewater Quality Parameters by the HMMs method

Similar to the application of the HMMs method to the SVI data, the HMMs method is applied to find the normal state and abnormal state for other wastewater water quality parameters. The normal state means the hidden state for the parameters when the sludge bulking does not happen; the abnormal state means the hidden state for the parameters when the sludge bulking happens. State 1 represents the normal state, and state 2 represents the abnormal state.

However, it is hard to define the threshold value of normal and abnormal states for every parameter. First, there is no exact definition for the abnormal value for some parameters. For example, no research has been done on the relationship of sludge bulking and ammonia. Further the value range of ammonia when the sludge bulking

happens is unknown. Second, for different wastewater treatment plants, the abnormal value for some parameters may vary. For example, temperature and influent wastewater flow rate vary among different wastewater treatment plants.

In this thesis, two criteria are used to determine the threshold value for normal and abnormal states: the first is the common knowledge and research studies in wastewater treatment technology; the second is the Mixture-Gaussian and EM algorithm functions in the HMMs method. However, it should be noted that the second criterion is only used under the condition of the failure of the first criterion.

After testing by the HMMs method, the proposed values for the normal state and the abnormal state for several wastewater quality parameters in Battery A are determined as listed in Table 6.1.

Table 6.1: Proposed values of normal and abnormal state for some wastewater quality parameters in Battery A

Parameters	Threshold Value	
	Normal State	Abnormal State
Influent Flow	85 MGD	55 MGD
Influent Ammonia	8 mg/L	15 mg/L
Effluent Ammonia	0.5 mg/L	2 mg/L
F/M ratio	0.05	0.02
BOD5	100 mg/L	140 mg/L
Influent DO	8 mg/L	1 mg/L
Effluent DO	8 mg/L	5 mg/L

After the determination of the normal and abnormal states, the cross tabulation and the correlation function analysis for SVI states data and other wastewater quality parameters were analyzed. The goal is trying to find what parameters are highly correlated with the SVI data.

6.2.2. The Correlation Function and Cross-tabulation Analysis for Hidden States of the SVI and Other Parameters Data

There are 5 states for the SVI data, state 1 represents the normal state, states 2 to 4 represent the pattern states, and state 5 represents the event state. There are two states for the other wastewater quality parameters, state 1 for the normal state, and state 2 for the abnormal state. After the testing by the HMMs method, the cross-tabulation and the correlation analyses were performed for the hidden states of the SVI and other wastewater quality parameters as listed in Table 6.2.

Table 6.2: Cross-tabulation and correlation analysis for the hidden states of the SVI and other parameters in Battery A from 2002 to 2009

Parameter	Cross-tabulation with SVI states						Correlation	
	State	1	2	3	4	5	R	P
F/M Ratio	1	1	0	0	0	0	0.0123	0.5072
	2	1899	166	166	166	524		
Detention Time, hr	1	1013	32	38	43	197	0.069	0.0002
	2	1088	81	75	70	285		
RSSS	1	73	11	11	11	1	0.0634	0.0006
	2	1639	227	227	227	495		
MLSS	1	678	48	50	47	273	-0.0311	0.0926
	2	1177	83	81	84	401		
Influent Ammonia, mg/L	1	1888	139	141	142	492	0.069	0.0002
	2	57	12	10	9	32		
Effluent Ammonia, mg/L	1	122	11	10	9	1933	0.0937	0
	2	12	3	4	5	813		
Influent DO, mg/L	1	784	55	52	56	248	0.0803	0
	2	910	148	151	147	371		
Effluent DO, mg/L	1	854	111	111	114	465	0.0195	0.2914
	2	569	128	128	125	317		
Influent BOD, mg/L	1	1840	84	81	78	702	0.0707	0.0001
	2	66	5	8	11	47		

It should be noted that these wastewater quality parameters were analyzed

separately. As discussed in Chapter 5, the HMMs method has a stability problem which could cause the transition probability matrix to be different for different wastewater quality parameters tests and the classification of SVI data into different states may vary among HMM runs. Hence, the number of values in states 1-5 vary in Table 6.2 for the comparison with the states of the other wastewater quality and operational parameters.

From Table 6.2, for the cross tabulation analysis, it can be seen that no strong evidence shows which parameter has a strong correlation with the SVI states, and the data of the cross tabulation is very chaotic. For example, in the normal state (state 1) for the SVI data, the HMMs method considered most of the influent ammonia at the same time as the normal state (1888 points). But in the pattern states and event state for the SVI data, the HMMs method still considered most of the influent ammonia at the same time period as the normal state. A similar problem can be seen from effluent ammonia and influent BOD. It can be concluded that the results of the cross-tabulation analysis were not acceptable, and it did not provide useful information. For the correlation function analysis, also no strong evidence could be found regarding parameter has a strong correlation with the SVI. The R-values of all parameters are less than 10%. From the statistical point of view, these R-values are very low and unacceptable. However, if the R-values are arranged from higher to lower, and the P-values (significance) are considered, some wastewater quality parameters were found to have higher correlation than others, such as detention time, effluent ammonia, influent ammonia, influent DO, and influent BOD.

Due to the weak results of the foregoing analysis, a time delay was considered in the retesting of the cross-tabulation and the correlation function analyses. Considering

the wastewater treatment process is a biological and chemical reaction process, it is possible that the other parameters will influence the SVI data with a time delay. So a time delay was considered for the other wastewater quality parameters, such as 1, 2, 3, 4, 5, 6, etc. days. Then the cross-tabulation and the correlation function analysis were repeated. Unfortunately, the results with the time delay were not better than the result with no lag, and in some cases even worse. The cross-tabulation results were chaotic, and the R-values for the correlation function were no greater than 0.10. So it can be concluded that a time lag is not an important consideration.

6.2.3. The Correlation Function of Hidden States for Wastewater Quality Parameters and the SVI in 3 Conditions

As previously mentioned, no parameter was found have a strong correlation with the SVI data. However, during the previous analysis, the whole data set (SVI and other wastewater quality parameters) was used. So the idea to only consider the data in the pattern states and event state for the SVI data and the other parameters was formulated. Three different conditions were set for data extraction for the pattern states and event states. The entire data extraction process was performed by the HMMs method.

Condition A: Pattern states and event state in the SVI data, parameter data for the same time

The pattern states and event state for the SVI data first were found, and then the corresponding SVI data were extracted for the pattern states and event state. Then the wastewater quality parameters data for the same time were extracted. In this case two sets data are obtained and their length should be the same.

Condition B: Event state in the SVI data, parameter data for the same time

The event state for the SVI data first were found, then the corresponding SVI data

for the event state were extracted. After that, the parameter data were extracted for the same time. These two data sets should have the same length.

Condition C: Event state in the SVI data, parameter data for the same time with a time delay

The procedure of this condition is the same as for condition 3; the difference is that a time delay is added. Different time delays were tried, from 1 to 20 days. Again, as previously mentioned, the time delay has little impact. For this reason, a time delay of 3 was chosen for illustration in the following summary of results.

I. Correlation function analysis for the 4 conditions

After the extraction process for the four conditions, the correlation analysis was done as summarized in Table 6.3.

Table 6.3: Correlation function analysis for three conditions in Battery A

Correlation with SVI						
Condition	F/M ratio		Detention time		Effluent NH3	
	R	P	R	P	R	P
A	-0.0034	0.9433	0.069	***0.0324***	0.1617	***0***
B	0.0028	0.9796	0.1808	***0.0109***	***0.233***	***0***
C, lag = 3	0.0423	0.1853	0.1674	0.3672	***0.2635***	***0***
Condition	Influent DO		Effluent DO		Influent BOD	
	R	P	R	P	R	P
A	-0.1258	***0.002***	-0.1604	***0***	0.0486	0.2452
B	***-0.2602***	***0***	***-0.3231***	***0***	0.0625	0.1003
C, lag = 3	-0.1443	***0.0001***	***-0.3277***	***0***	0.0441	0.3012
Condition	Temp		Effluent PH		Influent NH3	
	R	P	R	P	R	P
A	0.0611	***0.0623***	-0.0595	0.0894	0.0626	***0.0085***
B	0.146	***0.0002***	-0.026	0.1656	***0.2118***	***0.0033***
C, lag = 3	0.1447	***0.0003***	-0.0987	0.0013	0.1563	0.3386

From Table 6.3, the R-values for the wastewater quality parameters are all no higher than 0.35. From the statistical point of view, this is not a useful result. However, if the R-values were checked for different conditions, it can be found that the R-value in

condition B is better than the others, that means the wastewater quality parameters data have stronger correlation with SVI event states data. Such important numbers were underlined and bold in Table 6.3. Meanwhile, considering the significance (P value), it should be noted that several parameters have higher statistical correlation than the others, such as influent DO, effluent DO, influent ammonia, effluent ammonia, and temperature.

6.3. Analysis of the Combined Method for Batteries A, B and C

Several categories of wastewater quality parameters are considered to have impact on the sludge bulking problem, according to the analysis in the last section. From the HMMs method, the state of each SVI data can be obtained. And there are five states for the SVI data: 1 for the normal state; 2, 3, and 4 for the pattern states; 5 for the event state. The training SVI states data and the training data of selected wastewater quality parameters are used as the training set to build the multinomial logistic regression (MLR) model. Then the MLR model uses the test data of selected wastewater quality parameters to predict the probability of the SVI states.

From Table 4.13 in Chapter 4, 2005 has more sludge bulking events than other years, so 2005 is still selected to be included into the training data set. Battery B has more sludge bulking events than other batteries. For this reason, the SVI data in Battery B was first studied in the test. Then the tests for Batteries A and C are performed. Battery D will not be analyzed because o sludge bulking events happened from 2006 to 2009 in Battery D. The test is firstly performed using MATLAB to illustrate the result for each battery in different test years. However, the MATLAB can only show the analysis output and do not give details of the MLR model. So the SPSS was used to show the details of the MLR model of the test.

6.3.1. Analysis of Battery B Using MATLAB

Four different data combination tests were performed: (A) Training set: 2002 to 2005, Testing set: 2006; (B) Training set: 2002 to 2006, Testing set: 2007; (C) Training set: 2002 to 2007, Testing set: 2008; and (D) Training set: 2002 to 2008, Testing set: 2009.

I. Analysis of Data Combination A for Battery B

Unlike the improved TSDM and HMMs methods, no initial parameter values need to be set before applying the MLR model in MATLAB. The hidden state of each SVI data from 2002 to 2005 is obtained by the HMMs method. Then the MLR model was built for the hidden states data and selected wastewater quality parameters data from 2002 to 2005. Once the model is built, the test data of wastewater quality parameters in 2006 will be input to the model to calculate the future state data for each SVI value in 2006.

It should be noted that there are five states for the SVI data, one to five. To better illustrate the predicted SVI data in the figure, each state value will be multiplied by 35. These new values are called simulated state SVI value. So the new simulated state value for state 1 is 35 mL/g, for state 2 is 70 mL/g, for state 3 is 105 mL/g, for state 4 is 140 mL/g, and for state 5 is 175 mL/g. The simulated state SVI value for state 5 is larger than 150 mL/g, which is an event value for the sludge bulking problem. The simulated state SVI value for state 4 is 140 mL/g, which is nearly to 150 mL/g to arouse the attention of the operator of the WWTP to avoid the sludge bulking problem. These simulated SVI values are listed in Table 6.4. Figure 6.1 shows the test results of SVI in 2006 for Battery B by the combined method. The red line is the simulated SVI data line, which is predicted by the multinomial logistic regression model in the combination method. From

Figure 6.1, it is clear that the prediction result is not useful. No sludge bulking problem was detected or warned. Also, numerous false positive predictions are included in the result.

Table 6. 4: Simulated SVI value for each state

State	Simulated SVI value (mL/g)
1	35
2	70
3	105
4	140
5	175

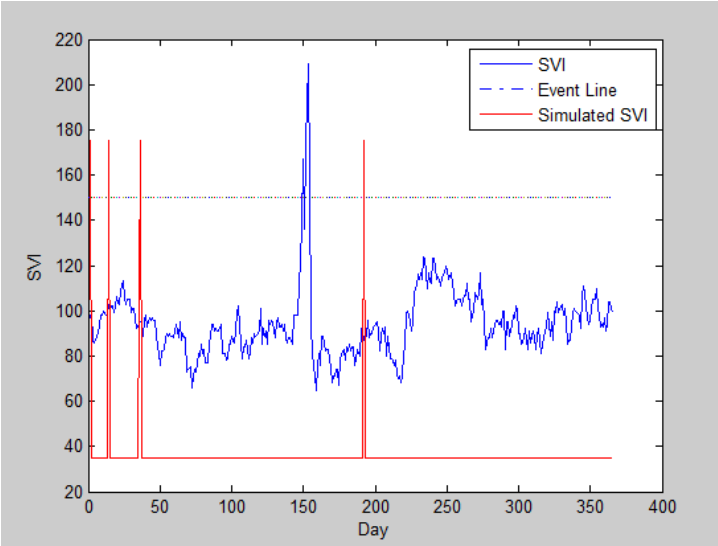

Figure 6.1: Test Result of SVI in 2006 for Battery B

II. Analysis of Data Combination B for Battery B

The process is the same as in Section 6.3.1.1. Figure 6.2 shows the test result for Battery B in 2007. The first two predicted events are false positive events. After that, the predicted events on day 269 and day 272 are true events, but they are detected after the sludge bulking problem has occurred. So the combined method failed to detect and prevent the sludge bulking problem in this case.

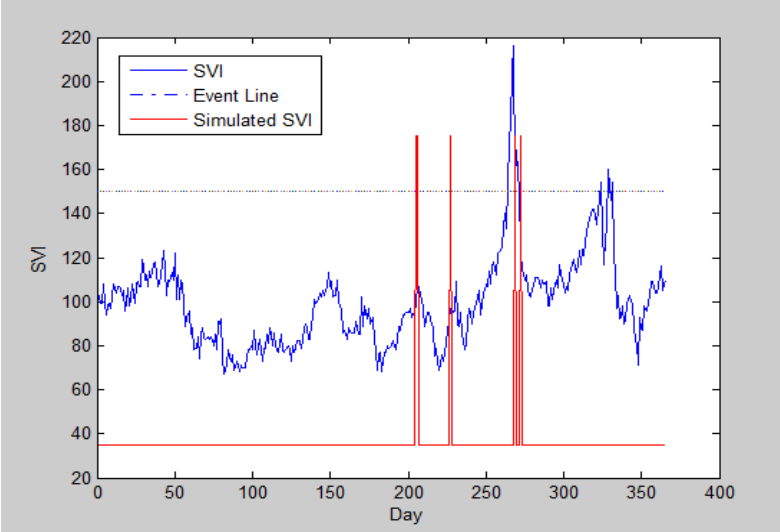

Figure 6.2: Test Result of SVI in 2007 for Battery B

III. Analysis of Data Combination C for Battery B

The process is the same as in Section 6.3.1.1. Figure 6.3 shows the test result for Battery B in 2008. From the figure, it can be seen that the predicted SVI values are all false positive results and no sludge bulking problem could be detected.

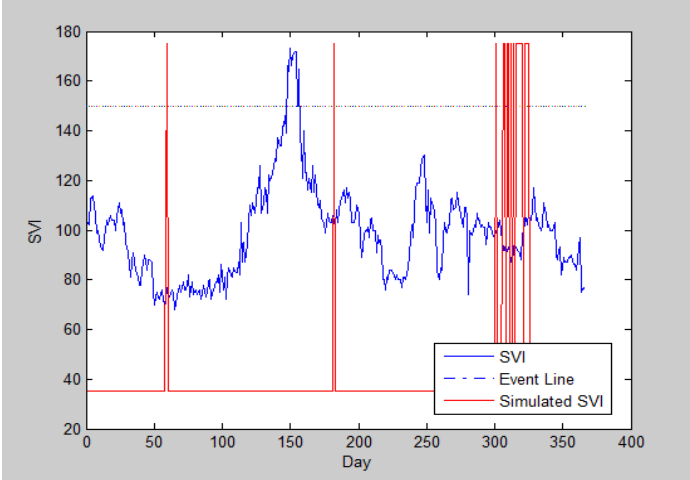

Figure 6.3: Test Result of SVI in 2008 for Battery B

IV. Analysis of Data Combination D for Battery B

The testing process is the same as in Section 6.3.1.1. Figure 6.4 shows the test result for Battery B in 2009. Also, from the figure, it can be seen that the predicted SVI values are all false positive results and no sludge bulking problem could be detected.

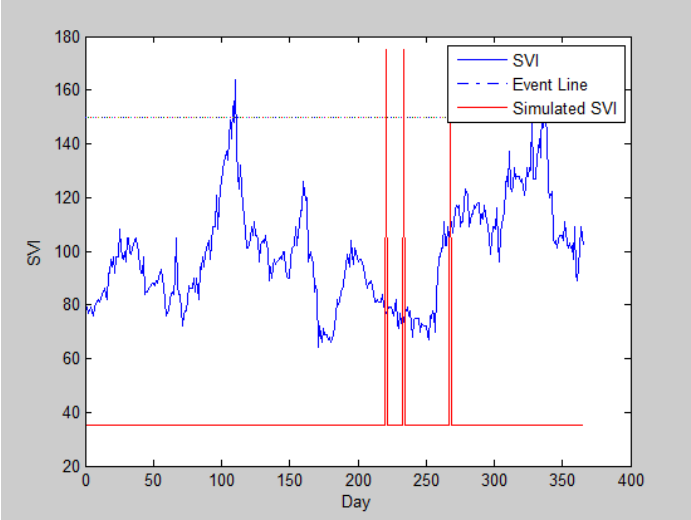

Figure 6.4: Test Result of SVI in 2009 for Battery B

6.3.2. Analysis of Battery A Using MATLAB

From the previous analysis of test results of Battery B, it can be seen the prediction results of the combined method are not useful. However, Battery A is still used to test the application of the combined method. Three different data combination tests are performed: (A) Training set: 2002 to 2005, Testing set: 2006; (B) Training set: 2002 to 2006, Testing set: 2007; and (C) Training set: 2002 to 2007, Testing set: 2008. No sludge bulking event occurred in 2009 (Table 4.13) and no testing process can be done for 2009.

Figures 6.5 to 6.7 show the test results for Battery A. From these 3 figures, only 2006 (Figure 6.5) has a better result. There is a predicted event before the sludge bulking event in Figure 6.5 that could provide warning information that could prevent the sludge bulking problem in 2006. Figures 6.6 and 6.7 only show false positive results.

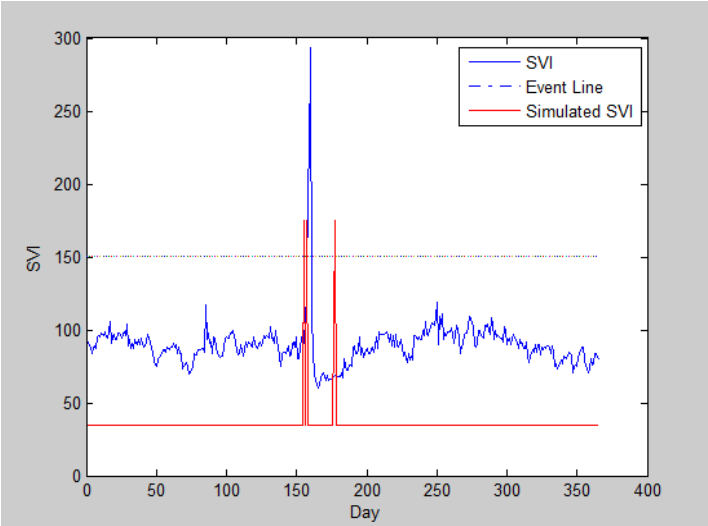

Figure 6.5: Test Result of SVI in 2006 for Battery A

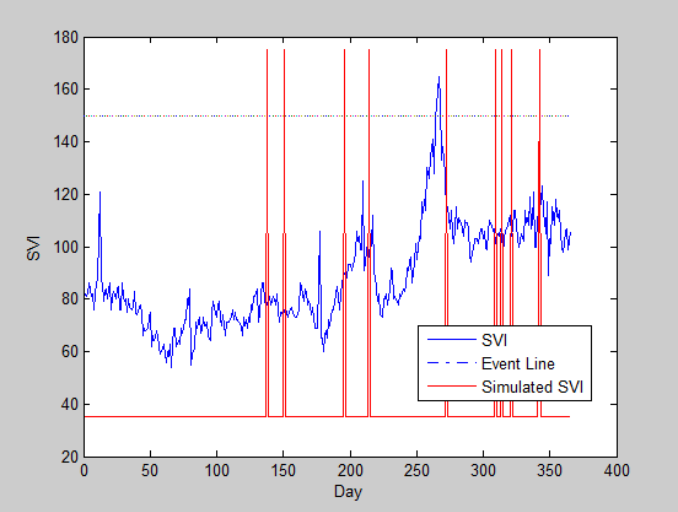

Figure 6.6: Test Result of SVI in 2007 for Battery A

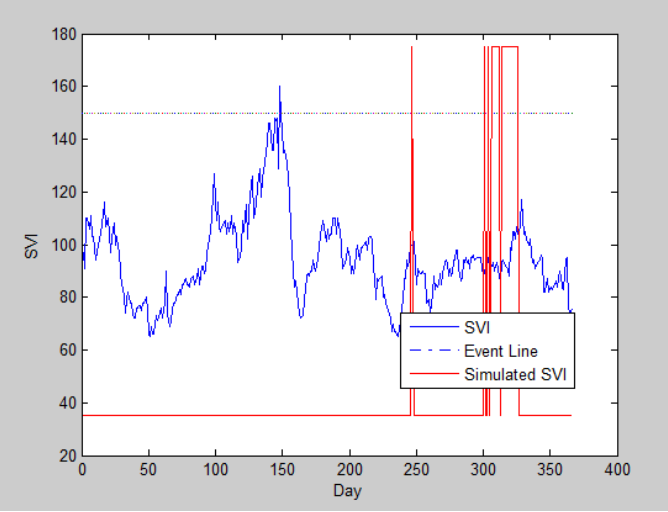

Figure 6.7: Test Result of SVI in 2008 for Battery A

6.3.3. Analysis of Battery C Using MATLAB

Two different data combination tests were performed: (A) Training set: 2002 to 2005, Testing set: 2006; and (B) Training set: 2002 to 2007, Testing set: 2008. No sludge bulking event occurred in 2007 and 2009 (Table 4.13). Figures 6.8 and 6.9 show the test results for Battery C. The result in 2006 (Figure 6.8) is chaos, and it has too many false positive results. From Figure 6.9, it can be seen that all the events predicted in 2008 are false positive results.

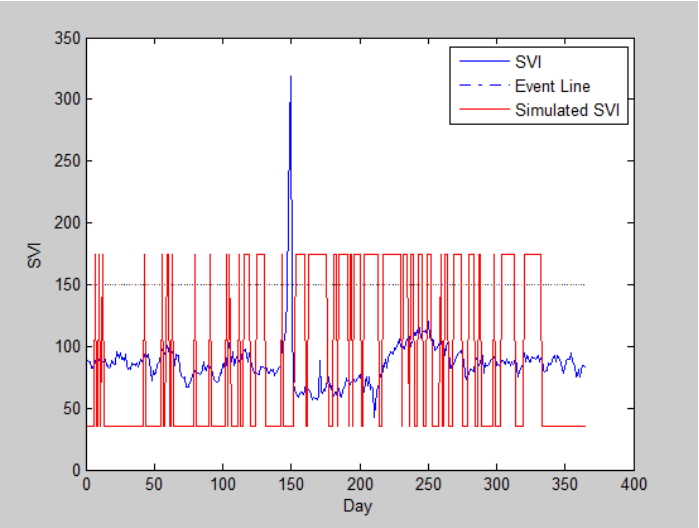

Figure 6.8: Test Result of SVI in 2006 for Battery C

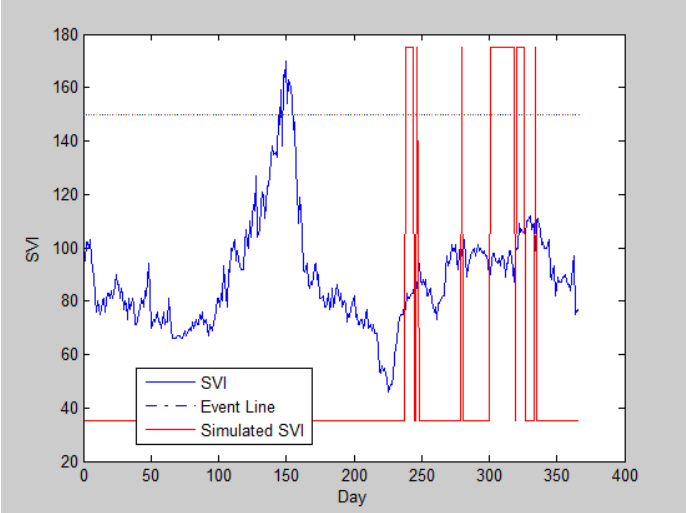

Figure 6.9: Test Result of SVI in 2008 for Battery C

6.3.4. Analysis of Battery B Using SPSS

The SPSS is statistical software package that can run the MLR model. The SPSS can also demonstrate what wastewater quality parameters have strong influence on the SVI state. According to the previous analysis, it can be seen that the combined method has poor performance in detecting and predicting sludge bulking problems. Unlike the operation process in MATLAB, the operation process in SPSS will use all the state SVI data and all wastewater quality parameters data in Battery B (2002-2009) to build the MLR model. With the built MLR model, then SPSS can reverse the operation to predict the state for each SVI data value. The goal of this whole process is to check the accuracy of the SVI state value predicted by the MLR model.

The procedure of MLR model can be found in Chapter 3. Table 6.5 lists the result for Battery B obtained by SPSS. The overall correct percentage is 70.6%, which seems to be an acceptable result. However, it should be noted that the correct percentage of state 1 is 99.4%, which leads to the overall correct percentage as 70.6%. Also, the correct percentage for the pattern states (2, 3, and 4) and event state (5) is very low. The correct percentage for the pattern states are all 0, and the correct percentage for the event state is only 1.8%. These results show that the MLR model fails to correctly detect sludge bulking problems.

Table 6.5: Test Results of State for the SVI Data for Battery B using SPSS

Observed state	Predicted state					
	1	2	3	4	5	Percent Correct
1	2050	0	0	0	13	99.4%
2	61	0	0	0	1	.0%
3	58	0	0	0	4	.0%
4	62	0	0	0	0	.0%
5	661	0	0	0	12	1.8%
Overall Percentage	99.0%	.0%	.0%	.0%	1.0%	70.6%

Table 6.6 lists the output of the estimation of selected wastewater quality parameters in the MLR model. The first category (state 1) was chosen as the reference set, so there is no output for state 1. From the β values in the Table 6.6, the equations for the MLR model can be obtained. $Z_2 = \sum X_i * \beta_2$ can be expressed as equation 6.1 where the subscript I refers to the different wastewater quality parameters.

$$Z_2 = X_i * \beta_2 = \text{-9.11 - 0.99*(Flow rate) +0.13*(Air flow rate)-28.648(F/M ratio)-2.07*(Detention time)+0.02*(SRT)+0.19*(Influent NH3)+0.22*(Effluent NH3)+0.06*(Influent DO)+0.02*(Influent BOD)+0.04*(Temp)+1.09*(Effluent pH)} \quad (6.1)$$

Similar as for state 2, the Z_j for the states 3, 4, and 5 can be obtained from the Table 6.6. With the Z_j for each state, the probability for each state can be obtained. For example, the probability for the state 4 is:

$$P(Y_i = 4) = \frac{\exp(Z_4)}{1 + \exp(Z_2) + \exp(Z_3) + \exp(Z_4) + \exp(Z_5)} \quad (6.2)$$

With all the calculated probabilities for each group of selected wastewater quality parameters, the highest probability for the state number was chosen as the detected state.

For instance, if the probability of state 4 is higher than that for the other states, so the MLR model will consider the state of the SVI value is state 4.

From Table 6.6, it can be seen that some quality parameters have significance relations (sig.<0.05) with the event state, including wastewater flow rate, detention time, F/M ratio, influent BOD, temperature, SRT, and effluent pH.

Table 6.6: MLRM Output of Estimation of Wastewater Quality Parameters in Battery B

Parameters	State							
	2		3		4		5	
	β_2	Sig.	β_3	Sig.	β_4	Sig.	β_5	Sig.
Intercept	-9.11	0.62	4.54	0.80	2.58	0.89	17.08	0.00
Flow rate	-0.09	0.26	-0.13	0.12	-0.04	0.64	-0.07	0.00
Air Flow rate	0.13	0.19	0.05	0.55	0.05	0.56	0.01	0.59
F/M ratio	-28.64	0.62	-31.50	0.66	-100.05	0.17	-40.15	0.00
Detention time	-2.07	0.06	-2.05	0.07	-1.66	0.14	-1.52	0.00
SRT	0.02	0.62	-0.01	0.85	-0.06	0.37	-0.09	0.00
Influent NH3	0.19	0.11	0.12	0.33	0.16	0.19	0.02	0.50
Effluent NH3	0.22	0.48	0.13	0.69	0.35	0.20	-0.08	0.29
Influent DO	0.06	0.79	0.01	0.98	-0.30	0.26	-0.08	0.12
Influent BOD	0.02	0.24	0.01	0.64	0.02	0.18	0.02	0.00
Temp	0.04	0.26	0.03	0.41	0.03	0.46	0.03	0.00
Effluent pH	1.09	0.54	0.41	0.81	-0.19	0.92	-1.21	0.00

6.3.5. Analysis of Battery A Using SPSS

All the SVI state data and all wastewater quality parameters data for Battery A (2002-2009) were used to build the MLRM by SPSS. The analysis process is the same as described in Section 6.3.4. Table 6.7 lists the test results for Battery A. From the table, it can be seen that the prediction accuracy of the event state is still very low at only 1.8%.

Table 6.7: Test Results of State for the SVI Data for Battery A by SPSS

Observed state	Predicted state					
	1	2	3	4	5	Percent Correct
1	2050	0	0	0	13	99.4%
2	61	0	0	0	1	.0%
3	58	0	0	0	4	.0%
4	62	0	0	0	0	.0%
5	661	0	0	0	12	1.8%
Overall Percentage	99.0%	.0%	.0%	.0%	1.0%	70.6%

Table 6.8 lists the output of the estimation of the selected wastewater quality parameters in the MLR model. From Table 6.8, it can be seen that some wastewater quality parameters have significant realtions (sig.<0.05) with the event state, including influent BOD, influent and effluent ammonia, SRT, detention time, F/M ratio and effluent pH.

Table 6.8: Output of Estimation of Wastewater Quality Parameters in Battery A

Parameters	State							
	2		3		4		5	
	β_2	Sig.	β_3	Sig.	β_4	Sig.	β_5	Sig.
Intercept	-10.324	0.364	-8.84	0.444	-15.725	0.185	7.447	0.051
Effluent pH	1.363	0.266	1.264	0.307	1.01	0.44	-1.268	0.002
Influent BOD	0.014	0.155	0.025	0.008	0.026	0.005	0.014	0
Effluent ammonia	-0.144	0.513	-0.175	0.433	-0.304	0.18	-0.263	0.001
SRT	-0.006	0.868	-0.001	0.967	0.005	0.878	-0.07	0
F/M	-31.283	0.441	-62.224	0.122	-78.629	0.045	-28.143	0.039
Flow	0.007	0.872	0.007	0.876	0.087	0.033	-0.001	0.948
Air flow	-0.052	0.221	-0.042	0.35	-0.006	0.895	0.028	0.124
Detention time	-0.349	0.562	-0.639	0.322	0.035	0.953	-0.443	0.041
Influent NH3	0.118	0.172	0.102	0.241	0.152	0.076	0.087	0.006
Influent DO	-0.455	0.011	-0.51	0.006	-0.675	0	-0.028	0.55
temp	-0.011	0.643	-0.009	0.694	-0.024	0.288	0.01	0.208

6.3.6. Analysis of Battery C Using SPSS

The SVI state data and wastewater quality parameters data for Battery C (2002-2008) are used to run the MLR model by SPSS. The year of 2009 is not included because no sludge bulking event happened in year 2009 and there is no need to contain a test year without events. The analysis process is the same as described in Section 6.3.4. Table 6.9 lists the test results for Battery C. From the table, it can be seen that the prediction accuracy of the event state is still low at only 16.2%.

Table 6.9: Test Results of State for the SVI Data for Battery C by SPSS

	Predicted					
Observed	1	2	3	4	5	Percent Correct
1	1737	0	0	0	42	97.6%
2	18	0	0	0	1	.0%
3	18	0	0	0	1	.0%
4	18	0	0	0	1	.0%
5	604	0	0	0	117	16.2%
Overall Percentage	93.7%	.0%	.0%	.0%	6.3%	72.5%

Table 6.10 lists the output of the estimation of selected wastewater quality parameters in the MLR model. From Table 6.10, it can be seen that some wastewater quality parameters have significant relations (sig.<0.05) with the event state, including pumped air flow rate, SRT, influent and effluent ammonia, influent DO, influent BOD, and effluent pH.

Table 6. 10: Output of Estimation of Wastewater Quality Parameters in Battery C

Parameters	State							
	2		3		4		5	
	β_2	Sig.	β_3	Sig.	β_4	Sig.	β_5	Sig.
Intercept	10.379	0.607	1.993	0.928	13.915	0.499	18.563	0
Flow rate	-0.019	0.787	0.025	0.737	-0.052	0.493	-0.011	0.45
Air Flow	0.074	0.216	0.058	0.384	0.014	0.833	0.032	0.018
F/M	1.877	0.983	24.229	0.702	23.116	0.768	18.986	0.14
Detention time	0.311	0.548	0.199	0.818	0.04	0.95	-0.114	0.504
SRT	-0.013	0.833	-0.015	0.801	0.002	0.921	-0.027	0.041
Influent NH3	-0.129	0.432	0.069	0.655	-0.039	0.805	0.093	0.004
Effluent NH3	0.118	0.781	0.1	0.794	0.153	0.703	-0.284	0.005
Influent DO	-0.712	0.033	-0.758	0.029	-0.406	0.2	-0.265	0
Influent BOD	-0.021	0.403	-0.01	0.61	-0.011	0.596	-0.006	0.088
Temp	-0.042	0.307	-0.028	0.498	-0.017	0.679	0.01	0.22
Effluent pH	-1.578	0.499	-1.288	0.609	-1.928	0.423	-2.861	0

6.4. Discussion and Conclusion

As previously mentioned, the sludge bulking problem is an extremely complex process. The previous research in the literature described in Chapter 2 focused on one or two wastewater quality parameters. So the idea for this chapter is that the sludge bulking could be related with more than 2 wastewater quality parameters. Most of the former prediction research used the point by point prediction of SVI values. The improved TSDM method and the HMMs method studied in this thesis all focus on predicting sludge bulking events defined by high SVI data values considering only the SVI time series data. The combined method studied in this chapter could use other selected wastewater quality parameters to detect the pattern states and event state that are obtained from the HMMs method.

In previous research on sludge bulking considering the SVI data and other wastewater quality parameters, some parameters have been found to have more impact on

sludge bulking than other parameters. These wastewater quality parameters are: influent DO, effluent DO, influent and effluent ammonia, temperature. These parameters are selected to be used to test the combined method.

The test results of the combined method are not useful for sludge bulking detection. From the application of the combined method using MATLAB and SPSS, the pattern states basically cannot be detected. And the accuracy of detection for the event state is lower than 20%, which is not useful to predict and prevent the sludge bulking problems. However, from the output of the SPSS, some wastewater quality parameters can be considered as parameters significantly related to sludge bulking. For Battery B, these parameters are wastewater flow rate, detention time, F/M ratio, influent BOD, temperature, SRT, and effluent pH. For Battery A, these parameters are SRT, influent and effluent ammonia, influent DO, and effluent pH. For Battery C, these parameters are pumped air flow rate, SRT, influent and effluent ammonia, influent DO, influent BOD, and effluent pH. From the output of all three batteries, some common parameters can be found, i.e. SRT and effluent pH. In conclusion, these parameters should be studied further in the analysis of sludge bulking.

CHAPTER 7 CONCLUSIONS AND RECOMMENDATIONS

This thesis focused on the detection and analysis of sludge bulking problems by application of three machine learning methods: the improved Time series Data Mining (TSDM) method, the Hidden Markov Model (HMMs) method, and a method combining the Hidden Markov and Multinomial Logistic Regression Models. The results and analysis are presented in the previous chapters. The improved TSDM method and the HMMs method show their capability to detect and predict sludge bulking events. These two methods have the notable advantage of focusing on detecting the temporal patterns for the events instead of point to point prediction. Such an advantage has more efficiency and could provide warning information to the WWTP operator. However, they are still new methods, which mean they are not perfect and they need to be improved in the future. The combined method demonstrates some useful information on the relationship between the SVI data and other wastewater quality parameters, though the combined method cannot effectively detect sludge bulking events at this time.

The improved TSDM method can have a sludge bulking event prediction accuracy between 60% and 100%, where a sludge bulking event is defined by a SVI value higher than 150 mL/g. Nearly all the long term sludge bulking period events can be detected except for that in 2006 for Batteries A and C. The sludge bulking event that occurred in 2006 was distinguished by a sudden jump in the SVI values, no similar jump happened before in the training data set to allow the TSDM method to learn how to detect such events. The analysis of the improved TSDM method reveals that it is a new method, and it has some special requirements for application to sludge bulking event prediction. For instance, some parameters of the improved TSDM method should be chosen

carefully before the testing process, e.g., the phase space embedding dimension, the enlarge radius for temporal pattern clusters, and the event value. For example, the event value could be reduced if the number of sludge bulking events in the training set data is not sufficient for the improved TSDM method to learn to find the temporal pattern clusters, e.g., as was done for the SVI tests of Battery C.

For the HMMs method, it should be noted that the advantage of this method is to detect the pattern states and event state separately. The improved TSDM method needs to detect the patterns by consistent SVI points. But the HMMs method detects the pattern state for each SVI point. This means the HMMs method could provide warning information to the WWTP operators, even if the HMMs method only detects the first state of the pattern. From the results and analysis presented in Chapter 5, once the first pattern state was detected, there was high probability (>80% in all cases, mostly > 90%) the event state (sludge bulking) would be occurred. It was also demonstrated that the HMMs method has capability and effectiveness to detect sludge bulking and provide warning information for impending sludge bulking events to the WWTP operators. Similar to the improved TSDM method, the HMMs method also has some parameters that need to be set before the testing process, e.g., initial value for each state. It also has some short comings including a stability problem. For example, the event state probability in the transition probability matrix cannot correctly converge, and such problem makes the HMMs method fails to predict the event state for the SVI data. Such a problems need to be investigated and improved in future research.

For the combined method, the new idea is to combine the HMM method and a Multinomial Logistic Regression Model. Although the testing results showed the

combined method was not effective in predicting sludge bulking events, it did provide useful information on the relation between the SVI data and some other wastewater quality parameters that have significant impact on the sludge bulking, i.e., sludge retention time (SRT), and effluent pH for all three batteries.

The improved TSDM method and HMMs method both demonstrate an ability to be applied to real world sludge bulking data. These methods could be useful for the WWTP operators possibly using both methods at the same time. Applying both methods could provide a double check on the possibility for impending sludge bulking and increase the detection and prediction accuracy. Applying both methods can also reduce the short comings for these two methods, i.e. the stability problem in the HMMs method and the need for a complete pattern for the TSDM method. Also, it is recommended that both methods should be run several times to obtain a comprehensive result, and this procedure could reduce the risk of failing to detect sludge bulking events. The training set used for both methods needs to include a sufficient number of events to properly train the methods, so it is recommended that at least 45 events should be included in the training data set.

In this thesis, the improved TSDM method and the HMMs method were applied to detect the temporal patterns in the SVI data alone. Since some wastewater quality parameters have been found to have a significant impact on the sludge bulking problem. It is meaningful to detect the temporal patterns relations to sludge bulking in these wastewater quality parameters in future research.

BIBLIOGRAPHY

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics,* 37(6), 1554-1563.

Bayo, J., Angosto, J. M., and Serrano-Aniorte, J. (2006). Evaluation of physicochemical parameters influencing bulking episodes in a municipal wastewater treatment plant. *Water Pollution VIII: Modelling, Monitoring and Management,* 95, 531-542.

Belanche, L., Valdes, J., Comas, J., Roda, I., and Poch, M. (2000). Prediction of the bulking phenomenon in wastewater treatment plants. *Artificial Intelligence in Engineering,* 14(4), 307-317.

Bhatla, M. N. (1967). Relationship of activated sludge bulking to oxygen tension. *Journal of the Water Pollution Control Federation,* 39(12), 1978-1985.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer Science+Business Media, LLC.

Bitton, G. (2005). *Wastewater Microbiology* (3rd ed.). Hoboken, NJ, USA: John Wiley and Sons, Inc.

Schelter, B., Winterhalder M., and Timmer, J. (2007). *Handbook of Time Series Analysis.* New York: Wiley.

Capodaglio, A., Jones, H., Novotny, V., and Feng, X. (1991). Sludge bulking analysis and forecasting: Application of system identification and artificial neural computing technologies. *Water Research,* 25(10), 1217-1224.

Chan, W. T., and Koe, L. C. (1991). A knowledge-based framework for the diagnosis of sludge bulking in the activated sludge process. *Water Science and Technology,* 23, 847-855.

Chen, J., and Beck, M. B. (1993). Modelling, control and on-line estimation of activated sludge bulking. *Water Science and Technology,* 28(11-12), 249-256.

Chiu, S. L. (1994). Fuzzy model identification based on cluster estimation. *Journal of Inteligent and Fuzzy Systems,* 2(3), 267-278.

Chudoba, J. (1985). Control of activated sludge filamentous bulking—VI. Formulation of basic principles. *Water Research,* 19(8), 1017-1022.

Combs-Orme, J. G. (2009). *Multiple Regression with Discrete Dependent Variables.* New York: Oxford University Press.

Feng, X. and Huang, H. (2005). A fuzzy-set-based reconstructed phase space method for identification of temporal patterns in complex time series. *Knowledge and Data Engineering,* 17(5), 601-613.

Forster, C. (1971). Activated sludge surfaces in relation to the sludge volume index. *Water Research,* 5, 861-870.

Grangier, D. (2008). *Machine Learning for Information Retrieval.* Martigny: Universit´e de Nice Sophia Antipolis, France.

Gujer, W., Henze, M., Mino, T., and Van Loosdrecht, M. (1999). Activated Sludge Model No.3. *Water Science and Technology,* 39(1), 183-193.

Henze, M., Gujer, W., Mino, T., Matsuo, T., Wentzel, M. C., and Marais, G. v. R. (1995). *Activated Sludge Model No.2.* Scientific and Technical Report No. 3. London.

Henze, M., Grady C. P., L., Gujer, W., Marias G. v. R., and Matsuo, T. (1987). *Activated Sludge Model No.1.* Scientific and Technical Report No.1., International Association on Water Pollution Research and Control, London.

Heukelekian, H. (1941). Activated sludge bulking. *Sewage Works Journal,* 13(1), 39-42.

Hiraoka, M., Tsumura, K., and Yamamoto, Y. (1988). Computer-based filamentous microorganism identification support system. *International Workshop on Artificcial Intelligence for Industrial Applications 1988*. Japan: IEEE, pp. 283-288.

Huang, H. (2001). *An Improved Time Series Temporal Pattern Identification Method.* Milwaukee, WI, USA: Master's Thesis, Marquette University.

Huang, X., Jack, M., and Ariki, Y. (1990). *Hidden Markov Models for Speech Recognition.* Edinburgh: Edinburgh University Press.

Jenkins, D., Richard, M. G., and Daigger, G. T. (2004). *Manual on the Causes and Control of Activated Sludge Bulking, Foaming and Other Solids Separation Problems.* Boca Raton, FL, USA: Lewis Publishers.

Kantz, H., and Schreiber, T. (2004). *Nonlinear Time Series Analysis* (Second ed.). Cambridge: Cambidge University Press.

Kao, J. J., Brill, E. D. Jr., Pfeffer, J. T., and Geselbracht, J. J. (1993). Computer-based environment for wastewater treatment plant design. *Journal of Environmental Engineering,* 119(5), 931-945.

Kappeler, J., and Gujer, W. (1994). Influences of wastewater composition and operating conditions on activated sludge bulking and scum formation. *Water Science and Technology,* 30(11), 181-189.

Madoni, P., Davoli, D., and Gibin, G. (2000). Survey of filamentous microorganisms from bulking and foaming activated sludge plants in Italy. *Water Research,* 34(6), 1767-1772.

Martínez, M., Rodríguez-Roda, I., Poch, M., Cortés, U., and Comas, J. (2006). Dynamic reasoning to solve complex problems in activated sludge processes: A step further in decision support systems. *Water Science and Technology,* 53, 191-198.

Martins, A. M., Pagilla, K., Heijnen, J. J., and Loosdrecht, M. C. (2004). Filamentous bulking sludge—A critical review. *Water Research,* 38(4), 793-817.

Menard, S. (2001). *Applied Logistic Regression Analysis* (2nd ed.). Thousand Oaks, CA, USA: Sage Publications, Inc.

Metcalf, and Eddy. (2003). *Wastewater Engineering Treatment and Reuse.* New York: USA: McGraw-Hill.

Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2008). *Introduction to Time Series Analysis and Forecasting.* Hoboken, NJ, USA: John Wiley and Sons, Inc.

Ng, W. J., Ong, S. L., and Hossain, F. (2000). An algorithmic approach for system-specific modelling of activated sludge bulking in an SBR. *Environmental Modeling and Software,* 15(2), 199-210.

Pipes, W. O. (1979). Activated sludge bulking and settling tank performance. *Journal of Water Pollution Control Federation.* 51(10), 2534-2537.

Povinelli, R. J. (1999). *Time Series Data Mining: Identifying Temporal Patterns for Characterization and Prediction of Time Series Events.* Milwaukee, WI, USA: Ph.D. Dissertation, Marquette University.

Povinelli, R. J. and Feng, X. (2003). A new temporal pattern identification method for characterization and prediction of complex time series events. *IEEE Transactions on Knowledge and Data Engineering,* 15(2), 339-352.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceeding of IEEE,* 77(2), 257-286.

Rensink, J. H. (1974, Aug.). New approach to preventing bulking sludge. *Water Pollution Control Federation,* 46(8), 1888-1894.

Sezgin, M., Jenkins, D., and Parker, D. S. (1978). A unified theory of filamentous activated sludge bulking. *Water Pollution Control Federation,* 50(2), 362-381.

Sin, G., Govoreanu, R., Boon, N., Schelstraete, G., and Vanrolleghem, P. A. (2006).

Evaluation of the impacts of model-based operation of SBRs on activated sludge microbial community. *Water Science and Technology,* 54(1), 157-166.

Snyman, J. A. (2005). *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-based Algorithms.* New York: Springer, Inc.

Soyupak, S. (1989). Effects of operational parameters on the settling properties of activated sludge. *Environmental Technology Letters,* 10, 471-478.

Takens, F. (1981). Detecting strange attractors in turbulence. In D. Rand, and L.-S. Young, eds., *Lecture Notes in Mathematics* (pp. 366-381). New York: Springer.

"wastewater treatment." *Encyclopædia Britannica. Encyclopædia Britannica Online Academic Edition*. Encyclopædia Britannica Inc., 2012. Web. 17 Apr. 2012. <http://www.britannica.com/EBchecked/topic/666611/wastewater-treatment>.

Yasuda, M. (1976). The influence of pH and organic loading on the filamentous bulking of activated sludge. *Transactions of JSCE,* 8(247), 51-59.