2013

Learning in the presence of sudden concept drift and measurement drift

William Rowcliffe
Iowa State University

Learning in the presence of sudden concept drift and measurement drift

by

William A. Rowcliffe

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:

Sigurdur Olafsson, Major Professor

Lizhi Wang

Dianne Cook

Iowa State University

Ames, Iowa

2013

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOTATION

$M_{t,j}$ measurement of attribute j at time t

$\psi_{t,j}$ drift from mean of attribute j at time t

$\hat{\psi}_{t,j}$ estimated drift from mean of attribute j at time t

$\boldsymbol{\beta}$ vector of linear model parameters

β_j linear model parameter associated with independent variable j

λ exponential weighted moving average weight

y iobserved outcome variable

$\hat{y}_t$ prediction of numerical value y_t

$x_{j,t}$ datastream value of attribute j at time t

$x_{j,t}$ datastream value estimate of attribute j at time t

Xj attrubute j

ws fixed window size used in implementation of the methods of this thesis

ACKNOWLEDGEMENTS

This would not have been possible without the guidance and support from a number of parties. Without these people and institutions, I would not have the opportunity to spend my hours realizing, developing, and expressing my ideas. This time was invaluable to me.

Many members of the faculty of the Iowa State University department of Industrial and Manufacturing Systems engineering have played a role in the development of this work. Out of these, the most influential has been Dr. Sigurdur Olafsson. His patience, guidance and ability to interpret my thoughts has been crucial to my ability to continue this work.

I can not forget to acknowledge Dr. Sarah Ryan and Leslie Potter for my introduction to graduate studies and appointing me the teaching assistantship which provided my financial support during this time. I thoroughly enjoyed the position.

I would like to thank my committee members Dr. Lizhi Wang and Dr. Dianne Cook. I have great admiration both. Their work ethic, excitement, and knowledge encouraged me to expand my understanding of mathematics and data analysis communication.

The humor, presence, and conversation of all my friends in the program were always a treat. I especially value the time shared with Pan Xu, Walter Bennett, and Youngrok Lee. The success of the friends I've known for many years provided determination to finish this work. You know who you are.

ABSTRACT

The current availability of vast data storage and the computational power to enact algorithms for interpreting that data in real time leads to the possibility of real time adaptive systems. Because change is nearly always inevitable, companies must strive to increase the adaptability of their manufacturing or service systems. To accomplish this, the methods for correcting the system and determining the correct change point must be studied.

The motivation of this thesis is advancing the ability of proper prediction and classification model learning on data streams containing change. This problem is known as concept drift. Motivation also stems from a study on a system with these properties, at an active manufacturing facility. After reviewing articles relating to the specific problem in the study, a similarity between the study and the studies performed in the research area of advanced process control became clear.

The underlying cause for the change in the manufacturing system is identified as measurement drift. The identification of measurement drift is explained. A discussion of the mathematical model representing measurement drift is provided.

Existing concept drift algorithms are adapted to fit the needs of the measurement drift problem. Their performance on the data from the study and synthetic data sets mimicking varying levels of drift magnitude and frequency is assessed. The results are compared to a popular advanced process control method, exponential weighted moving average adapting intercept (EWMA-I).

The advanced process control literature inspired the development of two new methods for learning in the presence of concept drift. The methods, ADMEAN and CD-EWMA (ADaptive MEAN and Concept Drift Exponential Weighted Moving Average), make changes to the incoming stream of independent variables. The performance of these algorithms on the measurement drift datasets and synthetic concept drift datasets is provided.

CHAPTER 1. INTRODUCTION

"Nothing endures, except change." - Heraclitus

1.1 Motivation

One of the keys to being a successful company is striving to have very high quality standards. Meanwhile, the efficiency of its systems is extremely important for the continued operation and growth of the company. One of the biggest obstacles hindering efficiency and product or service quality is good decision making. Because of the resources (time, employees, equipment) needed to obtain good decisions, decision making can be very costly and inefficient. When poor decisions are made, companies miss opportunities and sometimes need to waste even more resources to correct problems caused by the poor decisions. When good decisions are made, companies benefit from increased efficiency, quality, and sales.

In manufacturing, bad decisions can lead to defective products. If the defective product is detected before reaching the customer, there are three possible actions. 1) The product can be reworked. This rework cuts into time that the manufacturing system could be producing product and may require excess parts. 2) The defective product can be scrapped. This leads to one less unit produced in the time period allotted. The company will also be forced to absorb the cost of the resources attributed to the scrapped product. 3) The product can be passed, to meet demand. If the defective product is not detected, it has the same effect as a passed defective product.

It is important to address the problems as soon as possible. Since the cause of the defect may be a bad process, early detection allows prevention of future defective product. As a product travels through a manufacturing process, additional features may make rework more

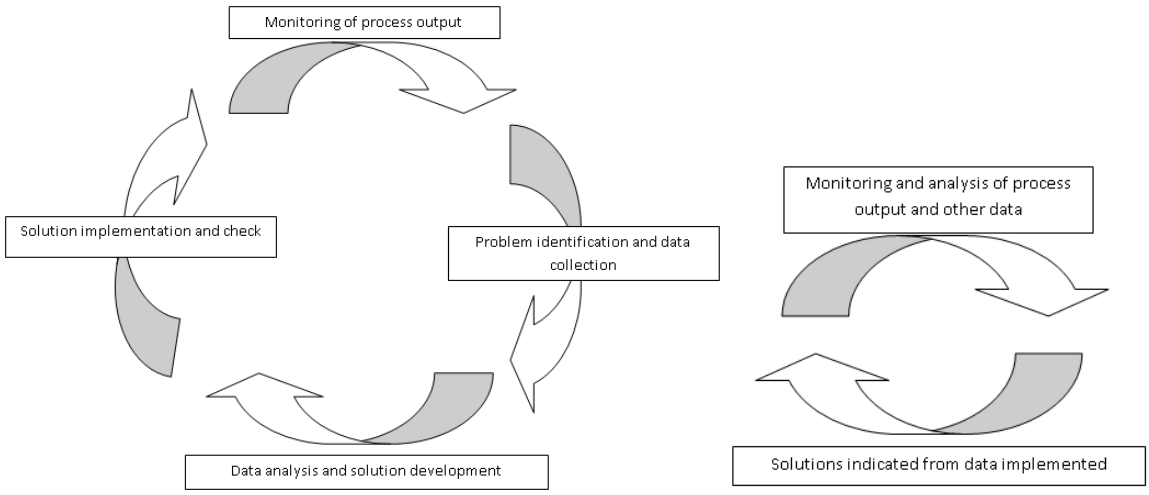
difficult. If the product is scrapped, all additional resources spent on the product since the defect occurred are unnecessarily lost. Because of these reasons for early detection and action, on-line monitoring and process adaptation can be extremely valuable to a manufacturer.

According to continuous improvement philosophies (popular bases for operating philosophies among companies today), customers receiving bad product is the biggest cost. In this case, the customer becomes a risk for the company, as they will likely share their experience with people they interact with. As public reviews have become more accessible, through the Internet, the audience for these dissatisfied customers can be quite large. The cost of appeasing this customer, reversing a public image, and experiencing lost sales can be much greater than detecting and addressing the defective products and processes before the customer receives the defective product. This increases the gains of on-line monitoring and process adaptation.

This millennium has seen the birth of big data. Companies collect more data than they can use. This is happening because collecting and storing large amounts of data has become extremely cheap. Companies struggle with using the data they collect. Making sense of this data has become a large focus in the research community. So, although this is a relatively new problem, there is a large body of knowledge to draw from. This thesis has been motivated by hopes to discuss and show methods that can exploit the explosion of data and knowledge to address the decision making problems within manufacturing systems.

According to experts in process improvement, variance in process output is due to common causes of variance and special causes of variance. Common causes in variance are things that can not be controlled, such as machining variance, humidity and temperature of the manufacturing environment, or material variance. They are inherent in the manufacturing system. Special causes of variance are not inherent in the system, such as an improperly trained operator or tool wear.

Traditionally, time must be spent defining a problem, so that the correct data can be collected. After collecting and analyzing the data, static solutions are formed to address the problems, despite the potential of future changes to the problems. While methods have been developed for manually monitoring the static solutions, the process restarts every time a change in the system is detected. Therefore, traditional problem definition and solution development



(a) Traditional approach (b) Online approach

Figure 1.1: Defect detection and process adaptation models

are difficult and resource intensive activities.

This thesis develops an underlying step toward updating this traditional approach into a new approach where data is collected and analyzed simultaneous to process monitoring. The on-line analysis of this data may point to the solution or even lead to better detection of special causes of variance.

Until the late 1990's, the cost of obtaining data was high. The electronic collection and storage of data was not feasible. Therefore, collecting insignificant data was a significant waste. Now, companies can collect almost unbounded amounts of data with little effort. The increase in data and the advances in analytics have given companies the power to make more accurate predictions and classifications, leading to better solutions.

Electronic monitoring and model storage allow for implementation of dynamic solutions, in which the system can adapt to change. Advanced process control, developed for improving chemical processes and semiconductor manufacturing, is an example of this type of monitoring and data analysis.

Throughout this thesis, changes in the data produced by a system will be referred to as concept drift. Concept drift is a term used to describe a change in the hidden context of a data stream. By definition, the special causes of variance described in process control literature

represent sources of concept drift.

In one part of this thesis, the focus is a system that relies on a continuous valued numerical decision. This decision is the product of input variables, which are subject to special causes of variance. Specifically, this problem is on-line detection and correction for measurement drift, a special cause of variance at the assembly line. The performance of existing advanced process control methods and adapted concept drift methods to this problem are assessed. This evaluation will be performed by observing the error of the algorithms applied to real data, described below, and synthetic data.

A second focus of this thesis is classification in the presence of concept drift. Companies are now able to classify their customers. This gives them the ability to recommend products to certain customers (e.g. Netflix and Amazon). They strive to identify if a customer will, or will not, like a product. Money lenders can use classification models to identify bad debt or deny lending. It is inevitable that the customers' tastes and needs change. This phenomenon is a type of concept drift. While the experiments in this thesis focus on the classic concept drift problems, the results may be useful to the specific classification problems that companies face. The algorithms are tested on synthetic data generated to contain concept drift.

1.2 Industrial study

The data used for this study was collected at an active manufacturing facility. The computer system used in the study received input from electronic measurement devices. These measurements were used by the system to provide output, in the form of a continuous numeric value. This value was directly related to a decision at the assembly line, thus, becoming a decision variable. The operator used the output to select a part, to be used in production. The goal was to create a non-defective assembly. A testing station immediately followed the work station and returned defective products for rework. The assembly was reworked as many times as needed to produce a product that would pass the testing station. The input to the model and the decision variable leading to the assembly passing the testing station were recorded. The operator made a new decision, based on the output of the testing station.

After analyzing the data collected from this study, a clear pattern emerged. It was this

pattern that pointed to the methods discussed in this thesis, for addressing drift detection and the adjustment of the prediction model, after drift. Further, data following variations of this data pattern are simulated and used in the testing of the methods.

1.3 Contribution

The following is a list of the contributions this thesis provides:

- Explanation of measurement drift and identification of measurement drift
- Application and evaluation of advanced process control to an assembly process experiencing measurement drift
- Adaptation and evaluation of concept drift algorithms to an assembly process experiencing measurement drift
- Introduction and evaluation of ADMEAN (ADaptive MEAN), an algorithm for addressing concept drift
- Introduction and evaluation of CD-EWMA, an algorithm for addressing concept drift

Ultimately, this thesis tries to point out the connection between the ideas of process improvement and data mining's concept drift.

1.4 Organization

In this thesis, Chapter 2 will introduce and review the important literature. The connection between the motivation for this research and previous research will be discussed. Relevant results of previous studies are shared.

Chapter 3 explores the proposed methods for addressing the problem and generating synthetic data. Details of the experiments and algorithms applied can be found in this section. Chapter 4 discusses the results of the experiments on the industrial study data, the synthetic measurement drift data, and the synthetic concept drift data.

Finally, Chapter 5 provides a conclusion of the thesis. This includes a generalization of the results of the experiments, opportunities for future work, and concluding remarks regarding the course of this research.

CHAPTER 2. REVIEW OF LITERATURE

"God does not play dice." -Albert Einstein

2.1 Process control

2.1.1 Control charts

Detecting change in a system has been studied for a long period of time. As a way of improving the quality in mass production, Walter Shewhart developed statistical quality control (Shewhart (1931) and Shewhart (1939)). A major part of statistical quality control involved control charts. Control charts are still a big part of quality control and give quality engineers the ability to identify sources of variance in a process.

Control charts track sample measurements, or sample proportions, and compare them to a population mean and standard deviation. The interpretation of control charts has been developed to include 7 rules that indicate causes of variance. Below a list defines the seven rules of control charting. They are expressed as functions of the standard deviation and mean, which are used to define zones of the control chart. The location of the zones refereed to in the seven rules can be changed depending on factors related to the data collected and/or strategies of the analyzing party. If a source of variance causes one of these rules to be broken, it is called a special source of variance.

- No single observation should exceed three standard deviations of the mean
- No two of three consecutive observations should exceed two standard deviations from the mean, on the same side of the mean
- No four of five consecutive observations should exceed one standard deviations from the mean, on the same side of the mean

- No eight consecutive observations should exist on one side of the mean
- No eight consecutive observations should exist a similar direction from immediate predecessors
- The number of observations above and below the mean should be reasonably equal
- No thirteen consecutive observations should remain within one standard deviation of the mean

To allow easy implementation and retain the ability to track small changes in the process, cumulative summation (CUSUM) charts were developed and studied. CUSUM charts track the sum of deviance from a mean (Page (1954)). Along with CUSUM charts, exponentially weighted moving average (EWMA) charts have been studied (Crowder (1987), Lucas and Saccucci (1990), Gan (1993), and Chandrasekaran et al. (1995)). EWMA charts give a weight to each observed instance. The weight is exponentially lower with respect to age of the instance. In Roberts (1959), the appeal of a weight equal to $\frac{2}{5}$ is presented.

2.1.2 Advanced process control

Advanced process control (APC) is a set of methods for on-line process monitoring and adapting. Some application areas discussed in literature are semi-conductor manufacturing, chemical mixing, and mechanical polishing. These processes rely heavily on precise actions. Within these processes, unpredictable changes can alter the needed action. However, if the processes can adapt to these changes, they can continue to operate with precision.

According to Lee (2008), good fabrication metrology is essential for success in the semiconductor industry. The current equipment utilizes APC methods to minimize the effects of sources of measurement error. Further, Lee's dissertation demonstrates the advantages of APC methods in semiconductor manufacturing. He finds that exponential weighted moving average (EWMA) controls provide robust and accurate incremental feedback.

EWMA approaches are popular in the semi-conductor literature, represented by many variations. Two of these variations are the exponentially weighted moving average with gain adaptation (EWMA-G) and exponentially weighted moving average with intercept adaptation

(EWMA-I). In EWMA-I, the adjustments to the intercept of a linear model is changed according to exponentially weighted previous observations. Similarly, EWMA-G adjusts the slope of the linear model according to exponentially weighted previous observations.

As Del Castillo (2001) points out, if the parameters of these approaches are not successfully chosen, the methods will become inaccurate. In the case of poorly chosen parameters, the EWMA-I will begin over correcting and give increasingly worse feedback. Wang et al. (2010) shows that the parameters must follow the conditions in 2.1, where β is the actual slope, b represents the estimate of β, and λ is the weight used for exponential weighting.

$$0 < \frac{\beta}{b}\lambda < 2 \tag{2.1}$$

EWMA-G is also susceptible to overcorrection. Wang and He (2008) shows that the feedback from EWMA-G approaches a consistent error as overcorrection takes hold. The parameter conditions are given in 2.2, where T is the target value, α is the actual intercept, and a is the estimate of the intercept.

$$0 < \frac{T-\alpha}{T-a}\lambda < 2 \tag{2.2}$$

The relationship between process control and feedback controllers has been discussed quite thoroughly. Faltin and Tuckers (1991), Tucker et al. (1992), and Vander Weil et al. (1992) describe "Algorithmic Statistical Control" in a very similar manor as the description provided in the introduction of this thesis of the on-line approach to process control. Some of the studies find EWMA controls the optimal choice for their applications, see Baxley (1990), Baxley (1994), Hunter (1986), and MacGregor (1987). These studies observe a process with constant linear change. In Sachs et al. (1995), control charts and feedback control are combined for on-line process control. Control charts serve as a diagnostic to determine a gradual or instant model update.

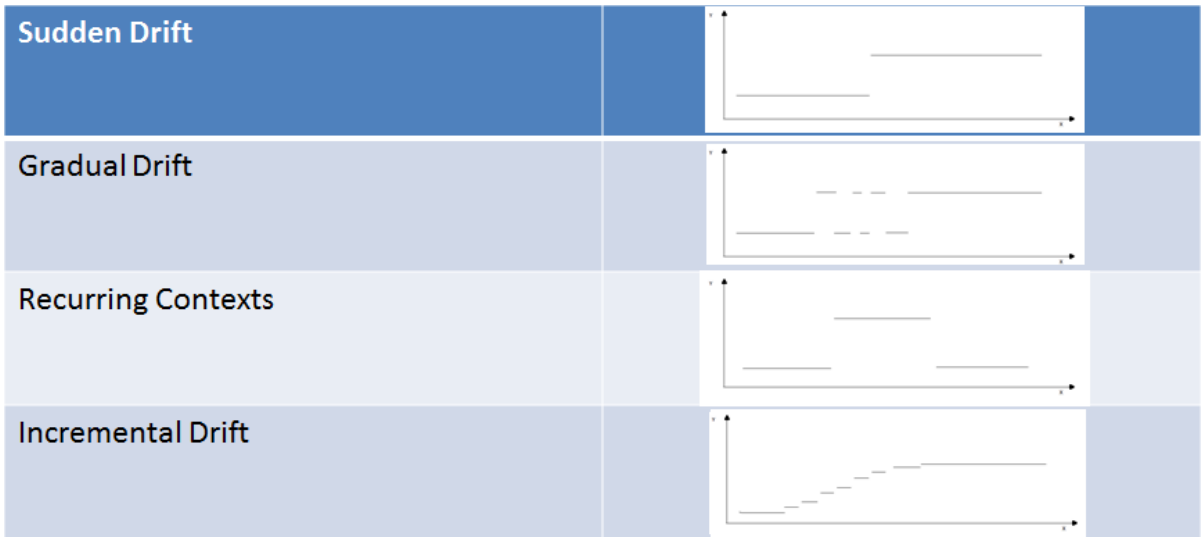


Figure 2.1: Types of concept drift studied in concept drift literature

2.2 Concept drift

2.2.1 Definition and examples

For a concept drift problem, there is a need to predict some unknown characteristic of an instance. The instance is preceded by multiple other instances, for which the characteristic in question is known. Concept drift is a term used to describe any change in the relationship between the attributes collected and the characteristic being predicted. Many algorithms have been formulated to address prediction and classification, but they rely on consistent relationships, across the data. If a concept drift has occurred, the models produced by these algorithms will be inaccurate.

An example of concept drift that is common in manufacturing may occur when an experienced operator is replaced with an inexperienced operator. In this situation, the physical output of a process or the recording of important measurements may experience an observable change. This change will affect the data recorded for the process output, and the important measurements. If an analysis of this data does not take this change into consideration, misleading information might be provided to decisions makers.

Concept drift has been divided into four distinct types of drift Zliobaite (2009). Identifying

the type of drift a dataset follows gives hints to the learning methods and models that will be successful. The types are expressed graphically as one measure verses time, in Figure 2.1.

The direction, magnitude, and time between drifts can give a hint to the nature of the data. In gradual drift, a steady concept begins to alternate with a new concept. Eventually the new concept becomes steady.

In recurring contexts, there is a constant switch between previously observed concepts. In the manufacturing example provided earlier, this could represent alternating between a set of operators.

Incremental drift is apparent when a steady concept begins to drift in one direction. As an extension of the manufacturing example, this might represent an increase/decrease in production as machines are incrementally added/removed to/from a bottleneck operation.

Finally, sudden drift is a shift in the underlying concept which occurs suddenly and remains stable for a period of time. This is the most basic type of drift and all other drift types contain this type of drift at some level. For instance, gradual drift experiences consistent sudden drifts between two contexts for a period of time. Incremental drift is a series of consistent sudden drifts. Recurring contexts is a a set of sudden drifts occurring as the distribution source switches between a number of contexts. If none of the other drift types can describe the data, the methods which perform well on sudden drift should be employed.

The realization of these different types of drifts has helped researchers identify the strengths and weaknesses of their methods. In application, identifying which type of drift is present can point to the best methods to apply. This will be reflected in the relevant results of previous studies, presented throughout this chapter.

Addressing change in systems from streaming data has become a popular research area. The problem is very common and, therefore, has been studied extensively in different fields. Some of the application areas are manufacturing, security, marketing, finance, medicine, and artificial intelligence. Hand (2006) points out that this has led to parallel research and many sets of terminology. This fact is used to support the argument that little advancement has been made in classifier technology. Zliobaite (2009) provides tables summarizing similar terms and application areas.

2.2.2 Learning and adapting in the presence concept drift

In order to create a system that can adapt to concept drift, the decision models that are used within that system may be adaptable. A change in the instances selected for learning the decision model could also allow a system to adapt to a concept drift. Most of the approaches studied in this thesis are referred to as incremental learning approaches. This means they make decisions about one or more current objects or situations based on past observations. The current model is evaluated by how it performs on the latest observation. Incremental learning and the term "concept drift" were introduced in Schlimmer and Granger (1986).

The process control problems studied in this thesis are specific instances of sudden and incremental concept drift. This thesis focuses on algorithms for addressing those types of drift.

2.2.2.1 Instance weighting

In instance weighting, it is common to decrease the weight of an instance indirectly to the age of the instance. For example, at each new observation, exponential weighting multiplies each instance's contribution by a constant $0 \leq r \leq 1$. Equation 2.3 represents an exponentially weighed mean at iteration t.

$$EWMA(\vec{x}) = \lambda \cdot \sum_{i=1}^{n}((1-\lambda)^{n-i} \cdot x_i) \tag{2.3}$$

2.2.2.2 Sliding windows

One simple and effective method for learning under concept drift is to learn a model on a sliding window. The sliding window forgets previous observations as new observations enter the set. This simple and easy-to-implement approach can quickly improve the accuracy of prediction and classification in the presence of concept drift. Because of this, it is the basis for many of the algorithms for incremental model learning under concept drift. A question arises, "what size should the window be?"

2.2.2.3 DDM and EDDM

One of the most popular methods for addressing window size is the use of a trigger. The trigger tells incremental learning algorithms that a concept drift may be present. The window size grows until a drift is detected.

The drift detection method (DDM) is a popular example of a trigger (Gama et al. (2004)). The method is formed from the idea that, given a stable data stream, the proportion of errors will decrease as a continuously learned model observes more instances. DDM detects drift by comparing the mean and standard deviation of the error in the most recently observed subgroup to the minimum error and standard deviation that has been observed.

Baena-García et al. (2006) introduces early drift detection method (EDDM). EDDM is a modification of DDM that detects drift from the expected time between errors, rather than the proportions. It assumes that as the number of observed cases increases, the time between errors will decrease, for a continuously learned model.

Both EDDM and DDM build control limits and signal a drift when an observation exceeds the control limits. This is similar to the first rule in control charting, which states that observations exceeding the control limits are likely from special causes. Like many control charts, the values three standard deviations away from the mean are the suggested control limits.

Both methods set a second level of limits. This second level is not as far from the mean and serves as a warning, rather than a trigger. The second level limits are often set at two standard deviations from the mean. Observations exceeding the second level are kept in memory. If a drift is detected, the stored observations are used to build the new model.

While statistical control charts have inspired the drift detection methods, the exact methods used in statistical control charting are not mirrored in the current popular methods. Before statistical control charts are initialized, the process for which they are meant must be stabilized. In concept mining, this means a large enough number of instances has to occur, before the algorithm can be initialized. The closest DDM and EDDM get to doing this is requiring 30 observations before detecting drift. The number 30 is chosen for the implications it has on sample statistics in the normal distribution. These implications are voided when a concept

drift has occurred during the 30 observations.

2.2.2.4 ADWIN

ADWIN (Bifet and Gavalda (2007)) splits a current window into two windows. The method tests for a difference between the two windows' means greater than ϵ_{cut} in equation 2.4. This cut is a modification of the Hoeffding bound that is adapted to the variance across the window (σ_w^2). The first split considered is the most recent. The method continues the search by considering splits at older time points, until a difference is detected or all data in the window considered has been searched. If a difference is detected, the oldest element of the total window is eliminated, and the search repeats.

$$\epsilon_{cut} = \sqrt{\frac{2}{m} \cdot \sigma_w^2 \cdot ln\frac{2}{\delta'}} + \frac{2}{3m} \cdot ln\frac{2}{\delta'} \tag{2.4}$$

$$m = \frac{1}{\frac{1}{n_0} + \frac{1}{n_1}} \tag{2.5}$$

To reduce the memory required, Bifet and Gavalda (2007) also presents ADWIN2. Besides computational costs, the algorithms are considered similar enough that ADWIN2 is used in all experiments.

A set of data streams are created of varying lengths from a Bernoulli distribution. The mean of the distribution begins to increase linearly, after a number of data points. When used as a change detector on the Bernoulli data streams, ADWIN2 outperforms the drift detection method (DDM) on longer data streams. Bifet et. al. explain that this is because of the influence of past examples within the DDM algorithm.

As a method for adjusting the training window for machine learning algorithms, ADWIN is generally applied to the error of a classifier. Zliobaite and Kuncheva (2009) implement ADWIN on each independent variable of an n-dimensional data stream. The independent windows are used to learn a nearest mean classifier.

2.2.2.5 $WRandWR^*$

The window resize (WR and WR^*) algorithms, proposed by Zliobaite (Zliobaite and Kuncheva (2009), Žliobaitė (2009), Žliobaitė (2011)), use a detection on raw data. In other words, the observation's position in attribute space is considered, rather than the error of the current model.

In the window resize algorithm (WR), each possible split of the data, since the last detected concept drift, is considered. A Hotelling's two-sample T^2-test is performed on the samples of from each class in the two subsets formed by each possible split. The split which minimizes the product of the Hotelling's two-sample T^2 p-values across all classes is considered the drift point, as in equation 2.6. In WR, the new window only considers data since the detected drift point (t_D), for training a classifier. In other words, the window size is $t - t_D$).

$$t_D = \underset{j=1}{\overset{t}{\operatorname{argmax}}}\{1 - \prod_{k=1}^{c} P(no\,change\,in\,\mu_k|d)\} \tag{2.6}$$

A modification of the window resize algorithm (WR^*) does not immediately correct the training window. Instead, the appropriate size (N^{WR^*}) given the expected error is calculated and the training window is only adapted when that number of observations have been observed, since the detected drift point($t = t_D + N^{WR^*}$).

2.2.2.6 Reoccurring contexts and gradual drift

The problems studied in this thesis follow sudden and incremental drift types. When generalizing the application of concept drift methods to process control and process improvement, it is clear that data following reoccurring contexts or gradual drift types are probable (e.g. two operators with different tendencies or an operator choosing between two available tools). For this reason, this thesis briefly explores the literature of these types of drifts.

In reoccurring drift, a previously abandoned model may become useful again, if new data is similar to a past concept. Methods like those in Bifet et al. (2009), Nishida and Yamauchi (2009), Becker and Arias (2007), and Zhang and Jin (2006) use ensembles to address this idea. In most ensemble methods, votes are given to different classifiers. Weights may be assigned to

the votes by characteristics of the current instance or recent accuracy of the classifiers.

FLORA (Widmer and Kubat (1996)) keeps a window of trusted examples and hypotheses. It also stores concept descriptions for re-occurring contexts. A heuristic, which monitors the behavior of the system, determines the contents of this stored information.

In the FISH algorithms ($FISH$, $FISH2$, and $FISH3$) time and attribute values are normalized to a unitary scale. After this, the distance measure is obtained by a scaled combination of Euclidean distance in the attribute space and the distance in time. Equation 2.7 shows the calculation of the distance measure (D_{ij}) between two observations, i and j. $d_{ij}^{(s)}$ and $d_{ij}^{(t)}$ represent the normalized distance in space and time, respectively. a_1 and a_2 are the weights used to scale the two distance measures. $FISH$ uses this distance measure and a k nearest neighbors approach to select instances for training. The FISH algorithms are especially effective for recurring contexts and gradual drift. Experiments on several datasets show that the best values for a_1 and a_2 vary highly, from $a_2 : a_1 = 1 : 0$ to $a_2 : a_1 = 0 : 1$.

Since $FISH$ requires the off-line evaluation of the best number of neighbors (k) to operate, $FISH2$ was developed. $FISH2$ considers subsets of size (s), where s increases by k until $s = n$. $FISH2$ learns classifiers on the s nearest neighbors of each observation (X_t). Then, the classifiers are tested on the set formed by the k nearest neighbors of the observation in question (X_{t_n}). Once this has occurred, the optimal size is the s which formed the classifier with the lowest error. $FISH2$ was compared to similar algorithms and a classifier built on all of the data. $FISH2$ was more stable than the other methods across different levels of drift and classification models.

$FISH3$ is a modification of $FISH2$ that also searches for the best weights ($a1$ and $a2$) to assign to attribute space and time distances. In Zliobaite's experiments, $FISH$, $FISH2$, and $FISH3$ have similar performance. $FISH3$ appears to give less error than $FISH$ and $FISH2$, overall.

$$D_{ij} = a_1 d_{ij}^{(s)} + a_2 d_{ij}^{(t)} \tag{2.7}$$

CHAPTER 3. METHODS AND PROCEDURES

This thesis work is based on the findings from the first analysis of the industrial study data. So, this chapter begins by discussing the initial analysis and identification of measurement drift.

Although the drift detection methods used in this work are built from existing ideas within concept drift detection literature, the exact methods are unique to this study. In particular, the application to incremental learning of regression models. The methods for generating synthetic measurement drift data is discussed. The adaptation of the concept drift algorithms is described. Two new algorithms, ADMEAN (ADaptive MEAN) and CD-EWMA (Concept Drift Exponential Weighted Moving Average) are presented. Finally, a method for generating synthetic concept drift data is introduced.

3.1 Measurement drift

This section begins by presenting the mathematical model representing the measurement drift problem. The role measurement drift plays in the prediction of y is given. Identification of measurement drift is discussed. Examples are provided from the industrial study. Finally, the method for creating synthetic measurement drift data for the experiments is explained.

3.1.1 Measurement drift model

In equation 3.1, the value of the input variable j at observation t, $M_{j,t}$, is given as a function of its components, measurement drift ($\psi_{j,t}$), the true attribute value ($x_{j,t} \sim N(\mu_j, \sigma^2_{x_j})$), and measurement error ($\epsilon_{j,t} \sim N(0, \sigma^2_e)$).

$$M_{j,t} = x_{j,t} + \epsilon_{j,t} + \psi_{j,t} \tag{3.1}$$

Since $M_{j,t}$ is our input variable, it is useful to observe its role in the linear model in equation 3.2 , where y_t represents the observed value of the response variable w.r.t. m input variables.

$$y_t = \beta_0 + \sum_{j=1}^{m} (\beta_j(M_{j,t}) - \beta_j(\epsilon_{j,t} + \psi_{j,t})) \tag{3.2}$$

When predicting the response variable, the true measurement error or measurement drift values are unknown. They must be estimated in some way. In equation 3.3, $\hat{y}_t$ represents the prediction of the response variable, where the prediction of ϵ is its expected value (i.e. zero) and the predicted measurement drift is $\hat{\psi}_{j,t}$.

$$\hat{y}_t = \beta_0 + \sum_{j=1}^{m} (\beta_j(M_{j,t}) - \beta_j(\hat{\psi}_{j,t})) \tag{3.3}$$

One of the key characteristics of the learning algorithms in this thesis is the ability to retain knowledge about a system from data occurring before a drift. Existing methods for retaining knowledge, within concept drift literature, focus on instance selection. The methods identify previous instances similar to the instance(s) that the model will predicted or classify. The instances identified as similar are used to build the prediction or classification model that will be used. This is particularly useful in the presence of reoccurring contexts, since the new data comes from a concept that has been previously observed. It can also be useful in gradual drift, since the intermediate steps of the full drift involves alternating between concepts.

The models proposed in this thesis, however, retain information from a previous model. Specifically, the information we want to keep is the slope of a linear model. The slope is retained while the intercept is changed. As seen in in equation 3.4, the difference between the predicted value and the real value is shown to be constant. This means that the change in the linear model is expressed in the intercept. Equation 3.5 is obtained from equation 3.2 and calculates the intercept value at time t.

$$\hat{y}_t - y_t = \sum_{j=1}^{m} (\beta_j(\epsilon_{j,t} + \psi_{j,t}) - \beta_j(\hat{\psi}_{j,t})) \tag{3.4}$$

$$\begin{aligned} INT_t &= \beta_0 - \textstyle\sum_{j=1}^{m}(\beta_{j,t} \cdot \psi_{j,t}) \\ &= y_t - \textstyle\sum_{j=1}^{m}(\beta_{j,t} \cdot M_{j,t}) \end{aligned} \tag{3.5}$$

In the presence of concept drift, the measurement drift value changes (i.e. $\psi_{j,t-1} \neq \psi_{j,t}$). The difference in $\psi_{j,t-1}$ and $\psi_{j,t}$ can be estimated before observing the response variable or after observing the response variable (i.e. proactively or reactively).

In the proactive approximation of the new intercept, the change in the input variable is observed. The result in equation 3.8 would allow adaptation before an erroneous prediction has been implemented. If drift is detected before making a decision, this is our best guess at the new value.

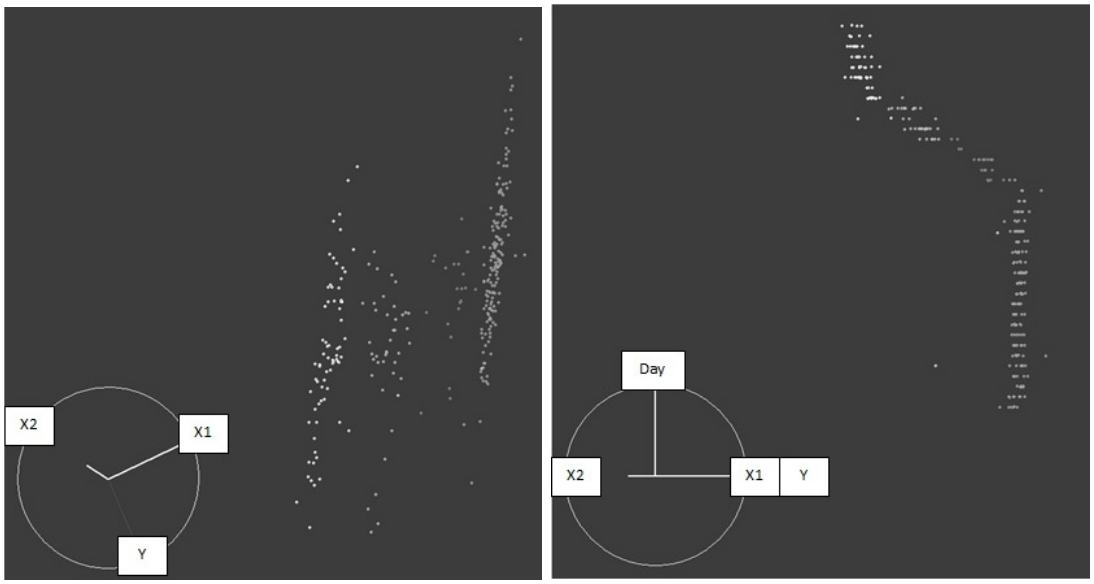
$$\hat{\psi}_{j,t} = M_{j,t} - \mu_j \tag{3.6}$$

Unfortunately, the effectiveness of the proactive approach relies on a large difference in measurement drift compared to the stacked variance in the measurement ($\sigma^2_{x_j}$) and the measurement error($\sigma^2_{\epsilon_j}$). It might be beneficial to wait until after the true response variable has been observed before adjusting the model. Equation 3.7 gives the intercept value at observation t. The approximation is effective when the drift is large w.r.t. the variance in measurement error($\sigma_{\epsilon_{j,t}}$). The reactive approximation is less sensitive and more accurate than the proactive approximation, but it relies on a steady concept between $t-1$ and t.

$$\hat{\psi}_{j,t} = y_t - \beta_0 - \sum_{j=1}^{m} \beta_j (M_{j,t-1}) \tag{3.7}$$

These findings will be considered in the frameworks of the concept drift methods. Variations of the measurement drift prediction will be used for adjusting linear prediction models in the presence of concept drift. Equation 3.8 shows the intercept as a function of $\hat{\psi}_{j,t}$.

$$INT_t = \beta_0 + \sum_{j=1}^{m} \beta_j (\hat{\psi}_{j,t} + \mu_{j,t}) \tag{3.8}$$



(a) Input and outcome variables (b) Attributes vs. time

Figure 3.1: Projections of study data with color indicating group (i.e. concept)

3.1.2 Identifying measurement drift

Now that measurement drift has been defined, the identification of measurement drift can be explained. An example of the data obtained for the industrial study can be seen in the 2-D projections in Figure 3.1. The projection in Figure 3.1 (a) emphasizes the linear relationship between the outcome variable and measurement attributes, which is held throughout the data. The visualization depicts very clear groups of data. This is the first hint that measurement drift may be present.

In Figure 3.1 (b) the dimension depicting time has been added. To emphasize the temporal nature of the data, the data is projected with the time axis perpendicular to all other axes. The data points retain the coloration given in Figure 3.1 (a). It is clear that, as time progresses, the data is moving with respect to the measurement and outcome variable axes. Further, it appears that the diata in each group remain very near each other. This gives further support for the existence of measurement drift. It also shows that the data follows the sudden drift type explained in section 2.2. If two groups appear to be occurring simultaneously, recurring contexts or gradual drift are likely present.

Finally, the existence of measurement drift can be seen in the histograms of the observed

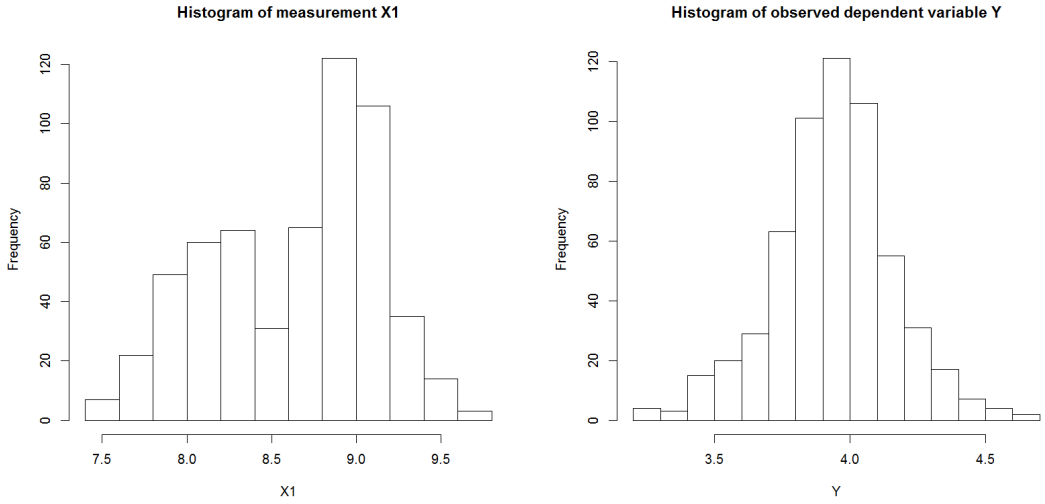


Figure 3.2: Measurement and passing y value histograms demonstrating measurement drift

data. Figure 3.2 provides an example of this from the study data. It seems apparent that the correct value of Y, needed to produce a non-defective part, originates from a single distribution. Meanwhile, the multimodal behavior of measurement $X1$ suggests that $X1$ does not originate from a single distribution. The combination of these two traits is a very good indication of measurement drift.

3.1.3 Measurement drift synthetic data

The data used in the study allows testing against a real dataset. However, it is only one representation of the problem. Data was synthesized to allow comparison of the methods to data with other characteristics.

All datasets contain 640 instances. Each instance has two measurements, with one outcome variable. Data is synthesized for four drift to variance ratios (0.5, 1, 2, and 4) against four levels of drift frequency (10, 20, 40, and 80), for a total of 16 synthetic datasets. The drift to variance ratio gives the magnitude of drift as a number of standard deviations from the mean of the outcome variable. Drift delay is the number of observations until a drift occurs. All drift is expressed in the positive direction away from the mean. This allows an evaluation of whether an algorithm is over adapting or under adapting by observing the mean in prediction

error. A positive value indicates a method is slow to adapt. Measurement error was included according to a normal distribution with mean equal to zero and variance equal to 0.025 times the attribute variance. An example of the data is shown in Figure 3.3.

3.2 Process control algorithms

The process control algorithms detect drift from attribute space or error between the predicted value and its observed true value. They then adjust the data from which a model is learned, adjust a regression model, or both.

In the presence of measurement drift, model error provides the best indication of drift. When observing the prediction error, variance in the true measurement values (σ_{x_j}) is removed from the equation. To visualize this, a measurement drift has been synthesized. The predicted values from a single model are shown in Figure 3.4 (a) against the actual values. The separation is clear. The error on this data, by a fixed prediction model, can be seen in Figure 3.5 (a). The difference in mean before and after the shift is directly applicable to the model.

The exponential weighted moving average reduces noise, allowing an even better distinction. This can be seen in Figure 3.6 (a). The chart method, in the experiments, takes advantage of this enhanced distinction.

3.2.1 Control charts

In section 2.1.1, different control charts are discussed. For the experiments in this thesis, a control chart like method is employed. For simplicity, the method is referred to as "Chart." Chart is a combination of a standard control chart, an exponential moving average chart, and the drift detection method (DDM).

The exponential moving average chart is used because of its ability to detect small drifts. As recommended in Roberts (1959), the weight (ω) is set to $\frac{2}{5}$. The error and an exponential moving average of the error are kept. The exponential moving average of the error is compared to the control level calculated from the standard deviation of the error(σ_{error}), the level k, and the number of observations (n) as $k \cdot \sigma_{error} \cdot \sqrt{\frac{\omega}{2-\omega} \cdot \left(1 - (1-\omega)^{2 \cdot n}\right)}$. The control level of the standard control chart in Chart is calculated as $k \cdot \sigma_{error}$.

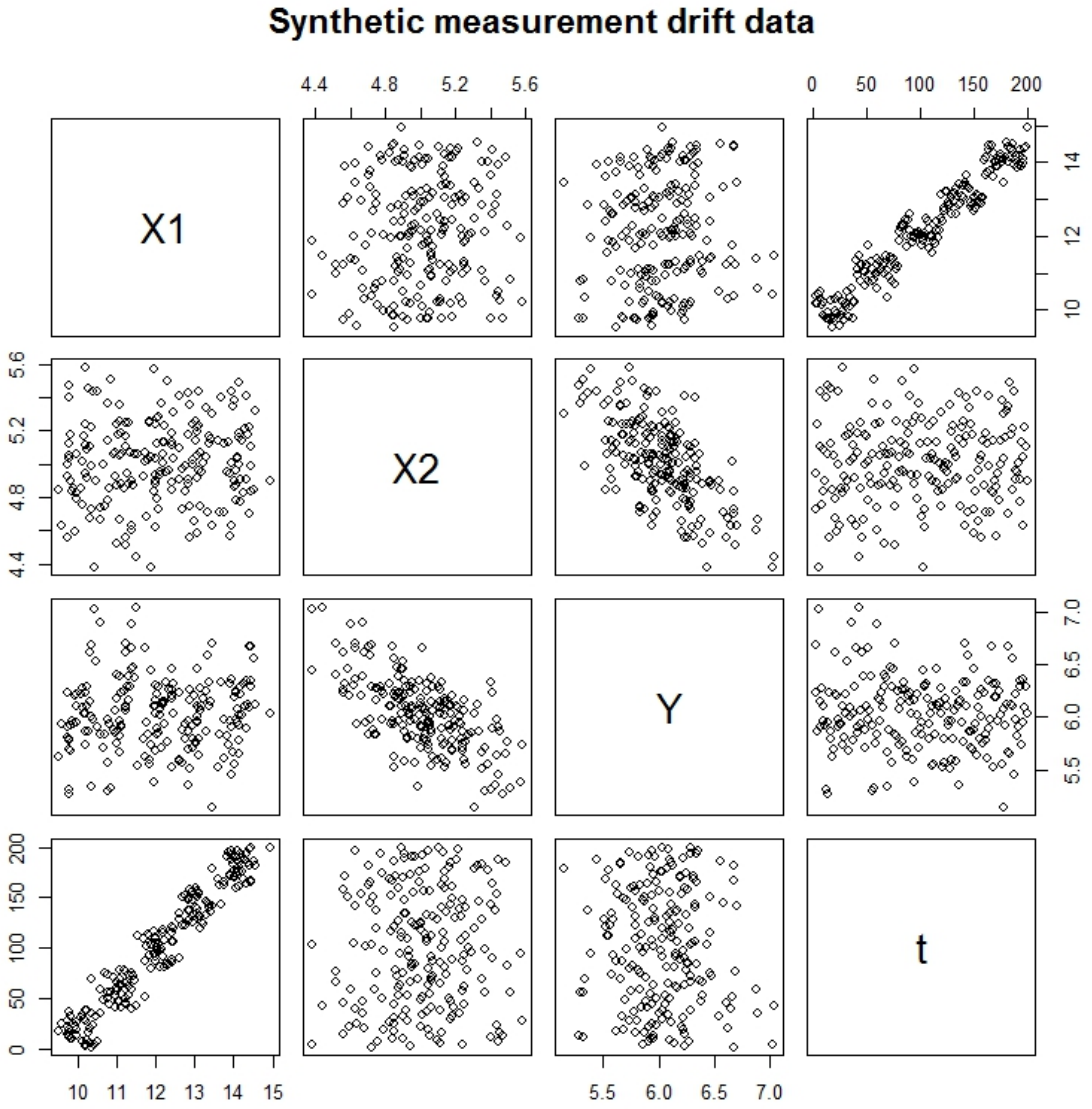
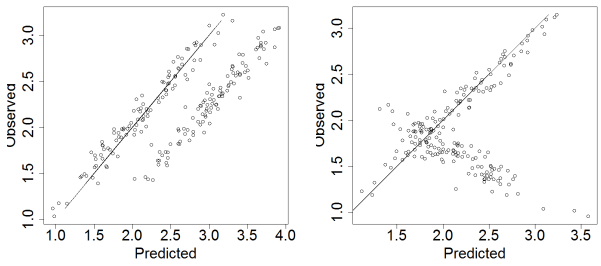
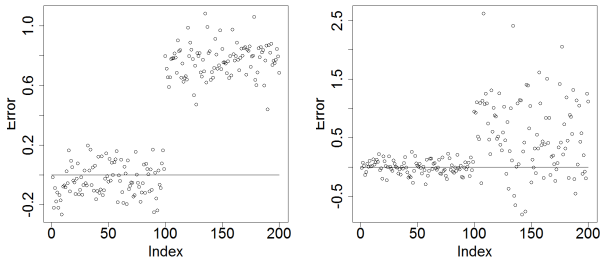


Figure 3.3: Example of synthetic data representing measurement drift (drift to variance = 4, drift delay = 40)



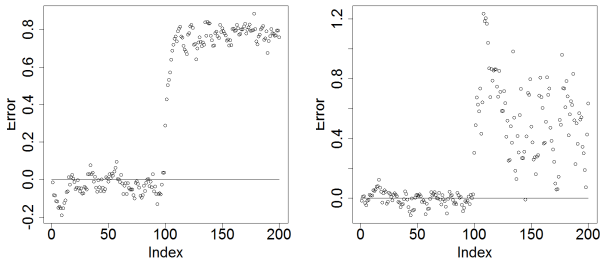
(a) (b)

Figure 3.4: Measurement drift (a) and gain drift (b) predicted versus observed



(a) (b)

Figure 3.5: Measurement drift (a) and gain drift (b) linear model error



(a) (b)

Figure 3.6: Exponential error for linear model prediction in the presence of measurement drift (a) and gain drift (b)

Chart retains two levels, similar to the drift detection method mentioned in section 2.2.2.3. The model enters the warning mode if either the EWMA-chart or the standard control chart records an observation exceeding the warning level ($k = 2$). The observation at which the model entered warning mode is remembered until any new observation does not exceed the warning level for either the EWMA-chart or the standard control chart. The window grows until the signal level ($k = 3$) is exceeded. After the signal level is exceeded, a new window starts and the model enters learning mode. If the model is in warning mode when the signal level is exceeded, the observation where the warning level was exceeded becomes the first observation of the new window. Otherwise, the observation exceeding the signal level is the beginning of the new window.

In stabilization mode, Chart uses a fixed sliding window to form a prediction model. For the experiments in this thesis a fixed window of 5 observations was used in the stabilization mode. This is necessary to allow learning when the dimensionality does not permit a proper model. A small window is used in case drift is not exactly sudden and takes some time to stabilize. Like DDM, after 30 observations, the model leaves stabilization mode. The new window is now used for learning a prediction model. Monitoring of the error and exponential moving average of the error begins, again.

Although the problem is not studied heavily in this thesis, the same visualizations are given for gain drift (Figure 3.4 (b), Figure 3.5 (b)), and Figure 3.6 (b). The control chart method would detect a gain drift, because of the change in variance of the error.

3.2.2 EWMA-I

As mentioned in section 2.1.2, exponential weighted moving average (EWMA) controls have lots of support within the literature and have been shown to perform well in the semiconductor industry. The EWMA algorithms are directly applicable to the measurement drift problem. Because the exponential weighted moving average for intercept adjustment (EWMA-I) fits the data structure and is popular in advanced process control literature, the experiments in this thesis will use EWMA-I as a representative of the advanced process control algorithms.

When applying EWMA-I to the measurement drift problem, the intercept calculation from

equation 3.5 and the EWMA calculation from equation 2.3 are combined. We want to know the intercept term for the current time, t_n. However, the intercept calculation requires an instance with an observed outcome variable. So, an assumption is made that the intercept at time t_n is approximately equal to the intercept of preceding observations. The purpose of using an exponential weighted moving average is to give the most recent values of INT a larger influence on the prediction of INT_t. Equation 3.9 gives the calculation of the predicted intercept value at time t_n.

For the measurement drift experiments on the industry data, λ was set to $\frac{2}{5}$ because of the performance of this weight in preliminary experiments. To show the range in results of varying weights, the experiments on the synthetic measurement data tested EWMA-I with the weights 0.2, 0.5, and 0.8.

$$\begin{aligned} \hat{INT}_{t_n} &= \lambda \cdot \textstyle\sum_{i=1}^{t_n-1}((1-\lambda)^{(t_n-1)-i} \cdot INT_j) \\ &= \lambda \cdot \textstyle\sum_{i=1}^{t_n-1}((1-\lambda)^{(t_n-1)-i} \cdot (y_i - \sum_{j=1}^{m}(\beta_{j,i} \cdot M_{j,i})) \end{aligned} \tag{3.9}$$

3.3 Concept drift algorithms

In this section, the relevant algorithms from the concept drift literature are shown in detail. ADMEAN and CD-EWMA are contributions. They are first proposed in this thesis.

3.3.1 WR and WR*

Zliobaite's window resize (WR) algorithm performs a backward search on the attribute space for change point detection. In these experiments Hotelling's T^2 test was implemented using the "Hotelling" package in R, the project for statistical computing. A 95% confidence level was used. As in equation 2.6, on page 14, the time at which splitting the window maximizes the probability of change ($\text{argmax}_t\{1 - \prod_{k=1}^{c} p_k(t)\}$) is the chosen change point.

The window size in WR includes all data from the time of the change point to the last instance with an observed outcome variable. If singularity does not allow linear discriminant analysis, the class is determined using nearest means classification. To conserve time, the backward search in WR extended only to 150 samples.

$W = W \cup x_{j,t_n}$
repeat
$W = W - 1$
$w_0 = elementsx_{j,t_n-W} \rightarrow x_{j,t_D-1}$
$w_1 = elementsx_{j,t_D} \rightarrow x_{j,t_n}$
until $pvalue\,(t - test\,(w_0, w_1)) > \alpha; \quad t_n - W + 1 \leq t_D \leq t_n - 2$ **return** t_D

Figure 3.7: ADWIN algorithm with using t-test with equal variance

In WR*, the data occurring after a detected change point is not used until N* instances have been observed. The methods for calculating N* are explained in Zliobaite and Kuncheva (2009).

For WR and WR*, a nearest means classifier is used until the covariance matrix of the data occurring after a change point is non-singular. During this time, N* is calculated according to an expectation of the error a nearest means classifier would experience on the new data. Once the data produces a non-singular covariance matrix, linear discriminate analysis can be performed, and a linear discriminate classifier (LDC) is applied. The expected error of a LDC is then used for the calculation of N*.

3.3.2 ADWIN

The ADWIN algorithm was applied to each attribute individually. As mentioned in the literature review, on page 13, the algorithm performs a backward search and shrinks the window until a search returns no significant differences between values before and after a split. Rather than determining significance with a cut point based on Hoeffding bounds, we implement a student's t-test with a 95% confidence level. Since the measurement drift model, given in section 3.1.1, does not include a change in variance, it is reasonable to assume equal variance before and after the split.

In the experiments in this thesis, ADWIN is applied to each attribute and the minimum size returned is used for learning. As with the implementation of WR, if the dimensionality and observed classes do not allow linear discriminant analysis, a nearest means classifier is used. To conserve time, the backward search in ADWIN extended only to 150 samples.

3.3.3 ADMEAN

ADMEAN is a combination of sliding windows and the ADWIN algorithm. A backward search for change point detection is applied to all attributes of the data stream. The backward search extends to ws elements, where ws is a fixed window size of the user's preference. For the experiments, $ws = 30$.

Since we cannot often guarantee a range $[a, b] \ni a \leq x_{j,t} \leq b$ over the attributes of the data studied, the suggested Hoeffding bound cut is not feasible. Instead, a t-test assuming equal variance at a 95% confidence level is used to describe a "significant enough" difference between the data before and after a split. If the null hypothesis is rejected for any split in the window, the suggested window size is decreased by one, and the backward search is performed again. This is repeated until no backward search over the suggested window leads to rejecting the null hypothesis of the t-test.

Once the suggested window size is determined, ADMEAN does not discard older data. Instead, all data contained by the sliding window used in the backward search are kept. ADMEAN splits the sliding window into data from the old source and data from the new source. The difference between means is subtracted from data appearing after the split. This is like removing the measurement drift ($\psi_{j,t}$) from the measurement. A model is learned over the fixed sliding window of modified data. Equation 3.10 and equation 3.11 show the calculations of the expected drift value ($\hat{\psi_{j,t}}$) and expected value of the new data without drift ($\hat{x}_{j,t}$).

$$\hat{\psi}_{j,t} = \frac{1}{t_n - (t_D - 1)} \sum_{t=t_D}^{t_n} M_{j,t} - \frac{1}{t_D - (t_n - ws)} \sum_{t=t_n-ws}^{t_D-1} M_{j,t} \tag{3.10}$$

$$\hat{x}_{j,t} = \begin{cases} M_{j,t} - \hat{\psi}_{j,t}, & t_D \leq t \leq t_n \\ M_{j,t}, & t_n - ws \leq t \leq t_D - 1 \end{cases} \tag{3.11}$$

Figure 3.8 illustrates the effect of applying ADMEAN to a dataset containing drift in one of the classes. The circles represent data occurring before the drift. The squares represent data occurring after the drift. The arrows are the length and direction of the mean before drift subtracted from the mean after the drift. The four pointed stars represent the data occurring

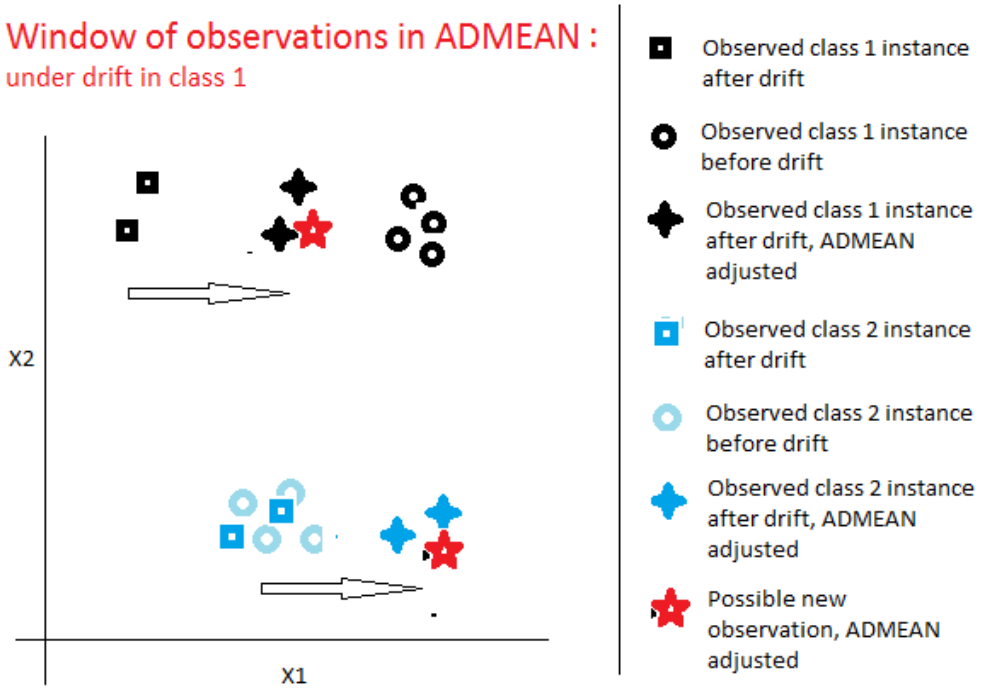


Figure 3.8: Illustrated data adjustment by ADMEAN

after the split with the difference in means subtracted. The five pointed star gives an example of a current observation, for which we do not know the class. The expected drift has been subtracted from the new observation.

3.3.4 CD-EWMA

Before the class of an instance is observed, information about that instance is given in the attributes we do have. Feed forward adaptation hopes to adapt the prediction model before making a prediction. CD-EWMA is introduced here as a method for feed forward adaptation. Although this type of adaptation subjects the model to variance in the attribute space ($\sigma^2_{x_j}$), on top of the variance in measurement error (σ_{ϵ_j}), it provides early adaptation to a concept or measurement drift.

The first consideration is that the concept drift problem is generally a classification problem. EWMA controls were developed for modifying a numerical prediction model. However, the assumptions remain the same. The instances occurring most recently will likely give the most

information about the current instance. This gives hope to applying EWMA at the attribute level.

CD-EWMA is a modified version of the EWMA controls discussed in section 2.1.2, on page 7. To allow application of the method to classification models, the method must predict the drift from the attribute values. When predicting drift from the attribute values in the data stream, the algorithm is subject to a larger variance. The variance in x_j (σ_{x_j}) is added to the variance given by measurement error (σ_{ϵ_j}). This means the prediction is less accurate. However, the attributes are available before the class. So, an opportunity exists to correct drift in the current observation.

EWMA observes the exponential moving average of the error given when applying a model, to provide a suggested change in the model. CD-EWMA applies an exponential moving average to the difference in attribute values and their mean within the data stream, as seen in equation 3.12. The result is subtracted from the current observation. The new value ($\hat{x}_{t,j}$), from equation 3.13, is used for prediction.

When applied to the measurement drift synthetic and real datasets, a single prediction model is consistent throughout the experiment. In the concept drift data, a classification model is learned on a window of the observed instances. A window size of 30 instances was used for the concept drift experiments.

Before learning the model, the attributes of each instance are modified by CD-EWMA. Different weights were tested before the experiments and a weight of $\frac{1}{20}$ was chosen for application of CD-EWMA. The significant decrease in weight, when compared to $\lambda = \frac{2}{5}$ used in the measurement drift experiments, is due to the significant increase in variance present at the attribute level.

$$\hat{\psi}_{t,j} = \lambda(x_{t,j} - \bar{x}_j) + (1 - \lambda)\hat{\psi}_{t-1,j} \tag{3.12}$$

$$\hat{x}_{t,j} = M_{t,j} - \hat{\psi}_{t,j} \tag{3.13}$$

3.4 Concept drift algorithms for measurement drift problem

Concept drift algorithms have been developed primarily to accomplish learning classification models in the presence of concept drift. In this thesis, the goal is to study learning prediction models in the presence of concept drift. To accomplish this goal, the first step is understanding the meaning and structure of the algorithms within the prediction model context.

When applied to the attribute space, the concept drift algorithms transfer simply into the measurement drift problem. The window size is selected in the same way as it was in the classification problems containing concept drift. The observations in the window are used for learning a model with regression analysis. As with linear discriminant analysis, dimensionality is important. This is already built-in to the models for learning linear discriminant classifiers.

It is possible to take advantage of the measurement drift data structure. To do this, ADWIN-I and WR-I use linear coefficients provided from off-line analysis, similar to EWMA-I. The algorithms adapt the intercept from the given window using equation 3.7, shown on page 18. By only modifying the intercept, these models are able to retain knowledge about the relationship between independent variables while adjusting to measurement drift.

3.5 Concept drift experiments

The concept drift experiments in this thesis use synthesized data from the literature. Also included are real data sets with synthetic drift. Zliobaite and Kuncheva (2009) synthesizes drift in 2-class data by relabeling attributes after an arbitrary interval. Attribute one becomes attribute two. Attribute two becomes attribute three, and so on. The last attribute becomes attribute one.

The experiments on real data with synthetic drift, in this thesis, are obtained in a new fashion. Multi-class data is used to create a 2-class dataset with drift. for the fist 120 instances, class one of the multi-class data is class one ("1") of the 2-class dataset. Class two of the multi-class data is class two ("0") in the new dataset. After 120 instances, class three of the multi-class data becomes class two ("0") of the new data. After 100 more observations, the source for class one in the new dataset is class three from the multi-class data. Instances for the new dataset

Index	"1"	"0"
1 - 120	1	2
121 - 220	1	3
221 - 320	2	3

Table 3.1: Class from mult-class data used for selection of class "1" and "0" of the synthetic dataset containing concept drift, for each interval

are randomly selected iteratively with equal probability of class one ("1") or class two ("0"). Selections are made with replacement (i.e. once chosen, an instance from the multi-class data can be chosen again). Table 3.1 gives shows the sources of the two classes in the dataset with synthetic drift.

The data was obtained from the UCI repository. All four, Iris, Wine, Wine2, and Olive oil contain three classes. Iris contains 50 instances of each class with four attributes. Wine has 13 attributes and contains 59 instances from class one, 71 instances from class two, and 48 instances from class three. Wine2 has 12 attributes and contains 681 instances from class one, 638 instances from class two, and 199 instances from class three. The classes for olive oil are determined by the subregion of origin. In the olive oil data, class one contains 323 instances, class two contains 98 instances, and class three contains 151 instances.

CHAPTER 4. RESULTS

4.1 Industrial study

Setting parameters for these algorithms from off-line analysis allows them some adaptability to drift magnitude and frequency, and noise levels. However, there are some general results from the experiments that can give an idea for the strengths and weaknesses of each algorithm.

The industrial study data can be separated into three sets with unique characteristics. The data in the first set fit the measurement drift description well and the results give us a good look at how well these algorithms work in application to a model problem. A box plot of the results can be seen in Figure 4.1.

The next set of data from the industrial study showed high variance on the input and was more difficult for prediction. This indicates that the data experienced a gain shift or much more measurement error was present than in the other datasets. The linear model provided for the data by an expert of the process did not appear to be the best model for the study. Because of this, the methods which learn a regression model over the suggested window performed better than intercept adjustment methods. ADMEAN appears to be the best method for predicting in this dataset. This can be seen in Figure 4.2.

As seen in Figure 4.3, the third dataset from the industrial study was the hardest for prediction because of the outliers and a section of the data stream that appeared to alternate, after one to three observations, between two contexts. The CD-EWMA algorithm was unable to adjust appropriately to the outliers and alternating concepts.

WR-I and intercept adjustment over a fixed sliding window of the most recent ten observations performed the best. They were not followed far in performance by other intercept adjusting algorithms or WR. This shows that these algorithms are more adaptable to the issues

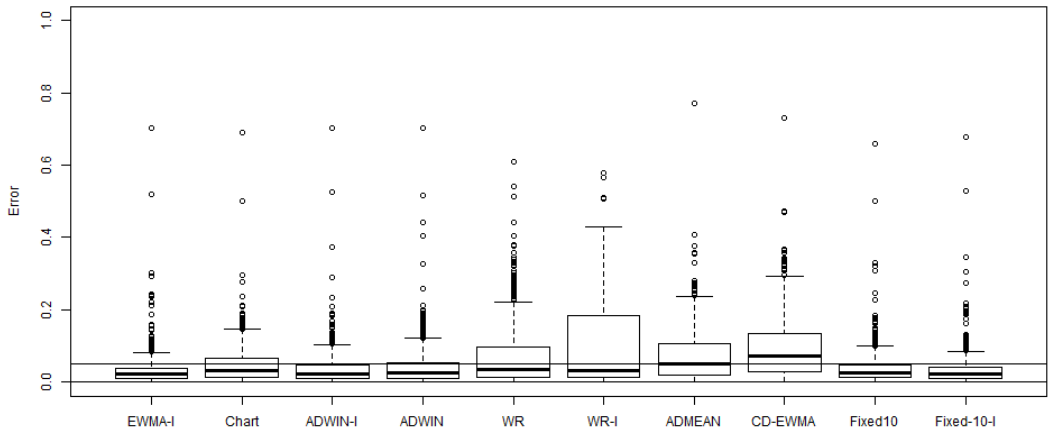


Figure 4.1: Absolute error for top performing algorithms on industrial study data exhibiting strong measurement drift patterns

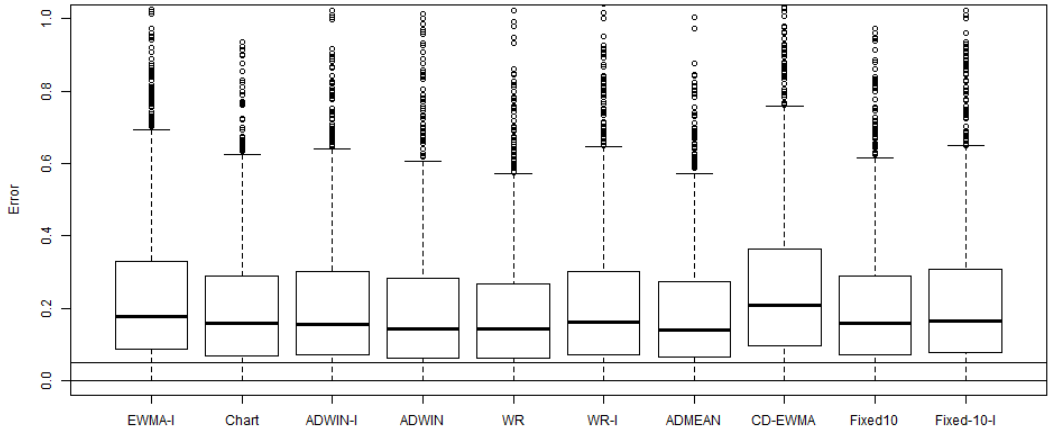


Figure 4.2: Absolute error for top performing algorithms on industrial study data exhibiting high variance and a possible change in linear relationship of dependent and independent variables

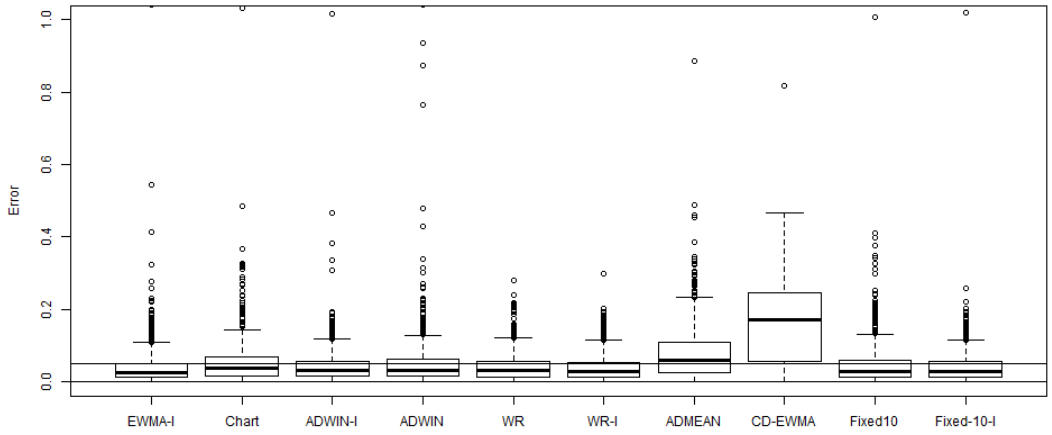


Figure 4.3: Absolute error for top performing algorithms on industrial study data exhibiting outliers and frequent alternating contexts

in the third data stream (i.e. outliers and frequent context alternations).

It is interesting to note that WR performed better than ADWIN. They are both learning a new model over a window but have different change point detection methods. This shows that change point detection on attribute space as a whole performed better than detection from each attribute individually.

When applied to the data from the industrial study, we find that the concept drift algorithms for mining data streams perform nearly as well, or better, than the EWMA-I control method. Overall, it is impossible to claim any method is superior to all other methods. Because of this, the simplicity of the algorithms becomes important. It appears that learning a regression model on a fixed window of size ten is a great choice for application to processes with the behavior seen in the industrial study.

The proposed method for learning in the presence of concept drift, ADMEAN, was competitive. However, it was only a good solution in the dataset containing high variance. If the variance was caused by a gain drift, there may not be a change point detection on the attribute space. However, the fixed size of the ADMEAN window allows an adaption to the new gain.

Figure 4.4 shows the size of the detected window. As expected, the window size for ADWIN is always equal to the window size for ADWIN-I when detecting from the attribute space.

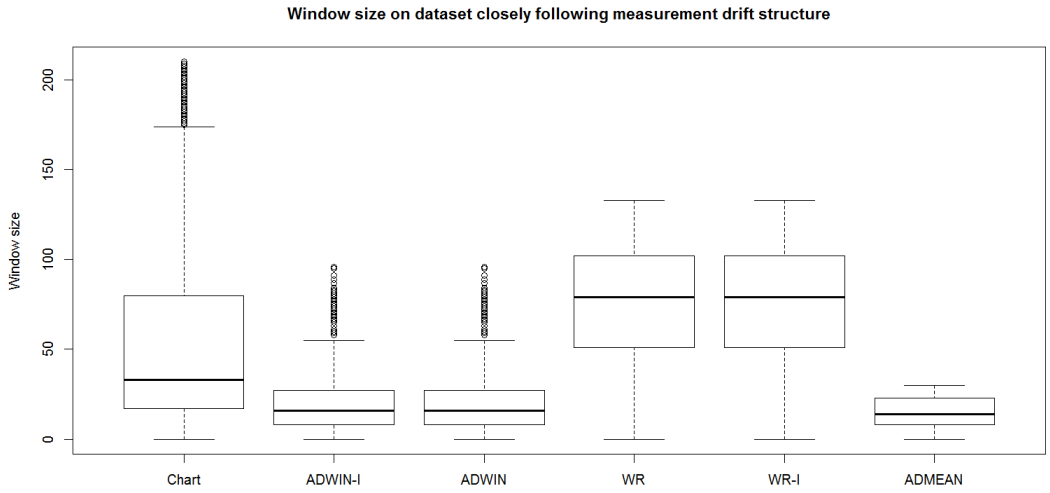


Figure 4.4: Box plot displaying window size for the data set which closely follows the measurement drift structure

Likewise, WR and WR-I will always have equal window sizes.

The window size for ADWIN is more likely to be smaller than the range it is searching over, on this dataset. On the other hand, WR is more likely to remain large. This shows that ADWIN's method for window sizing is more sensitive to the industrial study data than WR's window sizing method. When comparing the errors of the methods, this trait appears to give ADWIN the advantage for prediction in the presence of measurement drift (Figure 4.1). The results of applying the methods to data containing outliers were similar.

ADWIN continues to suggest a window smaller than WR in the dataset containing high variance (Figure 4.5). However, the window size suggested by the control chart method grows very large. The effect of this appears to be small, as the difference in error between the methods is insignificant (Figure 4.2).

4.2 Synthetic measurement drift data

The results of the application of experiments on the synthetic measurement drift data can be seen in the three dimensional visualizations of the absolute mean prediction error, in Appendix A. Also provided in Appendix A are visualizations containing two surfaces. The

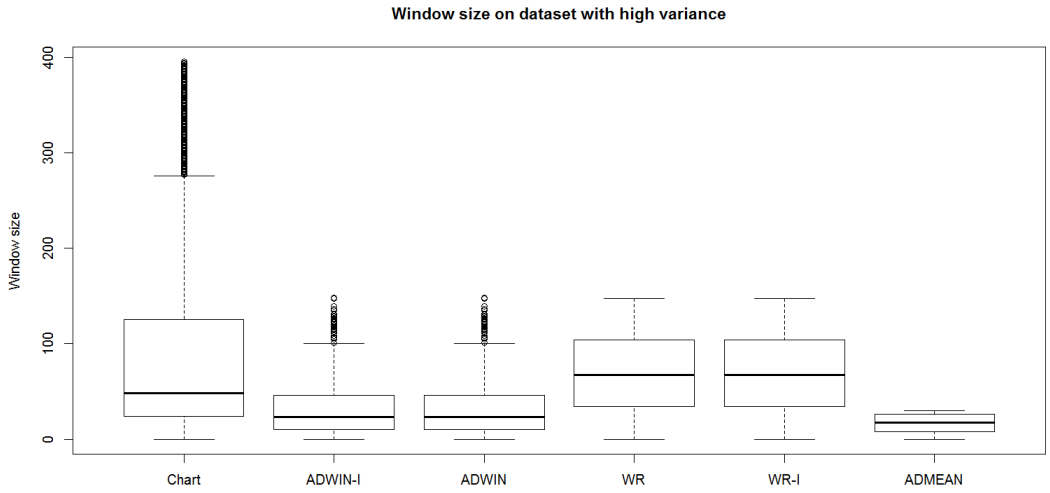


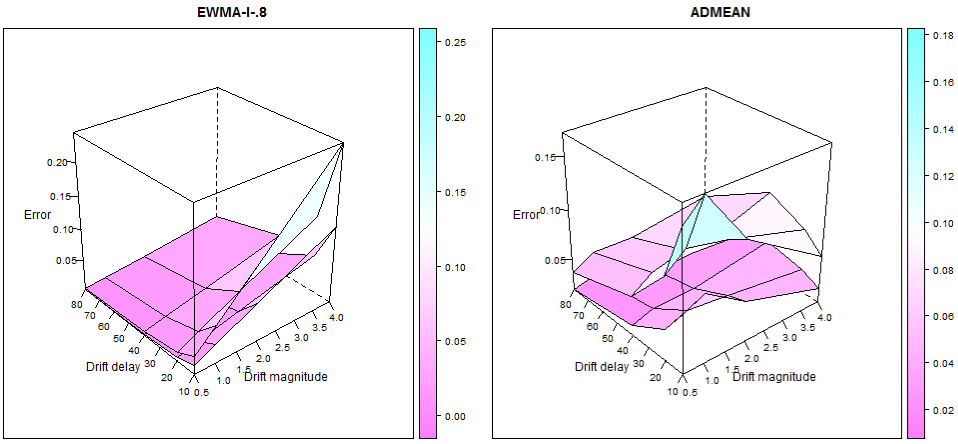
Figure 4.5: Box plot displaying window size for the data set which closely follows the measurement drift structure

topmost surface is the standard deviation of the prediction error added to the mean prediction error. The lower surface is the mean prediction error. The tables in Appendix B give the exact values of the means and standard deviations of prediction error for each algorithm tested.

ADMEAN performed well on the synthetic data. When compared to the best EWMA-I performer ($\lambda = 0.8$), an argument for implementing ADMEAN over EWMA-I can be made. In Figure 4.6, ADMEAN appears to be a great approach when the drift magnitude is high and drift occurs frequently. The worst performance for ADMEAN occurs at low drift magnitude with frequent drifts and remains low ($\approx .05$) relative to the highest error of other algorithms.

Overall, most algorithms were good at predicting in low magnitude drift occurring at large intervals. Except for CD-EWMA and fixed windows larger than 20 instances, all algorithms were near a mean absolute error of 0.10. The range of Y was approximately two. The algorithms also were generally better at prediction when the number of instances between drifts was large. When comparing the algorithms, the performance on data with frequent drifts (drift delay = 10 and 20) and high drift magnitude (2 and 4 times the variance) give the most diverse performance.

Of the concept drift problems, WR maintains the lowest prediction error across the synthetic



(a) (b)

Figure 4.6: Three dimensional of the mean and standard deviation of prediction error for EWMA-I $\lambda = 0.8$ (a) and ADMEAN(b)

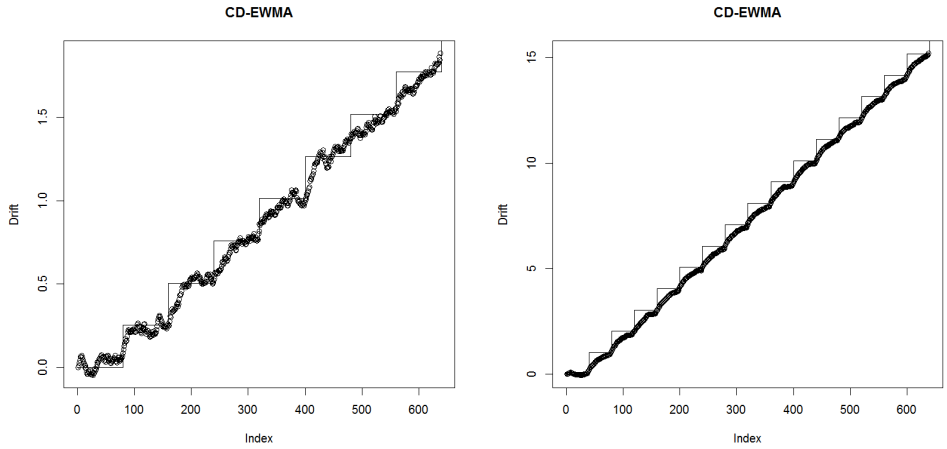
datasets. Outside of WR and ADMEAN, the concept drift algorithms are not competitive for predicting the outcome variable. CD-EWMA is especially poor. Slow adaption to the continuous drift in one direction left CD-EWMA reaching a prediction mean absolute error of 1.63 on an outcome variable with range approximately equal to two. This means continuously predicting the mean of the outcome variable would yield more a more precise prediction.

In the industrial study, a fixed window of the ten most recent instances yielded competitive results. On the synthetic data, a fixed window of size five was more accurate than ten. The mean and standard deviation of the prediction error incurred by a fixed window of size five shows that the method remains competitive with all other methods.

Ultimately, the EWMA-I method ($\lambda = 0.8$) displayed the best performance on the synthetic data studied. In cases where the drift magnitude to variance is large, ADMEAN is the recommended algorithm. If there is a chance for a large drift, but small drifts are also probable, ADMEAN remains a good choice for implementation.

The performance of ADMEAN and CD-EWMA rely on their ability to predict the measurement drift. In the synthetic data, the measurement drift is known. It is easy, then, to compare the predictions to the true value.

For CD-EWMA, the value is a stack-up of all drift. A plot of the predictions CD-EWMA



(a) Drift delay = 80, drift magnitude = 1.0

(b) Drift delay = 40, drift magnitude = 4.0

Figure 4.7: CD-EWMA predicted measurement drift

made for the measurement drift can be seen in Figure 4.7. A line represents the true drift. With a large drift delay and low magnitude (a), CD-EWMA is able to adjust to a drift. However, when the drift occurs frequently and with greater magnitude, CD-EWMA is not able to react quic enough. The area between the actual drift line and a line connecting the predictions would represent the total error. The error quickly adds up when the reaction is slow and the drift magnitude is high.

For ADMEAN, good predictions rely on two things. The window size must be selected to closely follow the drift pattern. Without identifying the correct window size, the diffence in means of the split will not be accurate. The second thing ADMEAN relies on is a good representation of the new distribution. The window size and accuracy can be seen in Figure 4.8 and Figure 4.9. Since ADMEAN learns a model on a fixed window, the only drift we are interested in is the drift contained by the window, represented by the solid line. The window size also has a maximum equal to the size of the fixed window. A solid line represents the suggested window size according to the drift occurrences.

When the drift is less frequent and of less magnitude (Figure 4.8), ADMEAN appears to often suggest a window size that is smaller than the true window size. There is at least one drift, at instance 480, that is almost completely missed. However, there are some very good

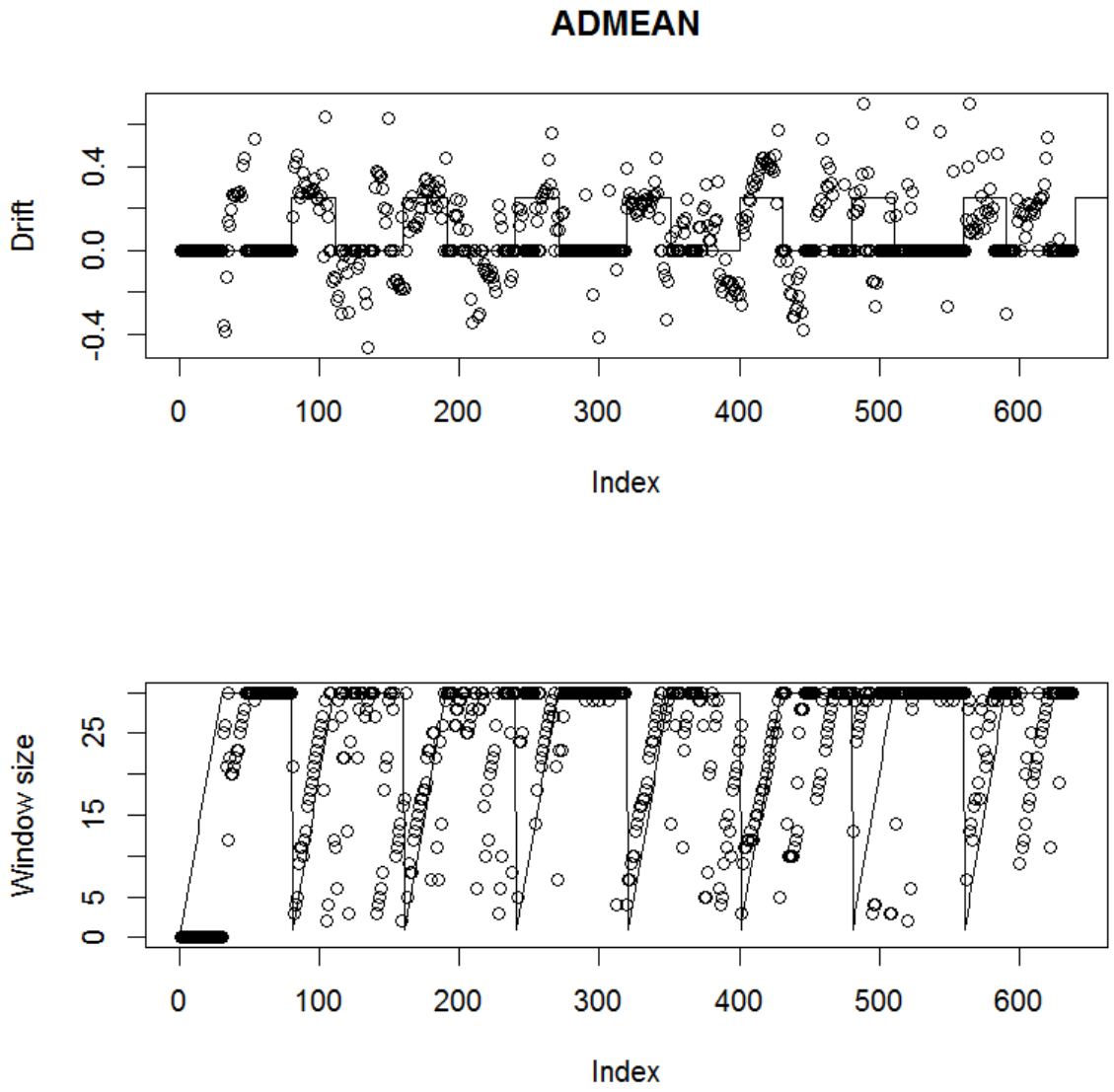


Figure 4.8: Drift delay = 80, drift magnitude = 1.0

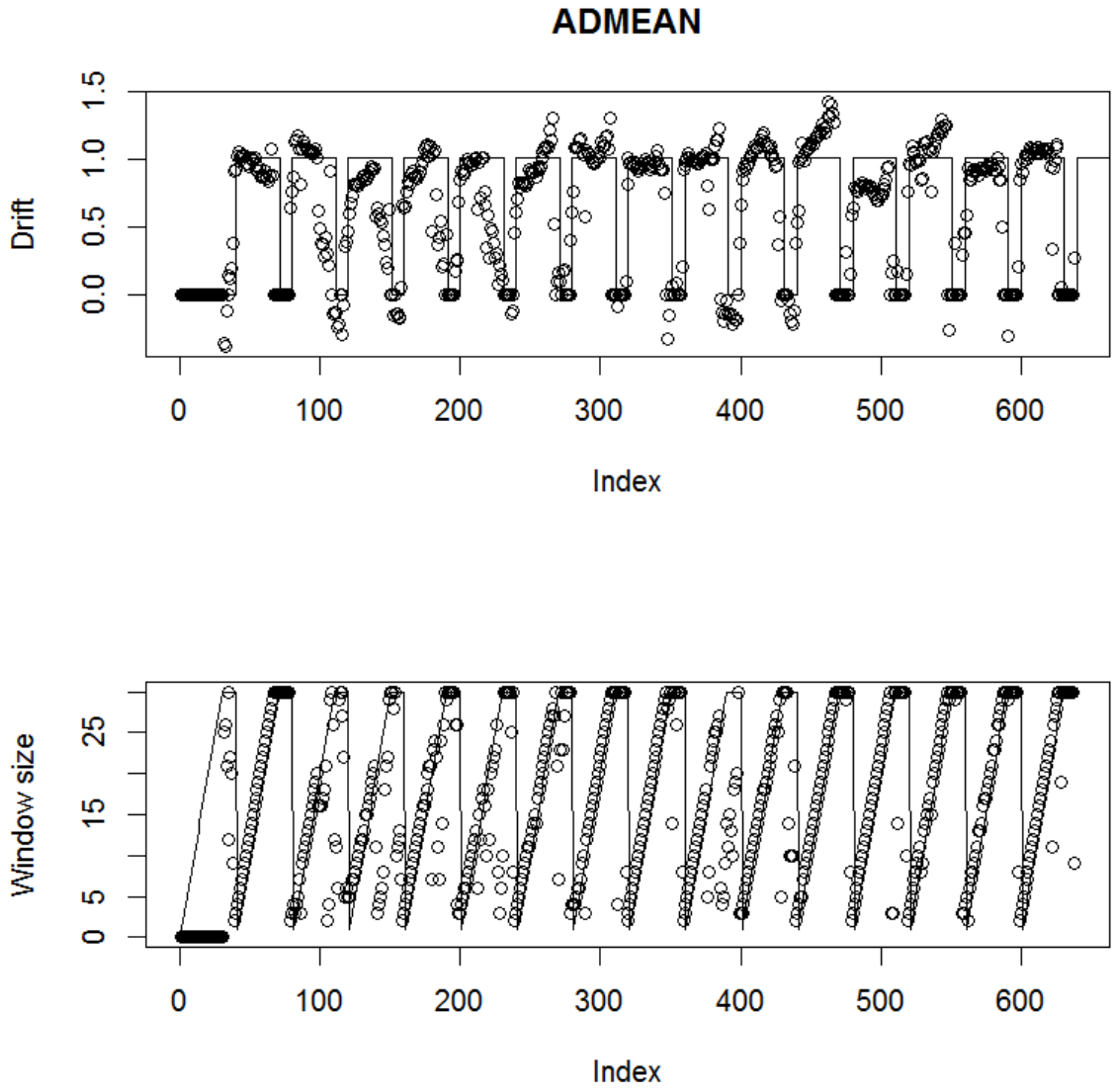


Figure 4.9: Drift delay = 40, drift magnitude = 4.0

Method	Sea	Gauss	Iris	Olive	Wine	Wine2
ADMEAN	9.7	1.6	4.5	4.8	4.1	27.9
CD-EWMA	9.1	1.1	3.1	3.1	3.8	30.0
WR*	9.5	7.5	5.9	6.9	12.4	32.4
WR	9.1	7.8	6.2	7.9	12.8	32.1
ADWIN	9.6	19.5	9.0	26.2	40.7	48.3

Table 4.1: Percent error of methods for classification of 2-class data containing concept drift

predictions between instances 240 and 430.

When the drift is of greater magnitude and frequency (Figure 4.9), there is more adherence to the true window size. The predictions of the drift appear to be closer to their true value as well. However, when an incorrect window size is suggested, the prediction error is large.

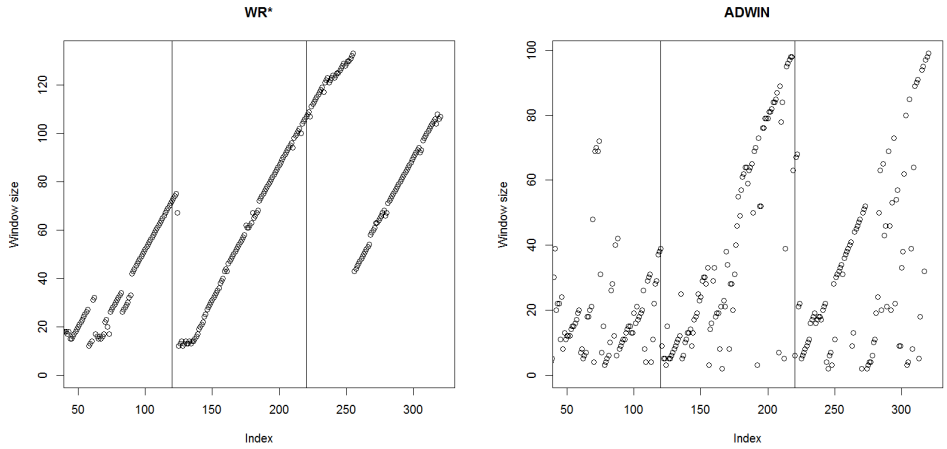
4.3 Concept drift problems

Table 4.1 gives the percent error for the concept drift experiments. Overall, ADMEAN and CD-EWMA show a stable improvement over the existing methods ADWIN and WR*. ADWIN appears to be the worst performer on the data tested.

The most improved predictions for ADMEAN and CE-EWMA are on the wine data with synthetic concept drift. The high dimensionality of the wine data makes learning a new model very difficult on small amounts of data. This gives ADMEAN and CD-EWMA the advantage, because they retain past data.

The window size for WR* seems to grow with consistency. Applying the ADWIN method to the attributes gives less consistent suggestions for window size. Recall that ADWIN observes each attribute as a separate stream. The smallest suggested window is adapted as the overall window size suggestion. This makes ADWIN more sensitive to false predictions. Figure 4.10 shows an example of this difference in window size selection for the olive oil data. The vertical lines indicate the location of the simulated concept drifts.

From Figure 4.10, WR* appears to experience a delayed reaction to the concept drift. The delay is a negative effect of the same low sensitivity that causes WR* to suggest a consistently



(a) (b)

Figure 4.10: Window size plotted against index for the WR* (a) and ADWIN (b) window sizing algorithms

growing window size. The high sensitivity of ADWIN causes the algorithm to experience a high error rate.

The ADMEAN method uses the same change point detection as ADWIN. However, an overall window is not suggested. Instead, the attribute values of the data occurring after the detected change point are adjusted according to the values within a fixed window occurring before the detected change point. Retaining the previous data counters allows stability despite the increased sensitivity. This allows ADMEAN quick reaction to change and less repercussions for false detections.

As an example, attribute three of the iris dataset is shown in Figure 4.11 with the window size and correction used in ADMEAN. The difference between this data and the synthetic measurement drift data is the existence of two classes, where only one of those classes experiences a drift. There appears to be a class balance problem between instances 200 and 220. This shrinks the window and is the most prominent source of error seen in this data. For this problem, the instances were equally probable to be either class. So, the imbalance is an anomaly. However, it is probable to assume there are changes in class balance in a real application.

The CD-EWMA predicted drift for attribute three of the iris data is shown in Figure 4.12. While the drift appears to be tracked by CD-EWMA, the variance in the prediction is large.

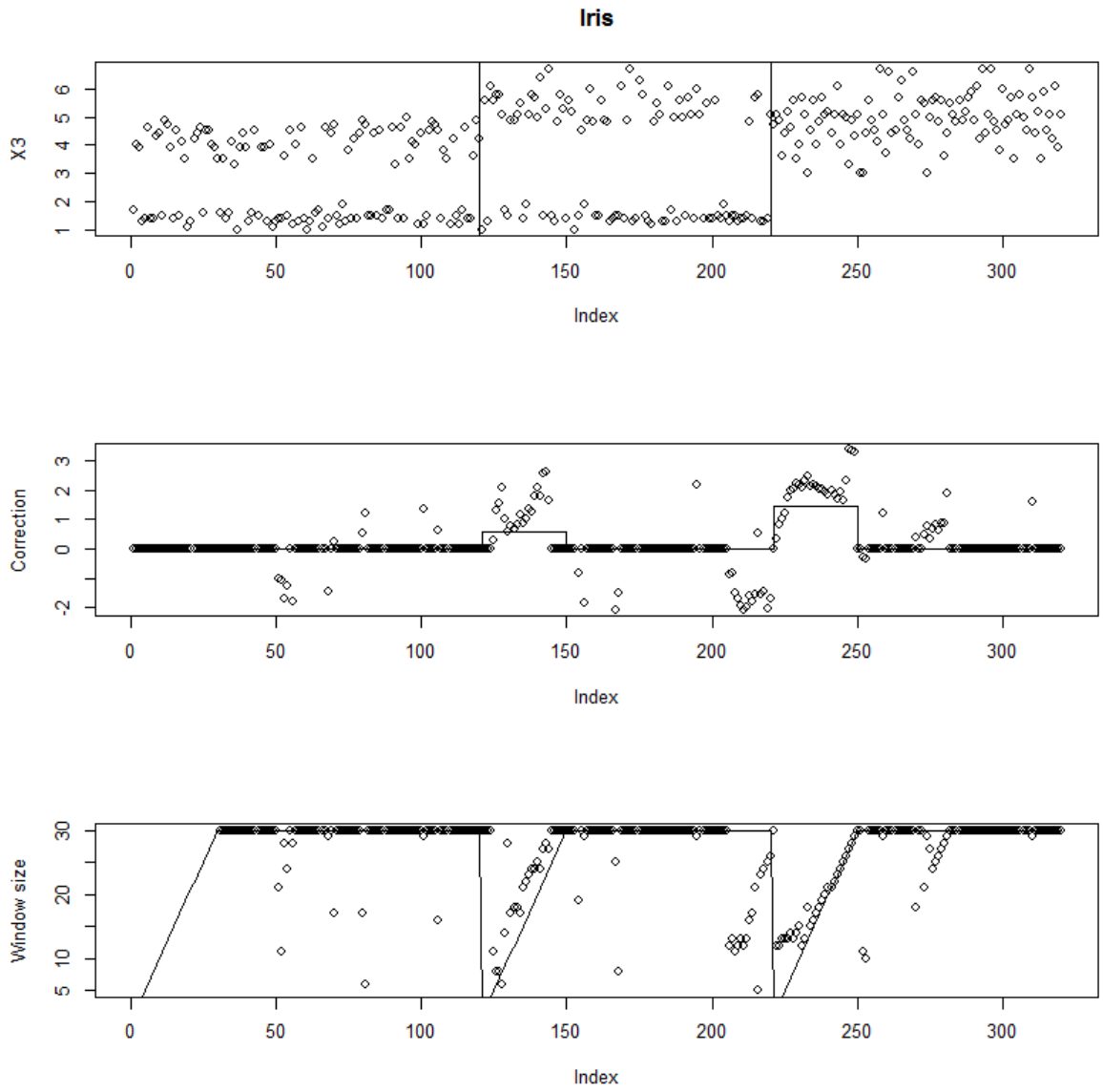


Figure 4.11: ADMEAN predicted drift and window size on attribute three of iris with synthetic concept drift

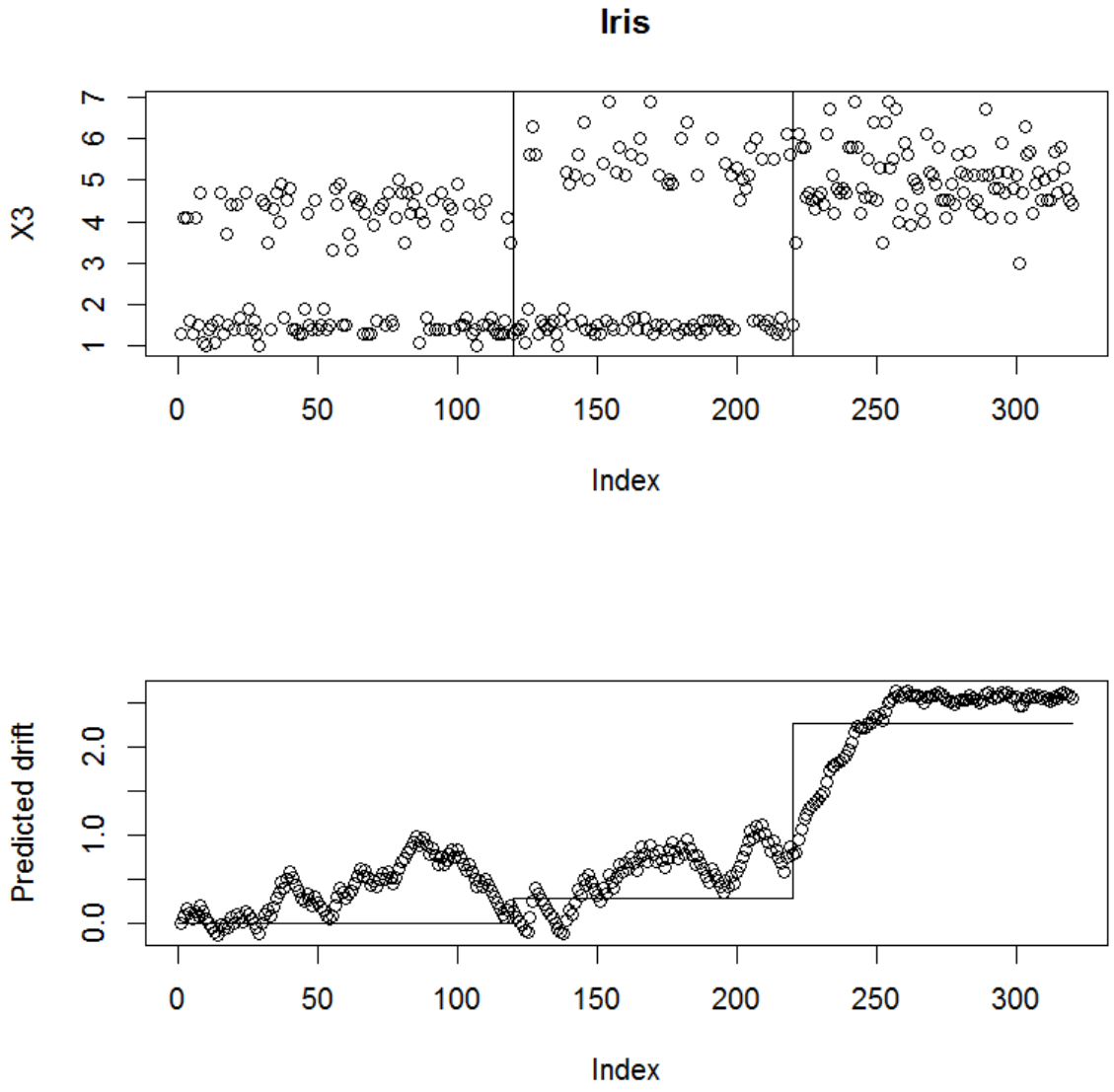


Figure 4.12: CD-EWMA predicted drift on attribute three of iris with synthetic concept drift

There are a few peaks, where the predicted value rises before falling back to the true drift value. The response to the large drift in Class 1 at instance 220 is apparent. However, the predicted drift is a little larger than the true drift.

CHAPTER 5. SUMMARY AND DISCUSSION

5.1 Conclusions

The advanced process control problem has been shown to be a specific case of the concept drift problem. Concept drift methods that were adapted to fit the problem of measurement drift in on-line process control were competitive. Adapting the measurement drift on-line process control algorithms for classification in the presence of concept drift proved more fruitful.

5.1.1 Measurement drift

The identification of measurement drift as the source for a change in model accuracy from projections of the data was given on page 19. The mathematical model representing this type of drift was described in section 3.1.1. These discoveries and review of the concept drift and advanced process control literature led to simple modifications of existing algorithms and the formation of two new algorithms. These methods were tested on three unique datasets provided by the industry partner as well as 16 synthetic data sets.

The results on the industry data indicated that the modified concept drift algorithms performed as well as the popular advanced process control algorithm EWMA-I. Overall, no algorithm was superior to the others. A simple fixed sliding window of small size (5-10) consistently performed comparably to the algorithm with least error in each of the three datasets. Its simplicity only aids the argument for its implementation in any situation similar to the industrial study in this thesis.

The results from the synthetic measurement drift data supported the results of the study data. Further, testing on data containing drifts that are large compared to the variance of the outcome variable revealed the strength of the proposed algorithm, ADMEAN, on this type of

data. A small fixed window size method remained competitive with other methods across the synthetic measurement drift datasets.

5.1.2 Concept drift

The proposed methods, concept drift exponential weighted moving average (CD-EWMA) and adaptive mean (ADMEAN) showed improvement on the generated concept drift datasets. While the existing methods performed well, the window resize method (WR) did not adapt to the new concept as quickly as the proposed algorithms. The sensitivity of the adaptive window (ADWIN) algorithm applied to each attribute as a separate data stream incurred too many false detections. This sensitivity benefited the ADMEAN algorithm, which used the same drift detection method as ADWIN. However, ADMEAN's retention of past data, adjusting the data occurring after the drift, allowed it to tolerate false detections and maintain a quick reaction to drift. CD-EWMA relied on its ability to adjust incoming data before classification to make a correct decision.

5.2 Future work: Process improvement

As mentioned in the introduction, this work has been developed to connect the data mining and process control fields. The hope is that model adaptations and ensembles can be implemented throughout a system, to continuously improve the products or services being produced and provided. The on-line model adaptation for part size selection, studied in this thesis, was promising because of the data collection already taking place. A clear measure of the attributes as well as the response were available.

Some of the answers that data mining research has provided could be helpful in on-line continuous process improvement. When assessing the performance of a system, solutions that favor outcomes on goods or services provided most frequently may be assessed as more appropriate, even if they cause many errors on the goods or services provided or produced with less frequency. Approaches to handle accuracy in the presence of underrepresented classes would allow a better assessment of the solution.

Sometimes there is a lot of information provided by the categorical attributes of a product. Data mining algorithms are often developed with application to categorical data in mind. Within service, it is important to understand the customer. The customer's demographics provide key categorical information.

There may be a chance for missing values in the data stream. There may also be long delays before learning of a response variable. Testing of a part or feedback from a provided service are may be costly or hard to obtain. Dealing with missing data has been an active area within data mining research.

5.3 Future work: Concept drift

CD-EWMA showed improvement on some of the concept drift datasets tested. However, there are methods in EWMA controls for selecting an optimal weight. Adapting one of these methods to the CD-EWMA algorithm may allow the CD-EWMA even better results.

In these studies, ADWIN selects the window of with the smallest size. If weights were assigned to the suggested window sizes from each attribute, perhaps ADWIN would be less sensitive to false drift detection.

The fixed window size of ADMEAN was set to thirty in the experiments of this work. Determining the fixed window size through off-line analysis of the number of variables and the data characteristics may improve the performance of the method. A dynamic window size, adjusted according to incoming and observed data, could also improve the method.

5.4 Concluding remarks

From this research, it was clear that solutions had been developed for a very specific problem. The solutions to that problem were unique and insightful. Identifying other areas where data mining can be applied can certainly be helpful if those applications lead to better results. However, that is not the only outcome. Identifying these areas can expand data mining research by bringing the approaches to a more general level. These generalized algorithms may prove to be a useful tool in the data miner's tool belt.

APPENDIX A. VISUALIZED SYNTHETIC DATA RESULTS

A.1 Absolute mean error plots

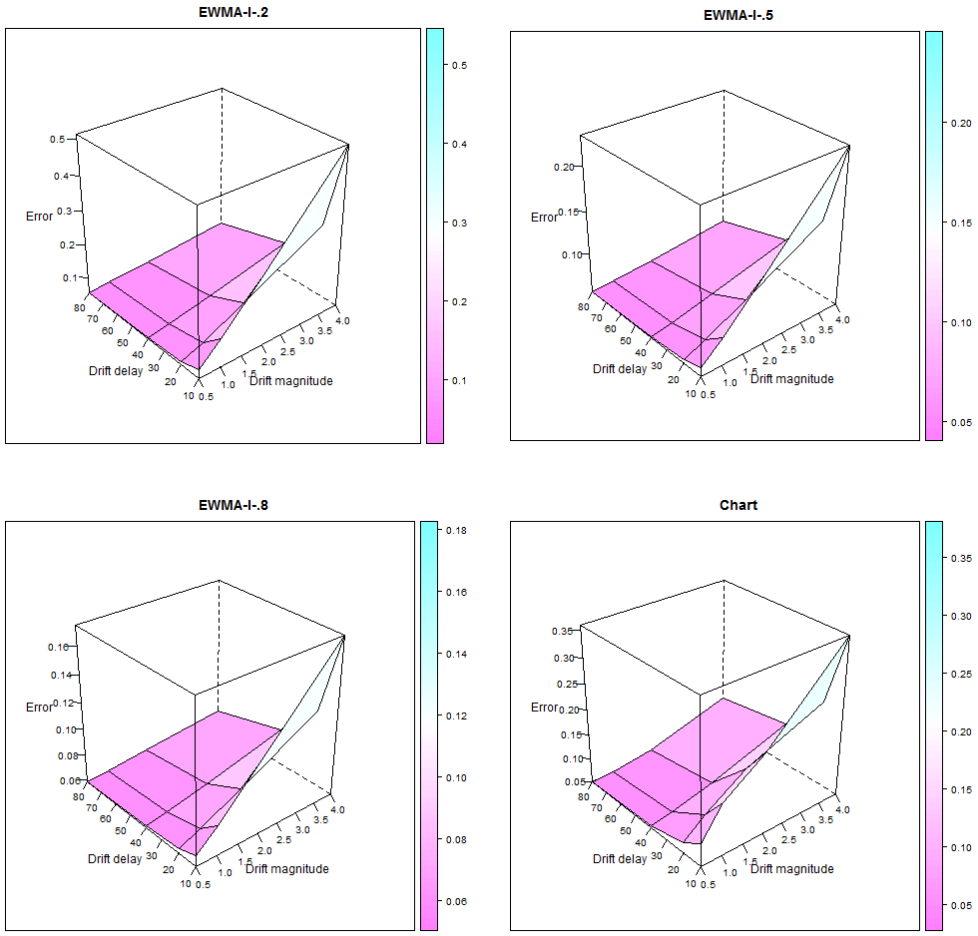


Figure A.1: Process control inspired algorithms applied to simulated measurement drift data

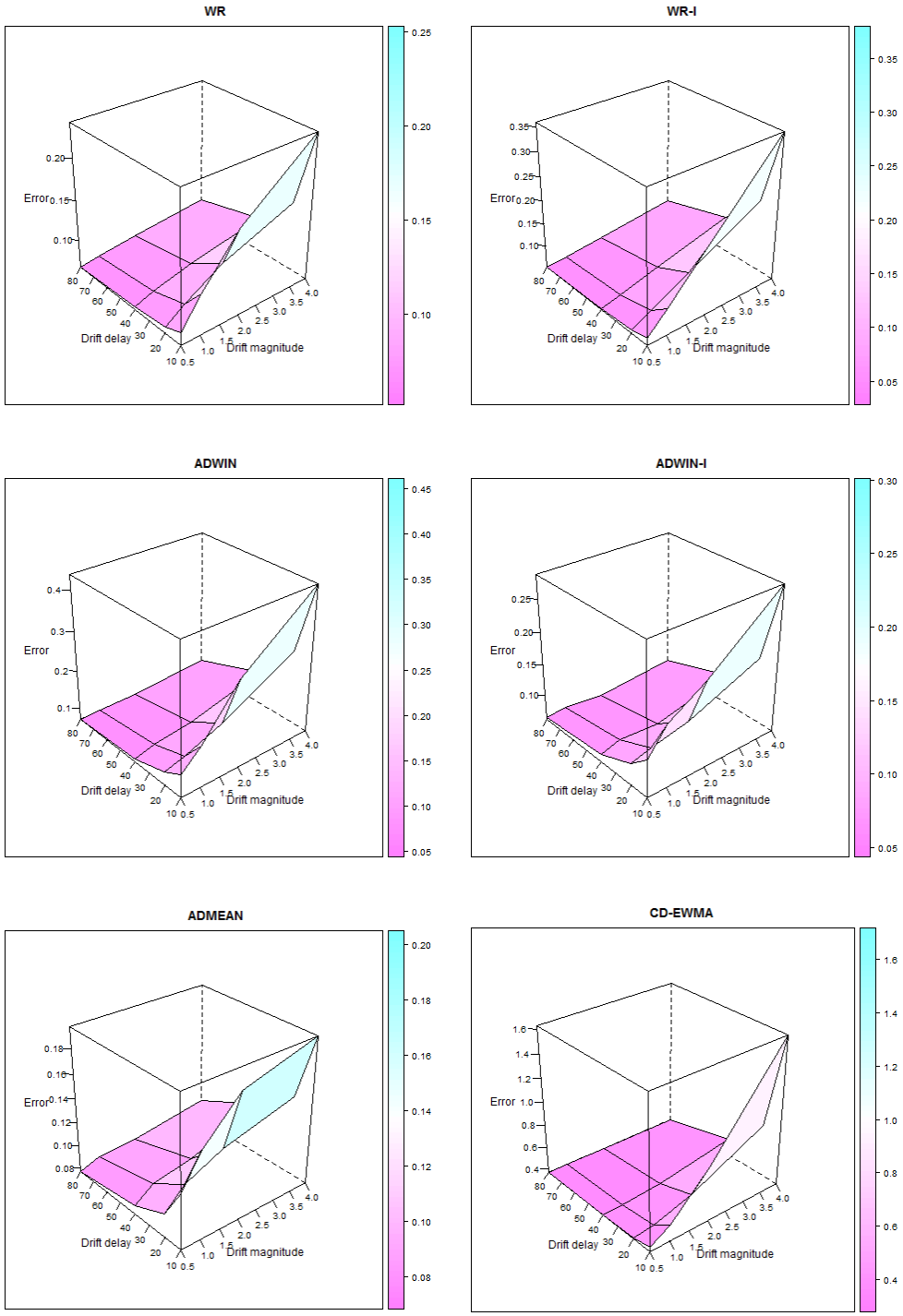


Figure A.2: Modified concept drift algorithms applied to simulated measurement drift data

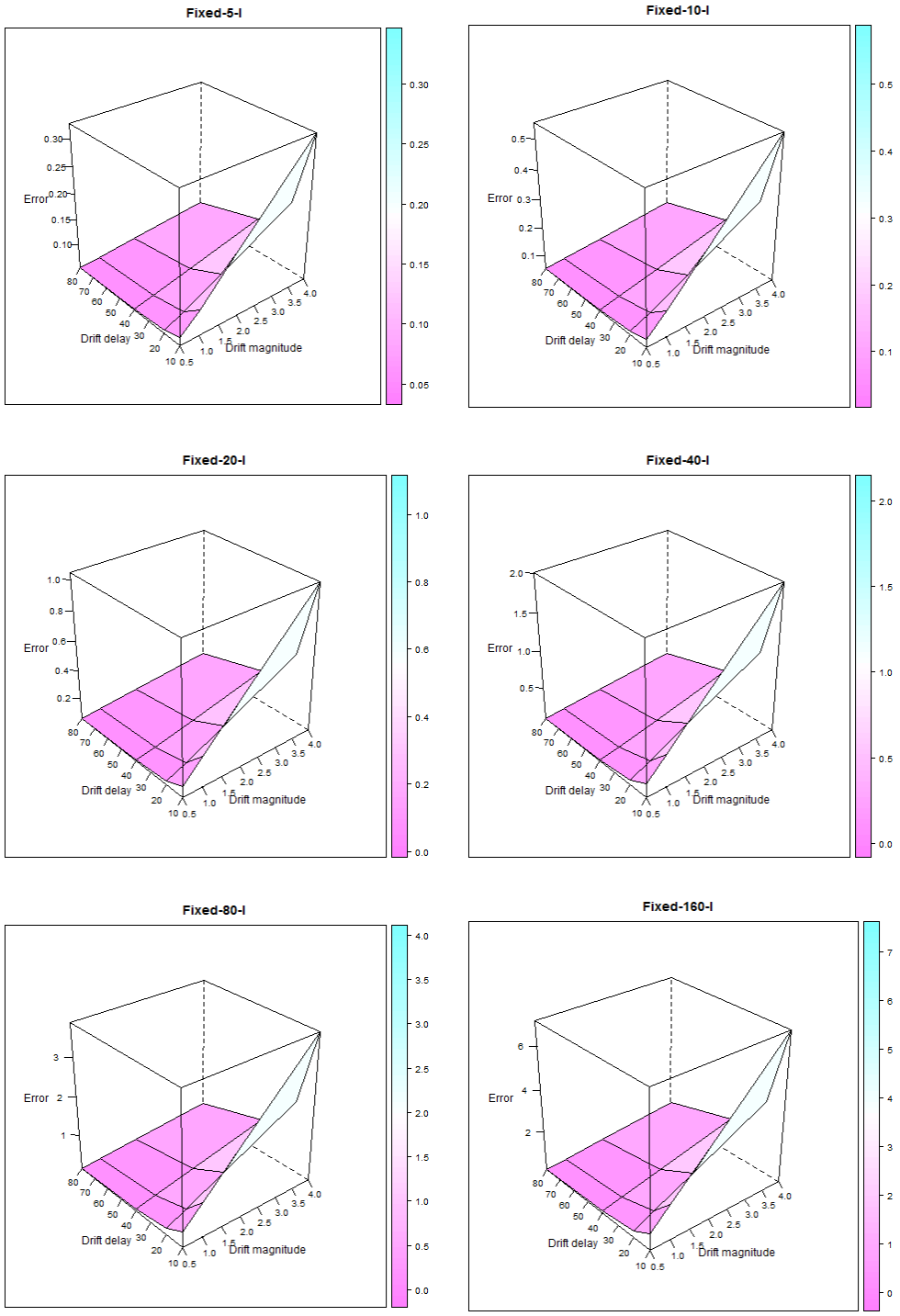


Figure A.3: Fixed window algorithms applied to simulated measurement drift data

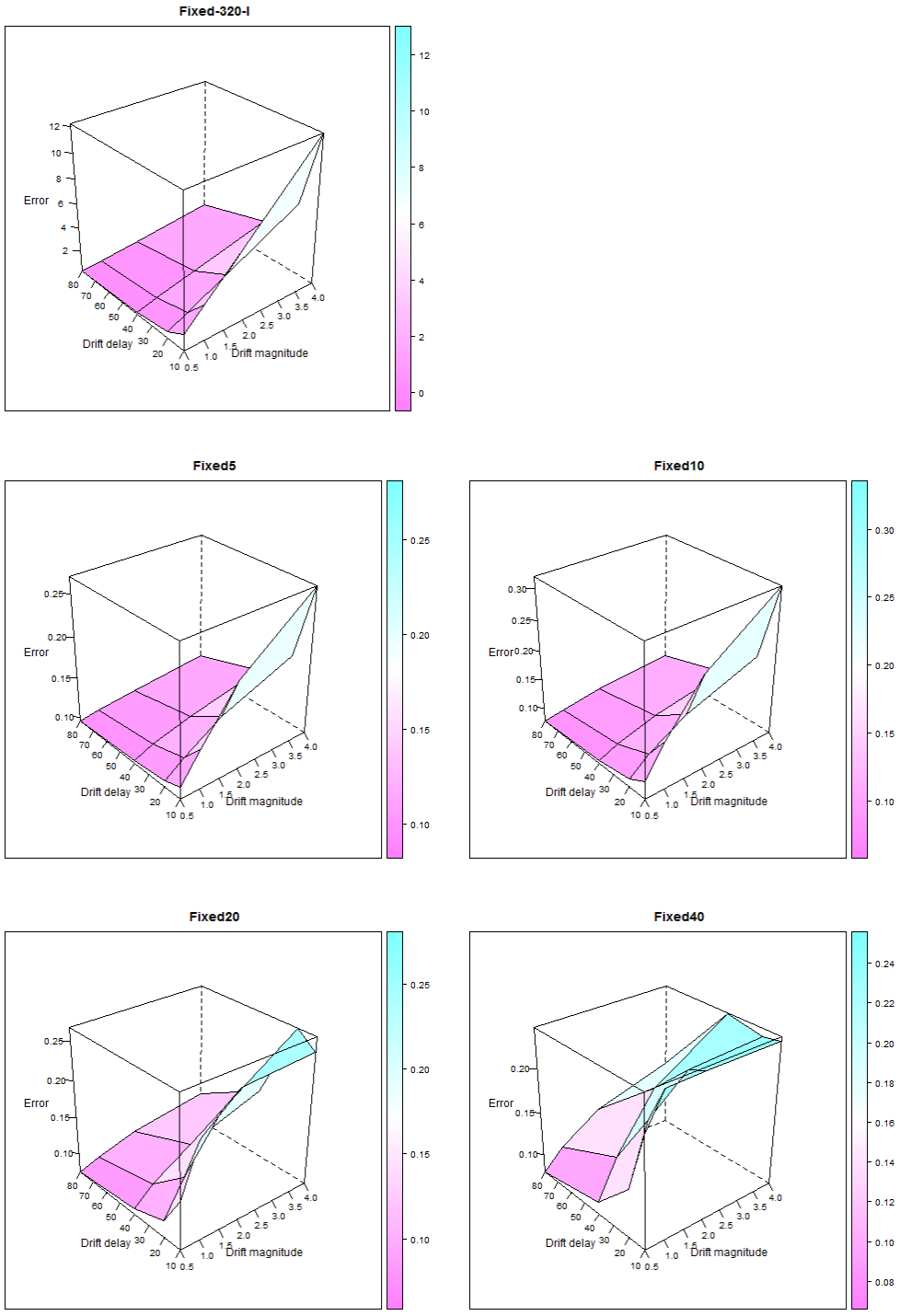


Figure A.4: Fixed window algorithms applied to simulated measurement drift data

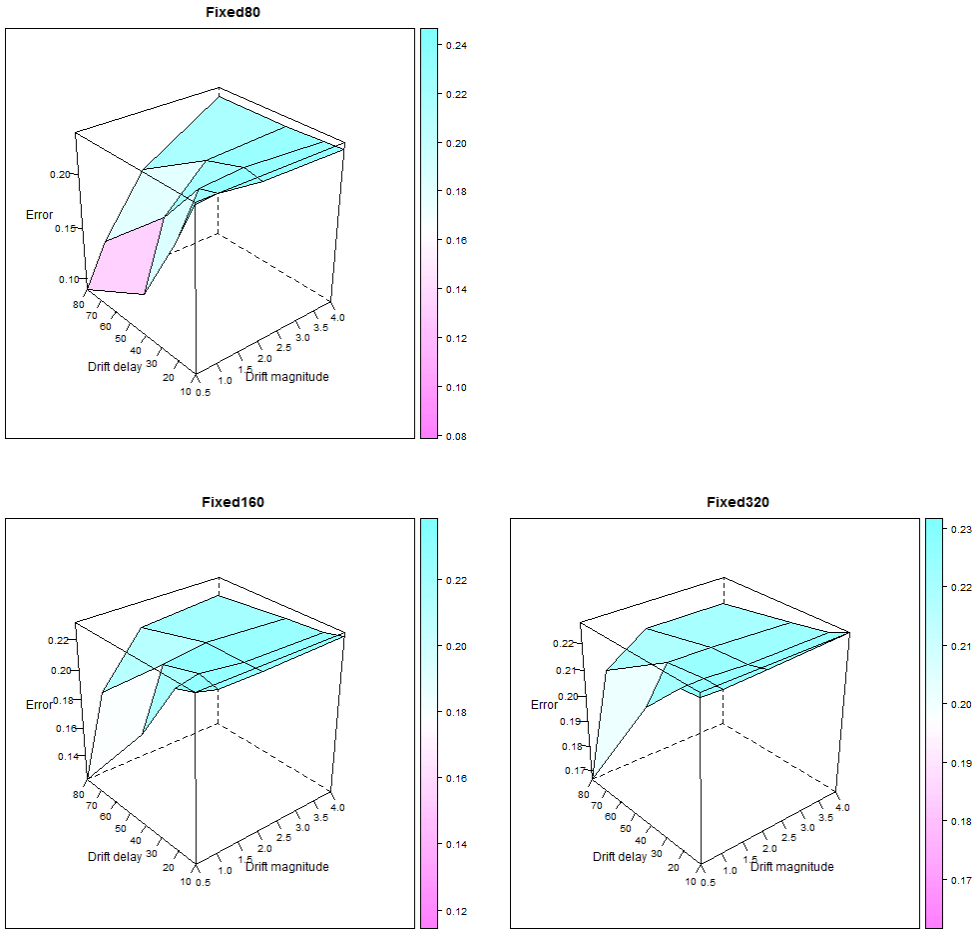


Figure A.5: Fixed window algorithms applied to simulated measurement drift data

A.2 Mean error and standard deviation plots

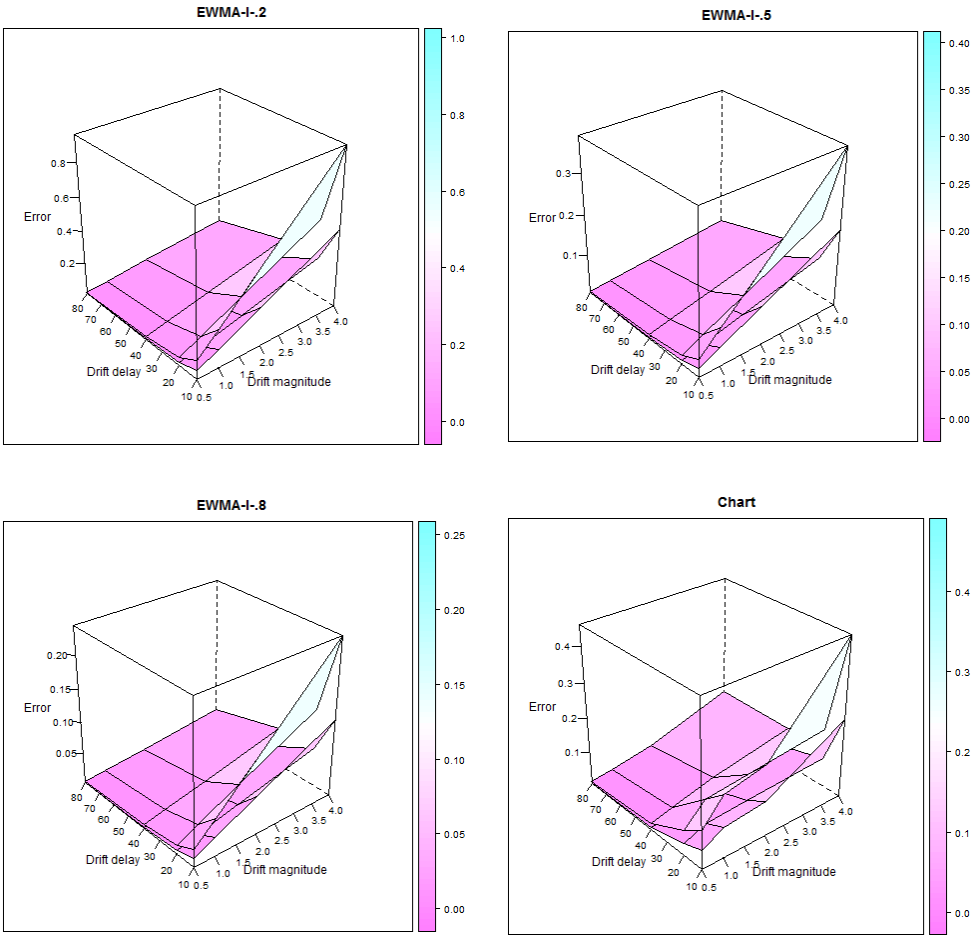


Figure A.6: Process control inspired algorithms applied to simulated measurement drift data with standard deviation

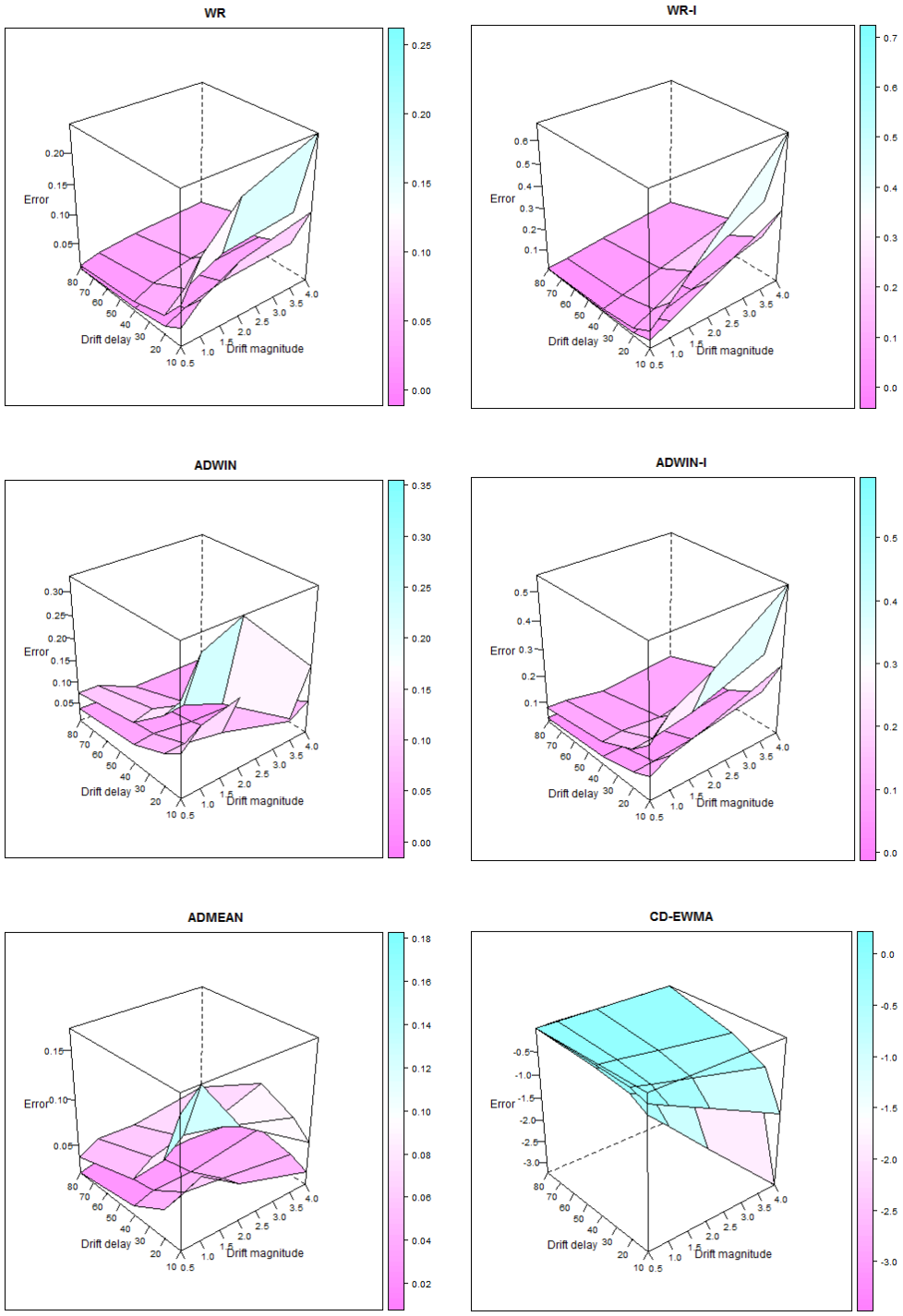


Figure A.7: Modified concept drift algorithms applied to simulated measurement drift data with standard deviation

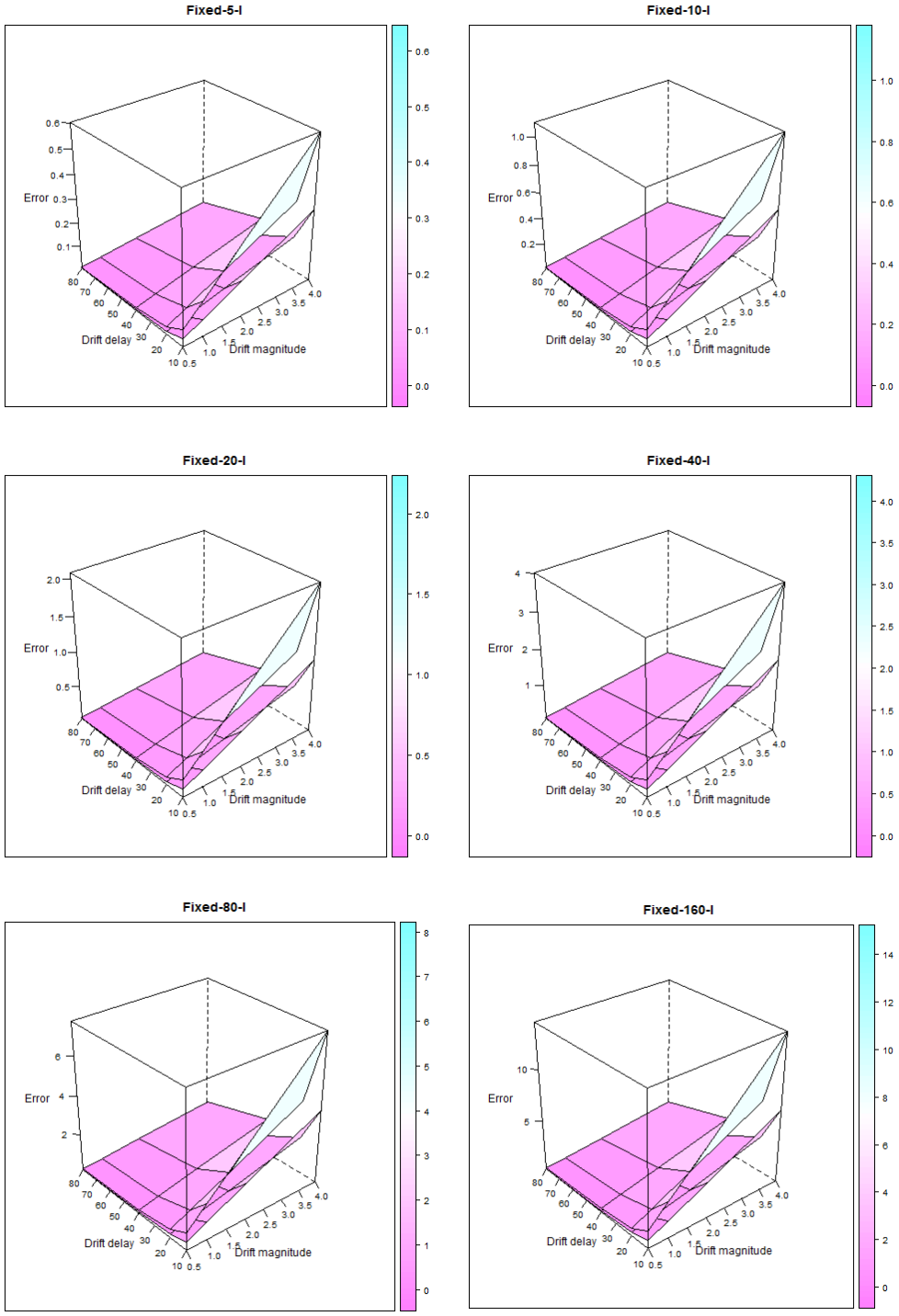


Figure A.8: Fixed window algorithms applied to simulated measurement drift data with standard deviation

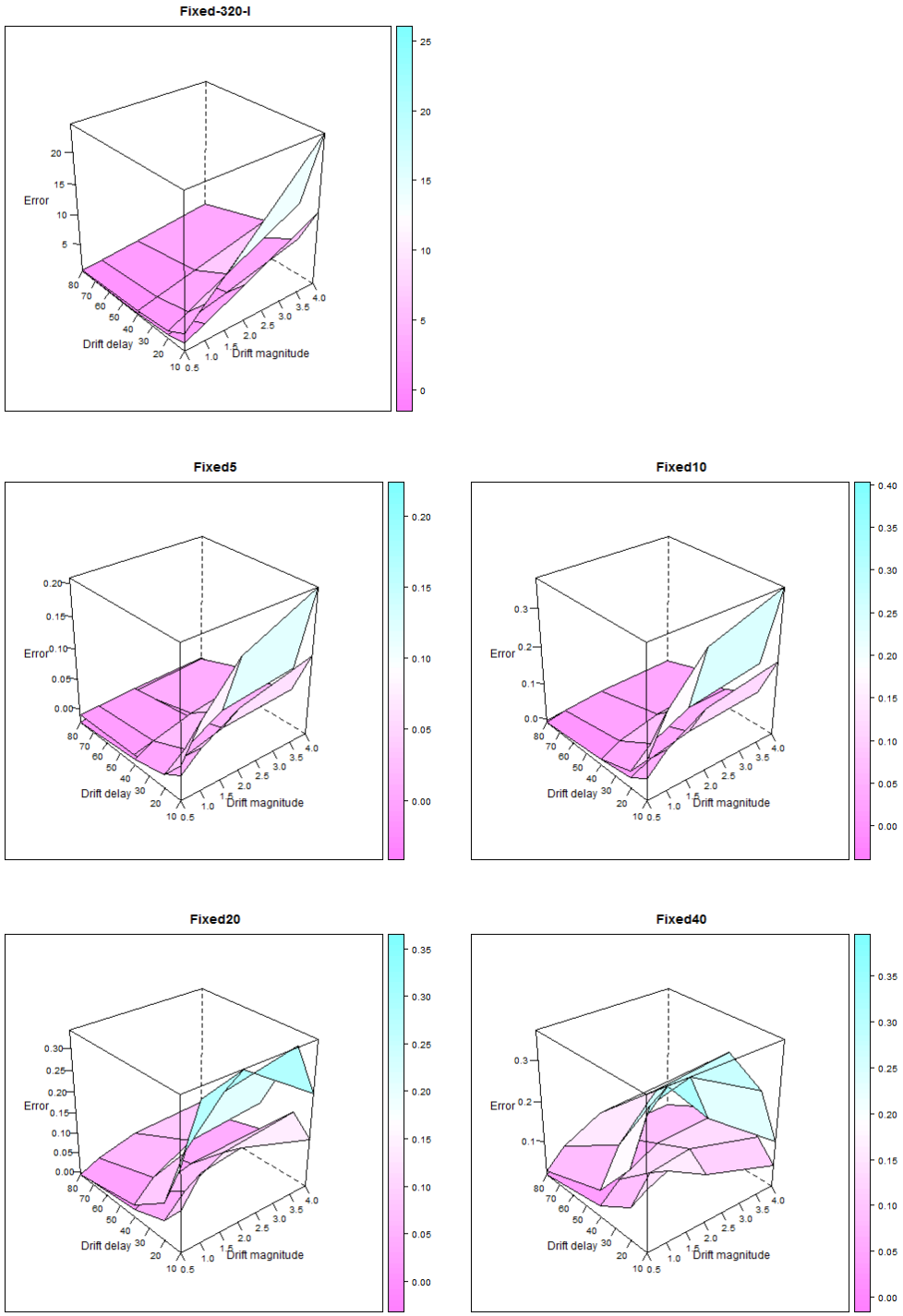


Figure A.9: Fixed window algorithms applied to simulated measurement drift data with standard deviation

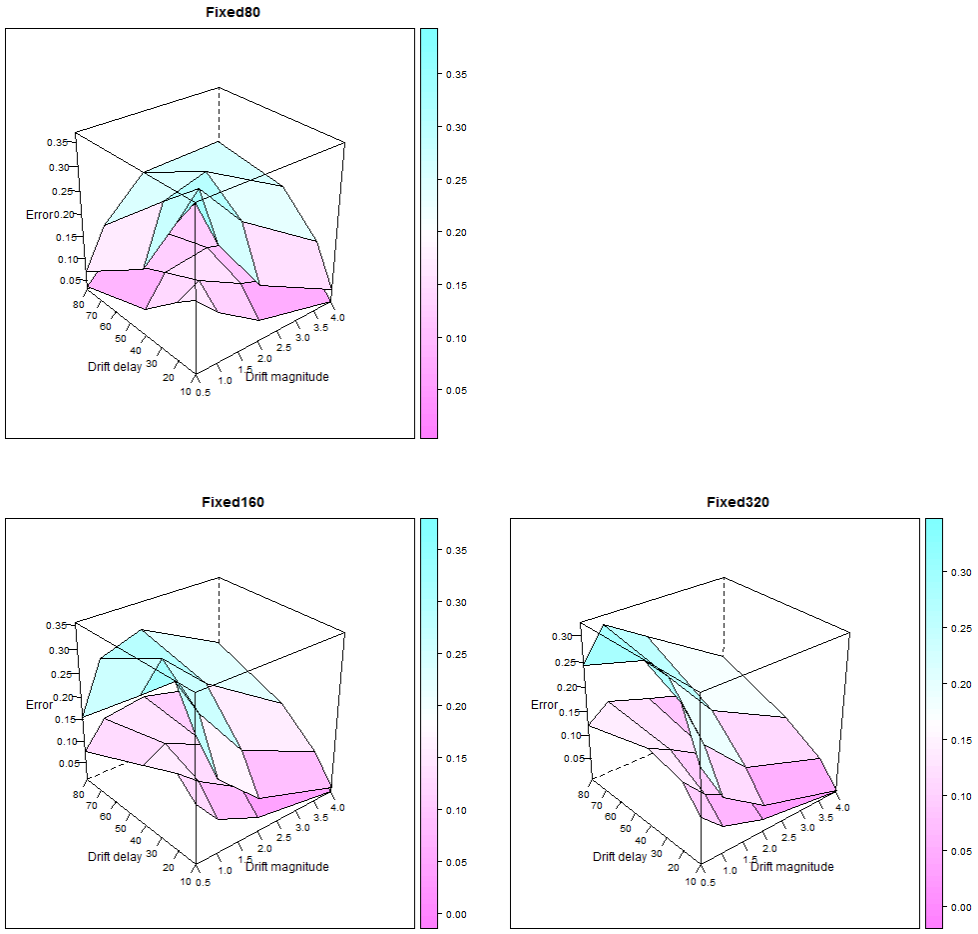


Figure A.10: Fixed window algorithms applied to simulated measurement drift data with standard deviation

APPENDIX B. NUMERIC SYNTHETIC DATA RESULTS

B.1 Simulation results drift delay 10-20: absolute mean error

Drift to variance	0.50	1.00	2.00	4.00	0.50	1.00	2.00	4.00
Drift delay	10.00	10.00	10.00	10.00	20.00	20.00	20.00	20.00
EWMA.I..2	0.08	0.13	0.26	0.51	0.06	0.09	0.14	0.26
EWMA.I..5	0.06	0.08	0.13	0.23	0.06	0.07	0.09	0.14
EWMA.I..8	0.07	0.08	0.11	0.17	0.06	0.07	0.08	0.11
Chart	0.10	0.15	0.21	0.36	0.07	0.11	0.16	0.22
ADWIN.I	0.12	0.16	0.19	0.29	0.09	0.10	0.12	0.16
ADWIN	0.13	0.17	0.29	0.44	0.10	0.12	0.16	0.25
WR	0.08	0.11	0.17	0.24	0.07	0.09	0.11	0.15
WR.I	0.07	0.11	0.19	0.36	0.06	0.08	0.12	0.20
ADMEAN	0.12	0.15	0.18	0.20	0.10	0.11	0.13	0.14
ADMEAN.I	0.12	0.15	0.19	0.26	0.09	0.11	0.13	0.16
CD.EWMA	0.41	0.51	0.84	1.63	0.38	0.41	0.50	0.81
Fixed5	0.11	0.14	0.20	0.27	0.10	0.12	0.15	0.18
Fixed10	0.10	0.16	0.23	0.32	0.09	0.11	0.15	0.19
Fixed20	0.14	0.20	0.24	0.25	0.10	0.14	0.20	0.27
Fixed40	0.20	0.24	0.24	0.24	0.13	0.20	0.23	0.24
Fixed80	0.23	0.24	0.23	0.23	0.19	0.23	0.23	0.23
Fixed160	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23
Fixed320	0.23	0.23	0.23	0.23	0.23	0.23	0.22	0.22
Fixed.5.I	0.07	0.10	0.18	0.33	0.06	0.08	0.11	0.19

Drift to variance	0.50	1.00	2.00	4.00	0.50	1.00	2.00	4.00
Drift delay	10.00	10.00	10.00	10.00	20.00	20.00	20.00	20.00
Fixed.10.I	0.08	0.14	0.28	0.55	0.07	0.10	0.16	0.30
Fixed.20.I	0.13	0.26	0.52	1.05	0.08	0.14	0.26	0.52
Fixed.40.I	0.25	0.50	1.00	2.01	0.13	0.25	0.50	1.00
Fixed.80.I	0.48	0.96	1.92	3.84	0.24	0.48	0.96	1.92
Fixed.160.I	0.89	1.78	3.56	7.13	0.45	0.89	1.78	3.56
Fixed.320.I	1.52	3.04	6.09	12.17	0.76	1.52	3.04	6.08

B.2 Simulation results drift delay 40 - 80: absolute mean error

Drift to variance	0.50	1.00	2.00	4.00	0.50	1.00	2.00	4.00
Drift delay	40.00	40.00	40.00	40.00	80.00	80.00	80.00	80.00
EWMA.I..2	0.05	0.07	0.09	0.15	0.05	0.06	0.07	0.10
EWMA.I..5	0.05	0.06	0.07	0.09	0.05	0.06	0.06	0.07
EWMA.I..8	0.06	0.06	0.07	0.09	0.06	0.06	0.06	0.07
Chart	0.05	0.06	0.08	0.13	0.05	0.05	0.06	0.11
ADWIN.I	0.07	0.08	0.08	0.11	0.06	0.07	0.06	0.07
ADWIN	0.08	0.09	0.10	0.15	0.07	0.07	0.07	0.09
WR	0.07	0.08	0.09	0.11	0.06	0.07	0.08	0.09
WR.I	0.05	0.06	0.08	0.13	0.05	0.05	0.06	0.08
ADMEAN	0.08	0.10	0.10	0.12	0.08	0.08	0.09	0.10
ADMEAN.I	0.08	0.09	0.09	0.11	0.07	0.08	0.08	0.09
CD.EWMA	0.37	0.38	0.41	0.51	0.37	0.37	0.37	0.37
Fixed5	0.10	0.11	0.12	0.14	0.09	0.10	0.10	0.11
Fixed10	0.08	0.09	0.11	0.13	0.08	0.08	0.09	0.10
Fixed20	0.08	0.10	0.13	0.16	0.07	0.08	0.10	0.11
Fixed40	0.09	0.13	0.19	0.24	0.08	0.10	0.13	0.15
Fixed80	0.13	0.19	0.22	0.23	0.09	0.13	0.18	0.23

Drift to variance	0.50	1.00	2.00	4.00	0.50	1.00	2.00	4.00
Drift delay	40.00	40.00	40.00	40.00	80.00	80.00	80.00	80.00
Fixed160	0.19	0.22	0.23	0.22	0.12	0.18	0.21	0.22
Fixed320	0.21	0.22	0.22	0.22	0.17	0.21	0.22	0.22
Fixed.5.I	0.05	0.06	0.08	0.12	0.05	0.06	0.07	0.08
Fixed.10.I	0.06	0.07	0.10	0.17	0.05	0.06	0.07	0.11
Fixed.20.I	0.06	0.09	0.15	0.28	0.05	0.07	0.10	0.15
Fixed.40.I	0.07	0.13	0.25	0.49	0.06	0.08	0.14	0.25
Fixed.80.I	0.12	0.24	0.48	0.95	0.07	0.12	0.23	0.46
Fixed.160.I	0.22	0.44	0.88	1.77	0.11	0.22	0.43	0.86
Fixed.320.I	0.38	0.76	1.51	3.03	0.19	0.37	0.74	1.49

B.3 Simulation results drift delay 10 - 20: error standard deviation

Drift to variance	0.50	1.00	2.00	4.00	0.50	1.00	2.00	4.00
Drift delay	10.00	10.00	10.00	10.00	20.00	20.00	20.00	20.00
EWMA.I..2.sd	0.08	0.11	0.20	0.39	0.07	0.10	0.16	0.31
EWMA.I..5.sd	0.08	0.11	0.18	0.35	0.07	0.09	0.14	0.26
EWMA.I..8.sd	0.08	0.11	0.18	0.34	0.08	0.09	0.14	0.24
Chart.sd	0.11	0.17	0.25	0.44	0.08	0.12	0.19	0.27
ADWIN.I.sd	0.09	0.12	0.18	0.31	0.08	0.11	0.15	0.25
ADWIN.sd	0.15	0.22	1.46	1.71	0.14	0.17	0.28	0.93
WR.sd	0.10	0.14	0.22	0.37	0.09	0.12	0.18	0.27
WR.I.sd	0.07	0.11	0.19	0.37	0.07	0.10	0.16	0.31
ADMEAN.sd	0.14	0.17	0.22	0.25	0.12	0.14	0.17	0.20
ADMEAN.I.sd	0.13	0.15	0.20	0.30	0.11	0.13	0.17	0.24
CD.EWMA.sd	0.47	0.47	0.48	0.50	0.47	0.47	0.47	0.48
Fixed5.sd	0.47	0.49	0.54	0.62	0.47	0.48	0.51	0.57
Fixed10.sd	0.47	0.47	0.51	0.59	0.46	0.47	0.49	0.54

Drift to variance	0.50	1.00	2.00	4.00	0.50	1.00	2.00	4.00
Drift delay	10.00	10.00	10.00	10.00	20.00	20.00	20.00	20.00
Fixed20.sd	0.47	0.48	0.50	0.53	0.46	0.47	0.50	0.55
Fixed40.sd	0.47	0.49	0.52	0.53	0.47	0.48	0.50	0.52
Fixed80.sd	0.49	0.51	0.52	0.53	0.47	0.49	0.51	0.52
Fixed160.sd	0.51	0.52	0.52	0.53	0.49	0.51	0.52	0.52
Fixed320.sd	0.51	0.52	0.52	0.53	0.50	0.51	0.52	0.52
Fixed.5.I.sd	0.08	0.11	0.19	0.37	0.07	0.10	0.16	0.30
Fixed.10.I.sd	0.07	0.09	0.16	0.30	0.07	0.11	0.18	0.35
Fixed.20.I.sd	0.07	0.10	0.17	0.32	0.07	0.10	0.16	0.31
Fixed.40.I.sd	0.08	0.12	0.22	0.42	0.07	0.10	0.18	0.34
Fixed.80.I.sd	0.12	0.22	0.43	0.85	0.09	0.14	0.26	0.51
Fixed.160.I.sd	0.28	0.54	1.07	2.14	0.15	0.29	0.56	1.12
Fixed.320.I.sd	0.66	1.32	2.63	5.26	0.34	0.67	1.33	2.65

B.4 Simulation results drift frequency 40 - 80: error standard deviation

Drift to variance	0.50	1.00	2.00	4.00	0.50	1.00	2.00	4.00
Drift frequency	40.00	40.00	40.00	40.00	80.00	80.00	80.00	80.00
EWMA.I..2.sd	0.07	0.08	0.13	0.24	0.06	0.07	0.10	0.18
EWMA.I..5.sd	0.07	0.08	0.11	0.19	0.07	0.07	0.09	0.14
EWMA.I..8.sd	0.07	0.08	0.11	0.18	0.07	0.08	0.09	0.13
Chart.sd	0.07	0.09	0.13	0.21	0.06	0.07	0.10	0.19
ADWIN.I.sd	0.07	0.10	0.12	0.20	0.07	0.09	0.09	0.14
ADWIN.sd	0.11	0.14	0.21	0.50	0.10	0.12	0.14	0.20
WR.sd	0.09	0.11	0.14	0.20	0.09	0.09	0.11	0.16
WR.I.sd	0.07	0.08	0.13	0.24	0.06	0.07	0.10	0.18
ADMEAN.sd	0.11	0.12	0.14	0.19	0.10	0.11	0.12	0.15
ADMEAN.I.sd	0.10	0.12	0.14	0.19	0.10	0.11	0.12	0.15

Drift to variance	0.50	1.00	2.00	4.00	0.50	1.00	2.00	4.00
Drift frequency	40.00	40.00	40.00	40.00	80.00	80.00	80.00	80.00
CD.EWMA.sd	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47
Fixed5.sd	0.47	0.48	0.50	0.53	0.47	0.47	0.47	0.49
Fixed10.sd	0.46	0.47	0.48	0.50	0.46	0.46	0.47	0.48
Fixed20.sd	0.46	0.47	0.48	0.51	0.46	0.46	0.47	0.49
Fixed40.sd	0.46	0.47	0.49	0.53	0.46	0.47	0.48	0.50
Fixed80.sd	0.47	0.48	0.50	0.52	0.46	0.47	0.49	0.53
Fixed160.sd	0.47	0.49	0.51	0.52	0.47	0.48	0.50	0.51
Fixed320.sd	0.48	0.50	0.52	0.52	0.47	0.49	0.51	0.52
Fixed.5.I.sd	0.07	0.08	0.13	0.23	0.07	0.07	0.10	0.17
Fixed.10.I.sd	0.07	0.09	0.15	0.28	0.07	0.08	0.12	0.21
Fixed.20.I.sd	0.07	0.10	0.18	0.34	0.07	0.09	0.14	0.26
Fixed.40.I.sd	0.07	0.10	0.16	0.31	0.07	0.10	0.17	0.33
Fixed.80.I.sd	0.07	0.11	0.20	0.38	0.07	0.10	0.17	0.33
Fixed.160.I.sd	0.10	0.17	0.32	0.63	0.08	0.12	0.22	0.43
Fixed.320.I.sd	0.18	0.35	0.69	1.37	0.11	0.20	0.38	0.75

BIBLIOGRAPHY

Baena-García, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavaldà, R., and Morales-Bueno, R. (2006). Early drift detection method.

Baxley, R. (1990). Discussion. *Technometrics*, 32:1–29.

Baxley, R. (1994). Applications of the ewma for algorithmic statistical process control. *Quality Engineering*, 7(2):397–418.

Becker, H. and Arias, M. (2007). Real-time ranking with concept drift using expert advice. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 86–94. ACM.

Bifet, A. and Gavalda, R. (2007). Learning from time-changing data with adaptive windowing. In *SIAM International Conference on Data Mining*, pages 443–448.

Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., and Gavaldà, R. (2009). New ensemble methods for evolving data streams. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 139–148. ACM.

Chandrasekaran, S., English, J., and Disney, R. (1995). Modeling and analysis of ewma control schemes with variance-adjusted control limits. *IIE transactions*, 27(3):282–290.

Crowder, S. V. (1987). A simple method for studying run–length distributions of exponentially weighted moving average charts. *Technometrics*, 29(4):401–407.

Del Castillo, E. (2001). Some properties of ewma feedback quality adjustment schemes for drifting disturbances. *Journal of Quality Technology*, 33(2).

Faltin, F. and Tuckers, W. (1991). On-line quality control for the factory of the 1990's and beyond.

Gama, J., Medas, P., Castillo, G., and Rodrigues, P. (2004). Learning with drift detection. In *Advances in Artificial Intelligence–SBIA 2004*, pages 286–295. Springer.

Gan, F. (1993). An optimal design of ewma control charts based on median run length. *Journal of Statistical Computation and Simulation*, 45(3-4):169–184.

Hand, D. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14.

Hunter, J. S. (1986). The exponentially weighted moving average. *J. QUALITY TECHNOL.*, 18(4):203–210.

Lee, H. (2008). *Advanced Process Control and Optimal Sampling in Semiconductor Manufacturing*. University of Texas at Austin.

Lucas, J. M. and Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, 32(1):1–12.

MacGregor, J. (1987). Interfaces between process control and on-line statistical process control. *Comp. Syst. Technol. Div. Comm.*, 10:9–20.

Nishida, K. and Yamauchi, K. (2009). Learning, detecting, understanding, and predicting concept changes. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pages 2280–2287. IEEE.

Page, E. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Roberts, S. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250.

Sachs, E., Hu, A., and Ingolfsson, A. (1995). Run by run process control: combining spc and feedback control. *Semiconductor Manufacturing, IEEE Transactions on*, 8(1):26–43.

Schlimmer, J. and Granger, R. (1986). Incremental learning from noisy data. *Machine learning*, 1(3):317–354.

Shewhart, W. (1931). *Economic Control of Quality of Manufactured Product*. D. Van Nostrand, New York.

Shewhart, W. (1939). *Statistical Method from the Viewpoint of Quality Control*. The Graduate School, Department of Agriculture, Washington, D.C.

Tucker, W., Faltin, F., and Vander Weil, S. (1992). Algorithmic statistical process control: An elaboration. *Statistical Research Report: AT & T Bell Laboratories*, 102.

Vander Weil, S., Tucker, W., and Faltin, F. (1992). Algorithmic statistical process control"concepts and an application. *Technometrics*, 34:286–287.

Wang, J. and He, Q. (2008). Ewma run-to-run controllers with gain updating: stability and sensitivity analysis. In *American Control Conference, 2008*, pages 2872–2877. IEEE.

Wang, J., Peter He, Q., and Joe Qin, S. (2010). Stability analysis and optimal tuning of ewma controllers–gain adaptation vs. intercept adaptation. *Journal of Process Control*, 20(2):134–142.

Widmer, G. and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. In *Machine Learning*, pages 69–101.

Zhang, Y. and Jin, X. (2006). An automatic construction and organization strategy for ensemble learning on data streams. *ACM SIGMOD Record*, 35(3):28–33.

Žliobaitė, I. (2009). Combining time and space similarity for small size learning under concept drift. *Foundations of Intelligent Systems*, pages 412–421.

Zliobaite, I. (2009). Learning under concept drift: an overview. Technical report, Overview, Technical report, Vilnius University, 2009 techniques, related areas, applications Subjects: Artificial Intelligence.

Žliobaitė, I. (2011). Combining similarity in time and space for training set formation under concept drift. *Intelligent Data Analysis*, 15(4):589–611.

Zliobaite, I. and Kuncheva, L. (2009). Determining the training window for small sample size classification with concept drift. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 447–452. IEEE.