ABSTRACT

The complexity of supply chains (SC) has grown rapidly in recent years, resulting in an increased difficulty to evaluate and visualize performance. Consequently, analytical approaches to evaluate SC performance in near real time relative to targets and plans are important to detect and react to deviations in order to prevent major disruptions.

Manufacturing anomalies, inaccurate forecasts, and other problems can lead to SC disruptions. Traditional monitoring methods are not sufficient in this respect, because complex SCs feature changes in manufacturing tasks (dynamic complexity) and carry a large number of stock keeping units (detail complexity). Problems are easily confounded with normal system variations.

Motivated by these real challenges faced by modern SC, new surveillance solutions are proposed to detect system deviations that could lead to disruptions in a complex SC. To address supply-side deviations, the fitness of different statistics that can be extracted from the enterprise resource planning system is evaluated. A monitoring strategy is first proposed for SCs featuring high levels of dynamic complexity. This presents an opportunity for monitoring methods to be applied in a new, rich domain of SC management. Then a monitoring strategy, called Heat Map Contrasts (HMC), which converts monitoring into a series of classification problems, is used to monitor SCs with both high levels of dynamic and detail complexities. Data from a semiconductor SC simulator are used to compare the methods with other alternatives under various failure cases, and the results illustrate the viability of our methods.

To address demand-side deviations, a new method of quantifying forecast uncertainties using the progression of forecast updates is presented. It is illustrated that a rich amount of information is available in rolling horizon forecasts. Two proactive indicators of
future forecast errors are extracted from the forecast stream. This quantitative method requires no knowledge of the forecasting model itself and has shown promising results when applied to two datasets consisting of real forecast updates.
Dedicated to my wife Yan and my family
ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. George Runger for his continuous support and guidance of my study and research at Arizona State University, and for his patience, motivation, enthusiasm, and immense knowledge. The guidance, encouragement, and academic freedom that he has given me help me all the time of research and writing of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I would like to thank my dissertation committee members, Dr. Esma Gel, Dr. Mani Janakiram, and Dr. Rong Pan for their insightful advices and comments. I have been honored to co-author with, David Bayba, Phillip Howard. Their unselfish help, insights, and feedbacks helped me improve my knowledge in the area. I am also grateful to my colleagues and friends: Amit Shinde, Houtao Deng, Bahareh Azarnoush, Anshuman Sahu, Lingjun Li, Bi Wu, Zhuoyang Zhou, Ning Wang, Ming Zhang and Austin Brent for their friendship and care, which helped me survive through these years.

Finally and most importantly, none of my achievement would have been possible without the love and patience of my wife and my family. My most sincere thanks go to my wife Yan Wu for her endless support and love during the past six years. I am very thankful to my parents, Peirong Liu and Yulan Xu, for their dedication and unconditional support throughout my life.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TABLE OF CONTENTS</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td></td>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Monitor Supply-side Deviations with High Dynamic Complexity</td>
<td>10</td>
</tr>
<tr>
<td>2.1</td>
<td>Background</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Monitor SC with Dynamic Complexity</td>
<td>15</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Control Strategy</td>
<td>16</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Control Chart Design</td>
<td>21</td>
</tr>
<tr>
<td>2.3</td>
<td>Experiments</td>
<td>25</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Initial Analysis</td>
<td>27</td>
</tr>
<tr>
<td>2.3.2</td>
<td>False Alarms</td>
<td>29</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Signal Detection</td>
<td>30</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Estimation of Control Limits</td>
<td>31</td>
</tr>
<tr>
<td>2.4</td>
<td>Conclusions</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>Monitor Production Execution in SC with High Dynamic and Detail Complexity</td>
<td>34</td>
</tr>
<tr>
<td>3.1</td>
<td>Background</td>
<td>37</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Complex Supply Chain</td>
<td>37</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Ship Rates</td>
<td>38</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Visualization of Multivariate Time Series</td>
<td>39</td>
</tr>
<tr>
<td>3.1.4</td>
<td>Multivariate Control</td>
<td>40</td>
</tr>
<tr>
<td>3.2</td>
<td>Control System</td>
<td>43</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Monitoring Statistic</td>
<td>43</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Heat Map Contrast</td>
<td>46</td>
</tr>
</tbody>
</table>
CHAPTER 3.2.3 Sensitize Monitor to Critical Failure .......................... 51
3.2.4 Deploy HMC to entire SC and Result Presentation .................. 53
3.3 Experiments ................................................................. 54
3.3.1 Simulation Model ....................................................... 54
3.3.2 Computation of $s_k(t), \tilde{s}_k(t)$ and $z_k(t)$ ....................... 57
3.3.3 SC Data Visualization ............................................... 58
3.3.4 Monitoring Underproductions ..................................... 59
  3.3.4.1 Comparison of Different Control Strategies ................. 59
  3.3.4.2 Detecting Underproduction Scenarios with HMC ........... 67
3.4 Conclusion ................................................................. 69
4 Proactive Monitor of Demand-side Deviations ............................ 71
  4.1 Background .............................................................. 73
    4.1.1 Estimation of Forecast Uncertainty ........................... 74
    4.1.2 Multivariate Control ............................................ 75
  4.2 Method ................................................................. 76
    4.2.1 Forecast Grid .................................................... 76
    4.2.2 Predictive Monitor of Forecast Uncertainty .................. 78
    4.2.3 Statistics ....................................................... 81
  4.3 Empirical Results ...................................................... 83
    4.3.1 Nominal GDP Forecast ........................................... 83
    4.3.2 Commodity Price Forecasts ................................... 85
  4.4 Enhancement .......................................................... 92
    4.4.1 Monitor Vision Change in Multiple SKUs with Heat Map Contrast 92
    4.4.2 Feature Extraction via Deep Learning ......................... 93
    4.4.3 False Alarm Pruning ............................................ 94
<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5 Conclusion</td>
<td>95</td>
</tr>
<tr>
<td>5 Conclusions and Future Work</td>
<td>96</td>
</tr>
<tr>
<td>5.1 Future Work</td>
<td>99</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>101</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>$ARL_0$ of EWMA-WSD and EWMA. $\lambda = 0.1$ and $L = 2.81$ Are Used.</td>
<td>29</td>
</tr>
<tr>
<td>2.2</td>
<td>$ARL_0$ of CUSUM and CUSUM-WSD Charts with $h = 0.5\hat{\sigma}_0$ and $H = 5\hat{\sigma}_0$.</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>$ARL_0$ in CUSUM and CUSUM-WSD Are Close to Expected.</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>$ARL_1$ and $SE_1$ for Four Control Charts. Column $ND$ Indicates the Number of Replicates Chart Fails to Detect Change within 800 Hours.</td>
<td>31</td>
</tr>
<tr>
<td>3.1</td>
<td>Equipment and Tool Information in the Simulation.</td>
<td>55</td>
</tr>
<tr>
<td>3.2</td>
<td>Baseline Ship Rate $b_k(t)$ for SKU1-SKU96 at $A_1$ in the Simulation Model.</td>
<td>56</td>
</tr>
<tr>
<td>3.3</td>
<td>Zone Score Conversion. If SKU $k$ Belongs to Group $g$, Converting $\tilde{s}_k(t)$ to $z_k(t)$ Using the Zone Widths Pair ${\tau^+_g, \tau^-_g}$.</td>
<td>58</td>
</tr>
<tr>
<td>3.4</td>
<td>Different Aggregations for SKU1-SKU96 at $A_1$.</td>
<td>62</td>
</tr>
<tr>
<td>3.5</td>
<td>Five Failure Cases Are Tested in Experiment 3.3.4.2. Each Failure Case Causes Different Number of SKUs to Deviate with Different Magnitudes.</td>
<td>68</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Overview of a Complex SC. A Large Number of SKUs Are Processed, Stored and Moved in the System.</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Inventory Building Curves of SKU S1-S16. Each SKU Builds Its Own Inventory to Anticipate a Delivery. Red Arrow: 15% Under Production in S14 From Time 350.</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Production of SKU k in a Manufacturing Site.</td>
<td>15</td>
</tr>
<tr>
<td>2.2 Shipment Quantity Plots of One SKU from a Semiconductor SC Simulator.</td>
<td>17</td>
</tr>
<tr>
<td>2.3 Illustration of Converting $s_k(u)$ to $r_k(t)$.</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Baseline Inventory and Actual Inventory of SKU k at a Manufacturing Site.</td>
<td>19</td>
</tr>
<tr>
<td>2.5 Build Plan Changes at $t_1, t_2$ and $b_k(t)$ Changes Accordingly. When Failures occur at $t$, $r_k(t)$ Deviates from $b_k(t)$.</td>
<td>20</td>
</tr>
<tr>
<td>2.6 Grouping SKUs Based on Their Parental SKUs or Shared Equipments.</td>
<td>25</td>
</tr>
<tr>
<td>2.7 Simulation Model of a Multi-echelon, Multi-product SC.</td>
<td>26</td>
</tr>
<tr>
<td>2.8 Different Skewness in $r_k(t)$ from Different $b_k$ Values. Low $b_k$ Values Lead to More Skewed and Variated $r_k(t)$. Window Size $w = 4$.</td>
<td>27</td>
</tr>
<tr>
<td>2.9 Due to Small Batch Production, Inventory Levels and Inventory Deviations Exhibit Strong Autocorrelation. Autocorrelations in Ship Rates Are Weaker.</td>
<td>28</td>
</tr>
<tr>
<td>2.10 Estimations of Control Limits for EWMA-WSD for $r_0(t)$ from Similar Baseline Build Plans with $w = \gamma = 4$ hours and $L = 2.81$.</td>
<td>32</td>
</tr>
<tr>
<td>3.1 Visualization of Multivariate Time Series.</td>
<td>39</td>
</tr>
<tr>
<td>3.2 Zone Score for Symmetrical Distributions.</td>
<td>44</td>
</tr>
<tr>
<td>3.3 Zone Score from Skewed Distributions. Left: Percentile Zone. Right: Simple Zone.</td>
<td>45</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>3.4</td>
<td>48</td>
</tr>
<tr>
<td>3.5</td>
<td>50</td>
</tr>
<tr>
<td>3.6</td>
<td>51</td>
</tr>
<tr>
<td>3.7</td>
<td>53</td>
</tr>
<tr>
<td>3.8</td>
<td>55</td>
</tr>
<tr>
<td>3.9</td>
<td>59</td>
</tr>
<tr>
<td>3.10</td>
<td>60</td>
</tr>
<tr>
<td>3.11</td>
<td>64</td>
</tr>
<tr>
<td>3.12</td>
<td>64</td>
</tr>
<tr>
<td>3.13</td>
<td>65</td>
</tr>
<tr>
<td>3.14</td>
<td>65</td>
</tr>
<tr>
<td>3.15</td>
<td>66</td>
</tr>
<tr>
<td>3.16</td>
<td>66</td>
</tr>
<tr>
<td>3.17</td>
<td>67</td>
</tr>
<tr>
<td>3.18</td>
<td>68</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>4.1 Rolling Horizon Forecast with Prediction Horizon Equals to Three.</td>
<td>73</td>
</tr>
<tr>
<td>4.2 Forecasting Grid $G_t$: Historical Forecasts and Realized Demand Information at $t$.</td>
<td>77</td>
</tr>
<tr>
<td>4.3 One Forecaster with Rolling Horizon and Survey of Professional Forecasters.</td>
<td>79</td>
</tr>
<tr>
<td>4.4 Four Different Changes Are Illustrated. (a) (b) Illustrate Not-well-apprehended Demand Changes; (c) Illustrates No Change in Demand, (d) Illustrates a Well-apprehended Demand Change.</td>
<td>80</td>
</tr>
<tr>
<td>4.5 Different Ways to Summarize $V_t$.</td>
<td>82</td>
</tr>
<tr>
<td>4.6 Training Data of US Nominal GDP for Random Forest.</td>
<td>84</td>
</tr>
<tr>
<td>4.7 Variable Importance from Random Forest Analysis.</td>
<td>85</td>
</tr>
<tr>
<td>4.8 Forecast Uncertainty Estimation from Three Different Statistics. Top: $\tilde{m}<em>{t-1}$, Middle: $\tilde{a}</em>{t-1}$, and Bottom: $MAE_{t-1}$.</td>
<td>86</td>
</tr>
<tr>
<td>4.9 Raw Data of Commodity Price Forecasts from the World Bank.</td>
<td>87</td>
</tr>
<tr>
<td>4.10 Plot of Actual Price (Black), One-period-ahead Forecasts (Blue), and $\hat{q}_t$ (Brown).</td>
<td>88</td>
</tr>
<tr>
<td>4.11 Scatter Plots of $e'_t$ and $\hat{q}'_t$. Points Are Less Vertically Spread Out When $</td>
<td>\hat{q}'_t</td>
</tr>
<tr>
<td>4.12 Plot of $</td>
<td>e'_t</td>
</tr>
<tr>
<td>4.13 Illustration of Monitoring $M_t$ with HMC.</td>
<td>93</td>
</tr>
</tbody>
</table>
CHAPTER 1

Introduction

This chapter provides an overview of our study and a summary of our contributions. Modern supply chains (SC) are complex adaptive systems that produce, store, and move stock keeping units (SKU) between geographically separated facilities [12]. Driven by product proliferation, rapid technological development, and complications in product design, the complexities of SCs have grown rapidly. Bozarth [6] divided these complexities into level of detail and dynamic components. Detail complexities are driven by the number of distinct finished products and their unique supporting parts; whereas dynamic complexities come from the instabilities in manufacturing tasks driven by the variabilities associated with the demand and manufacturing processes.

Figure 1.1 illustrates a SC that features high levels of dynamic and detail complexities. It offers greater variety in its products to appeal to more customers and satisfies the heterogeneity in their needs. Productions of each finished product involve many supporting parts and require the cooperation of multiple manufacturing sites. Demand forecasts for each individual product are fed into a planning system, and the planning system generates build and delivery plans to coordinate material movements among different sites. Such planning systems are capable of dealing with the complex interactions required to link productions with demands. They consider various factors such as material and equipment availability, bill of materials (BOM), production lead times, and demand variabilities.

Within each manufacturing site, a variety of SKUs are processed with production involving different sets of machines. Products often are made in small batches in order to accommodate low volumes of production [6]. As illustrated in Figure 1.2, each SKU has its inventory built to anticipate a delivery that consumes much of the available stock. The
Figure 1.1: Overview of a Complex SC. A Large Number of SKUs Are Processed, Stored and Moved in the System.

recurrent cycles of inventory build followed by deliveries repeat, but at infrequent, non-periodic intervals. Driven by changing delivery sizes, the inventory building curve can be very different from one period to another [44], [30].

The growing complexity within SC have had a profound impact on the operation. Since high levels of complexity also increase vulnerabilities to unexpected deviations. Even small deviations caused by reasons such as minor equipment failure or inaccurate forecasts can lead to SC disruptions. It is critical to quickly inform SC managers of such small deviations, because they occur frequently and can be found at both supply and demand sides.
• **Supply-side Deviations.** Build and delivery plans play a central role in successfully meeting customer demand. In order to reduce inventory cost, each manufacturing site is motivated to only hold a minimum safety stock to buffer normal system variations [62] [44]. When one site experiences small deviations in executing the production plan (e.g. underproduction caused by equipment failures), there may not be enough safety stock in the system which can lead to failures in delivery. Considering the interconnectedness in SC, delivery failures in one site will further impact its downstream sites and may cause serious SC disruption.

Figure 1.2: Inventory Building Curves of SKU S1-S16. Each SKU Builds Its Own Inventory to Anticipate a Delivery. Red Arrow: 15% Under Production in S14 From Time 350.
- **Demand-side Deviations.** To manage a SC, the planning system generates build and delivery plans based on demand forecasts, while the actual delivery is driven by the realized demand, which can be different from the forecast. Forecasters anticipate a range for the future demand (forecast uncertainty [21], [43]) and expect forecast error to fall within a certain range. The anticipated range of forecast errors is an important factor to be considered when determining safety stock levels. However, actual demand can lie well beyond the anticipated range when the variability in demand is underestimated. When this occurs, delivery failure is still possible even when production plans are perfectly executed.

Snyder [62] refers to minor SC disruptions such as material shortage, equipment failure, and inaccurate forecast as small scale SC disruptions. Compared to large SC disruption (e.g. disruption caused by earthquake), he notes that small scale SC disruptions occur more frequently and are harder to detect [62], [44], [38], [67], [68].

In practice, many companies leave the task of monitoring the execution of build plans to line operators or line leaders. For example, line leaders are expected to notice abnormalities such as Work-In-Process (WIP) piling up when there is an equipment failure [69]. Monitoring from such a perspective can be unreliable as people often overreact to normal variations that are present in real-world systems [61]. Additionally, when a large number of SKUs are processed with different build plans and deviations only occur to a small SKU subset, human detection can be very difficult (e.g. the deviation in the production of SKU S14 in Figure 1.2).

Some companies encode predefined deviations as events and use a Supply Chain Event Management (SCEM) system [31] to help automate the monitoring of deviations in the system. A SCEM system monitors operational metrics such as delivery time, and re-
ports events when deviations exceed their respective thresholds. Although SCEM automate
data extractions, the thresholds used for alerting are usually predefined and not sensitive to
small deviations in production [61].

To the task of estimating future forecast uncertainty, existing analytical methods
are limited in that they are primarily reactionary to large forecast errors that have already
occurred (such as mean absolute errors in previous forecasts [21] [52]) and assume the
magnitude of the uncertainty is the same regardless of market variability. These estimation
methods fail to link the range of past forecast errors to the future market conditions, thus
there is a need for proactive analytical methods which can monitor forecast uncertainty in
real-time. Such a method does not rely solely on current forecast errors, indicators of future
market variability such as trends in demand forecasts are also considered.

In this research, we propose surveillance strategies for complex SCs which allow
better monitoring of deviations that cause small scale disruptions. In Chapter 2, we first
propose a monitoring strategy to detect small deviations in executing build plans. The
strategy is based on statistical process control (SPC). In Chapter 3, we further explore
the topic and propose a monitoring strategy called Heat Map Contrasts (HMC), which
quickly detects deviations in productions when SC features both high dynamic complexity
and detailed complexity. In Chapter 4, we present a new approach for quantifying forecast
uncertainty by utilizing past forecast updates when rolling horizon forecast is implemented.
In Chapter 5, we provide conclusions and discuss future work. In the section below, we list
a brief synopsis of our contributions.
Monitor Supply-side Deviations in SC with High Dynamic Complexity

A major source of supply-side deviations are deviations in executing production plans. Traditional monitoring methods such as SCEM are insufficient in this respect, because when a SC features large instabilities in manufacturing tasks, small deviations can be easily confound with normal system variations. SPC is a promising monitoring technique for this task, but applying SPC to SC management has not been well studied and is challenging. To start with, the choice of statistic(s) is not clear. Although the ERP system records different production and logistic actions, these records reflect both deviations and normal variations. Additionally, SC data often feature strong autocorrelation, skewness, and other departures from traditional control chart assumptions. Secondly, most existing researches focus on simple SCs, applying SPC to a complex, multi-echelon SC is currently lacking in the literatures.

**Contributions**

- We evaluate the fitness of different control statistics that can detect deviations in the execution activities, such as inventory level[56], deviation in inventory level[30], and time-between-shipments[61], and propose one statistic, ship rate, that is robust to the complex system adaptations, sensitive to small deviations, applicable to general high dynamic SCs, and violating few control chart assumptions.

- To address the skewness caused by small batch production, we use weighted standard deviation EWMA and weighted standard deviation CUSUM charts to monitor the ship rates.

- We discuss estimations of control limits in practice, especially the estimation of limits from historical data with slightly different system settings, and how to deploy multi-
ple charts to a multi-echelon SC. These discussions complete a monitoring strategy that is sensitive to small over or under-productions and robust with dynamic complexities.

- This work fills the gap of applying SPC to monitor complex SC, and presents an opportunity for SPC tools to be applied in a new, rich domain of SC management.

Monitor Supply-side Deviations in SC with High Dynamic and Detail Complexities

When a SC features not only dynamic complexities but also detail complexities, new challenges in SC management arise. First, since each manufacturing site processes many SKUs, an equipment failure or material shortage only impact a small subset of these SKUs. Traditional multivariate control techniques are designed for low or medium dimension data and can lose their power to detect signals in individual or small subset of variables quickly when the dimension gets large [59] [72]. Second, after an out-of-control signal is detected, a fault diagnostic is needed to identify which SKUs are responsible for the deviation. Multivariate control chart does not come with such fault diagnostics, so additional diagnostic steps are needed. With the complex non-linear relations in manufacturing, to identify the right subset of SKUs from such a environment is also a challenging task.

Contributions

- We propose a control strategy for multi-echelon SC system with both high level dynamic and detail complexities, by establishing a multivariate control algorithm called Heat Map Contrasts (HMC). HMC converts the monitoring problem into a series of classification problems. It is robust to autocorrelations in SC data and avoid potential classification biasness. It handles high dimensional data and is capable of detecting small deviations occurring in only a small subset of SKUs. A fault diagnostic func-
tion is built inside HMC to identify deviated SKUs every time a signal is detected. We also provide discussions of deploying HMC to a complex multi-echelon SC.

- We propose a simpler technique to efficiently remove skewness in the ship rate statistic and reduce the computational burden when the number of SKUs is large. We also discuss creating SKU aggregations and aggregated statistics that are more sensitive to certain failures for better control performance.

- We propose a visualization to present information of multiple SKUs over a period in one chart. Compared to traditional visualizations as line and scatter plots [9], our visualization provides a more compact representation for easier trend detection and cross-product comparison.

Proactive Monitoring of Demand-side Deviations

Existing analytical methods for understanding demand forecast uncertainty rely solely on current forecast errors and are primarily reactionary. To adjust supply chain operational decisions for possible large forecast errors in the future, a proactively estimation of forecast uncertainty is needed. Proactive estimations of forecast uncertainties exist in econometrics research, but they either require domain knowledge to scale or skew past forecast accuracy (e.g. mean absolute forecast errors (MAE) or mean square forecast error (MSFE)) based on future market variability [37] or require the use of multiple forecasts arising from different sources (e.g. Survey of Professional Forecasts)[17], [20], [5]. These methods are not applicable for SC with high levels of detail complexities. When many SKUs are offered, it is impossible to manually scale and skew past forecast accuracy for each SKU at each time point. Unlike econometric forecasts, multiple forecasters are normally not available in demand forecasting. This requires the need for estimating demand forecast uncertainty in
real-time without relied on human knowledge or multiple forecasters, however no existing solution can be found in the current literature.

Contributions

• Many companies forecast demand in a rolling horizon fashion, but often do not adequately utilize the rich information buried in these forecasts. We quantify forecast uncertainty by utilizing past forecast updates and propose two statistics that serve as proactive indicators of future forecast errors. The statistics have shown promising results when applied to two datasets consisting of real forecast updates. Our method does not require prior knowledge of the forecast model or market conditions drive the forecast uncertainty to change.

• Unlike existing methods for analyzing forecast uncertainty, our approach allows operational adjustments to be made prior to the occurrence of forecast errors rather than reacting after they have occurred. It is more applicable to demand forecasts because it does not require multiple forecasters.

• We also discuss potential extensions such as using HMC to monitor forecast uncertainties for multiple products, false alarm pruning and extraction of more complex patterns when long forecast horizons are used.
CHAPTER 2

Monitor Supply-side Deviations with High Dynamic Complexity

Driven by rapid technological development and complications in product design, the complexities of supply chains (SC) have grown rapidly and have had a profound impact on their operation [12]. In a complex SC, production of each individual finished product involves many supporting parts requiring the cooperation of multiple manufacturing sites. A planning system generates build and delivery plans to coordinate materials movements between manufacturing sites. Accurately executing those plans play a central role in successfully meeting customer demand, and even small deviations in plan execution can lead to failures in delivery, further impacting downstream sites and causing serious SC disruption.

An important source of deviations in plan executions are production failures such as equipment failures, material shortages and operator errors. SC disruptions caused by such failures are referred to as small scale SC disruptions [62]. Researchers also point out that, compared to large scale SC disruptions (disruptions caused by hurricanes or earthquakes), small scale SC disruptions occur more frequently and are much harder to detect, especially in a SC with high dynamic complexity [38, 44, 67, 68].

When a SC features high dynamic complexity, productions often are made in small batches in order to accommodate low volume productions. For each production cycle, a SKU builds its own inventory to anticipate a delivery that consumes much of the available stock. The recurrent cycles of inventory build followed by deliveries repeat, but at infrequent, non-periodic intervals. Driven by changing delivery sizes, equipment and material availability, the inventory building curve can be very different from one to another. Under such context, a small underproduction can be easily confounded with normal production variations (Figure 1.2).
For example, when an equipment is still function but lost a small fraction of its capacity and causes a small underproduction, detecting the small capacity loss is difficult. In practice, it is common that productions involve different sets of machines, and people working in the field are expected to notice abnormalities such as Work-In-Process (WIP) piling up when there is an equipment failure [69]. Even in case that performance data of all individual equipment are automatically extracted and monitored, the monitoring is often very simple. Because the performance data recorded are inconsistent with different equipment suppliers, firms often use event management [31] and signal only when the performance measure (such as processing time) exceeds a threshold, which is also predefined by people work in the field. Monitoring from such a perspective can be both insensitive to small deviations and overreactive to normal variations that are present in real-world systems [61].

Statistical process control (SPC) is a promising monitoring technique for small scale SC disruptions. Control charts can be designed to be sensitive to small deviations or trends, and normal variations are appropriately handled to effectively reduce the risk of overreaction. However, it is a challenge to apply SPC to a SC with high dynamic complexity while still obtaining desired results, and only limited research has considered the problem. First of all, the choice of monitoring statistic(s) is not clear. From an implementation perspective, to avoid additional data collection, the statistic needs to be derived from data in enterprise resource planning (ERP) system. ERP systems record different production and logistic actions and multiple statistics can be extracted from ERP records. They often feature strong autocorrelation, skewness, and other departures from traditional control chart assumptions that reduce model detectability. For example, a quantity such as inventory level is often strongly autocorrelated, and autocorrelation affects detection, which is illustrated later. Given these difficulties, finding an appropriate statistic that is robust to the complex system adaptations, sensitive to small deviations, applicable to gen-
eral high dynamic SCs and violates least control chart assumptions is challenging. Second, variations in SC data often depend on the system settings such as production speed and equipment utilizations. With high dynamics present, control limits need to be updated frequently, and it is also not clear that how to estimate the control limits from historical data and when they should be updated.

Motivated by these facts, we introduce a SPC based monitoring strategy for SCs with high dynamic complexities to detect small scale disruptions that can potentially lead to more serious SC problems. We conceptualize the supply chain system, then evaluate the fitness of various statistics commonly reported in an ERP system or used in other applications. We first evaluate inventory level related statistics [29, 45, 56, 63] and exclude due to their strong autocorrelation. We then evaluate sojourn time [61] and exclude it for not being robust to complex manufacturing. Finally, we propose a single and effective control statistic that has desirable properties.

We evaluate different charting strategies and propose a strategy of deploying multiple univariate control charts. Each univariate chart is designed to be robust to SC data and sensitive to small deviation, and we discuss how to estimate control limits in practice. We also provide a discussion on charts deployment in multiple-echelon SCs, and for which the only weakly-related existing literature is chart allocation for serial-parallel multistage manufacturing process [35], that research is developed to control product quality and is different from our goal.

The remainder of this study is organized as follows: the following section provides a brief background on SC monitoring. The next section illustrates the selection of monitoring statistics, charts design and deployment. Then an experiments section conducts three sets of experiments to demonstrate the viability of this SPC strategy with a representative data from a semiconductor SC. The final section concludes our work.
2.1 Background

This section provides a brief background on complex SC and current monitoring techniques, SPC and its applications in SC management, and control charts for skewed population. Figure 1.1 illustrates a complex SC. Productions of each finished product involve many supporting parts and require the cooperation of multiple manufacturing sites. Demand forecasts for each individual product are fed into a planning system, and the planning system generates build and delivery plans to coordinate material movements among different sites. Such planning systems are capable of dealing with the complex interactions required to link productions with demands. They consider various factors such as material and equipment availability, bill of materials (BOM), production lead times, and demand variabilities.

Statistical process control (SPC) is one of the most important and widely used monitoring tools in quality control. It is useful to detect deviations from a baseline state and various control charts have been designed to be sensitive to different types of deviations given different data characteristics. However, SPC approaches have only been used to a limited extent in SC management.

A monitoring method based on individualized trace data (ITD) recorded in radio frequency identification (RFID) application was proposed by Shu and Barton [61]. They monitored the sojourn time between two RFID readings, however, their method requires the entity to be traceable and it becomes more challenging for a process where one SKU can be processed into other SKUs (like a wafer SKU being cut into different die SKUs). Furthermore, some processes hold lots, while others proceed, and such a purposeful delay would need to be accounted for in the sojourn times. With many arrangements of the schedule this becomes a greater challenge. The authors did not discuss how control charts
should be deployed. More importantly, they assumed the system is stable and didn’t discuss how to extend it to a dynamic process where inventory is built to anticipate different sizes of deliveries.

Spearman et al. [63] proposed Statistical Throughput Control to determine whether or not a production quota is likely to be achieved. For a production period of length \( R \) and a target production volume \( Q \), at a time pint \( t \), they monitor the deviation in actual cumulative production volume \( N_t \) from a baseline defined by a linear production \( \frac{Qt}{R} \), that is \( d_t = N_t - \frac{Qt}{R} \). Spearman et al. [63] assumed \( d_t \) to be independently normally distributed. However, in small batch productions, the \( d_t \) can be highly autocorrelated, which inflates false alarms which we illustrated later. Additionally, they assume the largest variation in \( d_t \) always occurs at the end of the production period, and this assumption can also be violated because operators often trade capacities between SKUs.

Other applications include an SPC-based replenishment policy. Pfohl et al. [56] built control charts for demand and inventory level. Replenishing rules are set according to historical inventory and demand data to optimize the re-order policy. Lee et al. [45] modified Pfohl’s method by examining the bull whip effect caused by order batching and compared the traditional event-triggered and time-triggered inventory policies against the SPC-based replenishment method for a two-echelon SC. However, both studies focus on simple SC with one product characterized by stable demand and are hard to be extended to systems with high dynamic complexity. Other SPC applications include detection of forecasting errors [3] and monitoring inventory-record accuracy [22], which are different topics from our focus.

Another monitoring technique used today is SCEM [31, 53]. Events are monitored and a signal is triggered when deviation in an operational statistic exceeds a predefined threshold, which is often determined by SC managers. An important concern is that a pre-
defined threshold is not sensitive to a small underproduction being confounded with system adaptations. Without a statistical basis, SC managers can overreact to normal variation.

This paragraph reviews monitoring techniques for a skewed population. According to Change et al. [10], there are three approaches to tackle this problem. One approach is to increase the sample size to sufficiently large value so that the sample mean becomes approximately normally distributed. It is expensive, and not applicable to our case of real time monitoring. A second approach is to assume the underlying distribution is known and use contour charts to specify the desired false alarm rate. This is not applicable for the unknown distribution. A third method is to use a heuristic to obtain a control chart. One heuristic is a weighted standard deviation (WSD) method proposed by Chang [10]. The basic idea is to split a skewed distribution at its mean, then use different distributions to model each segment. Atta et al. [1] compared different WSD charts and concluded that both EWMA-WSD and CUSUM-WSD are sensitive to small deviations.

2.2 Monitor SC with Dynamic Complexity

Consider one manufacturing site in the SC in Figure 1.1 and the production of SKU $k$ is illustrated in Figure 2.1. Materials received from its upper stream sites are first stored in the received inventory, then produced in SKU $k$ based on a build plan received from

![Figure 2.1: Production of SKU $k$ in a Manufacturing Site.](image)
the planning system. After production, finished products (SKU \(k\)) are moved to the finished inventory and wait to be delivered in bulk to downstream sites. When high dynamic complexity presents, the productions are often in small batch sizes to accommodate low production volumes [6]. When a batch of SKU \(k\) finishes processing and enters the finished inventory at time \(u\), we refer to the batch as a shipment, and denote the batch size as \(s_k(u)\), where \(k\) is the SKU ID and \(u\) is the process finishing time. We further denote finished inventory level of SKU \(k\) at time \(u\) as \(I_k(u)\), and use term “delivery” to refer to a bulk of SKU \(k\) being transported to a downstream site.

2.2.1 Control Strategy

The control strategy we consider requires adjustments to the dynamics in inventory build driven by different delivery sizes and we consider the following characteristics in our design. From a practical perspective, the control statistic should be compatible with the ERP system. It is either directly retrieved from the ERP system or can be derived from the ERP data. The control statistic should also be reported frequently so that quick detection is possible. Furthermore, it is desirable that the control statistic be approximately normally distributed with weak autocorrelation, because standard SPC control charts are based on these assumptions.

Two possible control statistics are inventory quantity [56] and deviations in inventory quantity [63]. They are often strongly autocorrelated, and autocorrelation can cause high false alarm rates. This is aggravated in a monitoring system with multiple control charts. Furthermore, increasing the report frequency to improve detection yields an even stronger autocorrelation. Another possible control statistic is the shipment quantity \(s_k(u)\). A shipment quantity is available whenever a new batch of SKUs finishes processing, so it is updated frequently and this can facilitate a responsive control chart. In some industries, the
time interval between two consecutive shipments can be as short as one minute (or less). However, shipment quantity has its own disadvantages. First a skewed distribution is common. We have studied shipment data in semiconductor industry and found most shipments are small while large shipments are rare (Figure 2.2). This characteristic is expected to be common to other industries. The second disadvantage is that a shipment quantity is not recorded periodically, so that there can be large gaps without any values.

![Shipment Quantity Over Time](image)

Figure 2.2: Shipment Quantity Plots of One SKU from a Semiconductor SC Simulator.

Shipment quantity is a step closer to a useful control statistic, but a key element for SC monitoring is whether the rate of production is sufficient to meet the build plan. That is, sufficient product should be available at the time of a delivery. By analogy to water flow in a pipe, the volume per unit time needs to be sufficient to meet the total volume required at a specified time. Consequently, this leads one to consider the rate of production. Towards
this end, a time window of length $w$ is selected. The ship rate at $t$ is defined as the integral (sum) the the shipment quantities from time $t - w$ to $t$, divided by the window length $w$. This is analogous to a flow rate for a fluid measured in cubic meters per second. We denote the ship rate of SKU $k$ at time $t$ with window size equals to $w$ as

$$r_k(t) = \frac{\int_{t-w}^{t} s_k(u)\,du}{w} \quad (2.1)$$

We compute $r_k(t)$ at discrete time points $t$ and each $t$ represents a time period of length $w$. The time span between $t$ and $t + 1$ is called a stride $\gamma$; typically we let the stride $\gamma$ equal to $w$, but $\gamma < w$ can also be used. In this way we obtain a series of $r_k(t)$ for $t = 1, 2, \cdots, T$, and control charts based on discrete time can be applied. Figure 2.3 shows an example of converting shipments $s_k(u)$ to ship rates $r_k(t)$.

Figure 2.3: Illustration of Converting $s_k(u)$ to $r_k(t)$.

In our experiments, ship rates are weakly autocorrelated so that control charts can be applied directly. Different values for $w$ can be selected based on the industry and problem context. Generally speaking, when applying a larger $w$, $r_k(t)$ is smoother and more
symmetric, but updated less often. A smaller $w$ typically leads to a more skewed $r_k(t)$ with frequent updating. We provide illustrations and use ship rate $r_k(t)$ as the monitoring statistic for further model building.

![Diagram](image)

**Figure 2.4:** Baseline Inventory and Actual Inventory of SKU $k$ at a Manufacturing Site.

For each site, a delivery plan indicating upcoming deliveries for different SKUs is known in advance, and a build plan specifies the production volumes of different SKUs. When build plans are accurately executed, each delivery consumes most available inventory when it occurs. For a perfectly balanced production process, the inventory of SKU $k$ grows linearly between two consecutive deliveries. We refer to the linear inventory increase as the baseline inventory, and denote the baseline inventory level at $t$ as $BI_k(t)$. Driven by different demand signals, the baseline inventory can ramp at different speed for different production cycles. The actual inventory may not be perfectly balanced and can fluctuate around $BI_k(t)$. We denote the actual inventory level at $t$ as $I_k(t)$.

Figure 2.4 illustrates inventory curves of SKU $k$ at three deliveries occurring at $t_1$, $t_2$ and $t_3$. We use $t_1 -$ and $t_1 +$ to indicate the time right before and after the delivery at $t_1$, respectively.
and same applies for \( t_2 \) and \( t_3 \). Assuming a starting inventory \( I_{t_0} \), the baseline inventory is expected to reach \( BI_k(t_1-) \) according to the build plan. Then, delivery at \( t_1 \) consumes most of the available inventory and brings the baseline inventory to \( BI_k(t_1+) \). The actual inventory curves increases when shipments arrive at the finished inventory, and due to the normal variation in the production system, \( I_k(t) \) fluctuates around \( BI_k(t) \). It reaches \( I_k(t_1-) \) at \( t_1 \) which is slightly larger than \( BI_k(t_1-) \) and delivery is made successfully. The planning system updates the build and delivery plans at \( t_1+ \) and start another production cycle. The next delivery is anticipated to occur at \( t_2 \), and compared to period \([t_0,t_1]\), both \( BI_k(t) \) and \( I_k(t) \) ramp much faster in \([t_1,t_2]\) to anticipate a much larger delivery. The actual inventory level reaches \( I_k(t_2-) \); although it is less than \( BI_k(t_2-) \), safety stock held is used for the slight underproduction caused by normal system variations, and delivery is still made successfully. At time \( t \), an equipment issue causes \( I_k(t) \) to deviate from \( BI_k(t) \), and this leads to a significant underproduction at \( t_3 \). The safety stock is no longer sufficient and a stock out occurs. However, with better control of the inventory build up, both the safety stock and the risk of stock out can be potentially reduced.

![Figure 2.5: Build Plan Changes at \( t_1, t_2 \) and \( b_k(t) \) Changes Accordingly. When Failures occur at \( t \), \( r_k(t) \) Deviates from \( b_k(t) \).](image-url)
As mentioned, the \( r_k(t) \) statistic features less autocorrelation and is a useful statistic for control charts. For a production cycle \([t_i, t_{i+1}]\), a baseline ship rate of SKU \( k \), \( b_k(t) \), can be derived from \( BI_k(t) \) as

\[
b_k = \frac{BI_k(t_{i+1}^-) - BI_k(t_i^+)}{t_{i+1} - t_i}w \tag{2.2}
\]

The actual ship rate \( r_k(t) \) is the production rate in period \([t - w, t]\), and fluctuates around its baseline \( b_k \) when the system is under control. The baseline ship rate \( b_k \) changes when the slope of \( BI_k(t) \) changes, and we expect the in-control means of \( r_k(t) \) to change accordingly. This type of change should not be flagged, besides, it is common to have normal variations depend on production system settings such as production speed. Therefore, a control model of \( r_k(t) \) needs to be adjusted when \( b_k \) changes. When issues like equipment failures or material shortages occur, \( r_k(t) \) starts to deviate from the corresponding baseline \( b_k \) more than the natural variation. We extract \( r_k(t) \) and \( b_k \) from Figure 2.4 and plot them in Figure 2.5. The baseline ship rate \( b_k \) is recomputed at \( t_1^+, t_2^+ \), as the slope of \( BI_k(t) \) changes. At time \( t \), \( r_k(t) \) starts to deviate from its baseline \( b_k \) because of the equipment issue and falls below lower control limit.

2.2.2 Control Chart Design

Different failures result in deviations in different subsets of SKUs. Therefore the \( r_k(t) \) of multiple SKUs need to be monitored simultaneously. Using a multivariate control chart to monitor \( r_k(t) \) in \( K \) SKUs faces several difficulties. Covariance matrices \( \Sigma \) of \( r_k(t) \) in multiple SKUs depend on system settings such as equipment assignments and production rates. In order to obtain a reasonable estimation of \( \Sigma \), we need sufficient in-control data from periods with identical system settings. For example, to estimate \( \Sigma \) for 10 SKUs, one needs a historical period when all 10 SKUs were produced with the same production rates and equipment settings as in the target period. Because production of each SKU is driven
by its own demand, with high demand variabilities present, finding such training data is a challenge. Additionally, multivariate control chart does not come with fault diagnostics, so additional diagnostic steps are needed after an out-of-control signal is detected [32, 46, 50]. With the complex non-linear relations in manufacturing, identifying the right subset of SKUs from such an environment is also a challenging task.

Considering the difficulties in implementing multivariate control chart, we propose to use multiple univariate control charts. We chart one SKU at a time and do not need the covariance among SKUs.

We first discuss the design of control chart for ship rate in an individual SKU and the estimation of the control limits. We make the reasonable assumptions that the delivery plan is known in advance and we consider the design of a control chart for \( r_k(t) \) of a specific period. During this period, the slope of \( BI_k(t) \) does not change and the baseline ship rate is a fixed value \( b_k \).

Ship rate is expected to be skewed in many applications because there can be cases with many small shipments followed by a few large ones. The usual control limit estimates in EWMA and CUSUM charts can produce a chart that is insensitive to ship rate decreases (which is the primary concern in practice). Atta [1] compared different chart designs for skewed data and concluded that EWMA-WSD and CUSUM-WSD [10] are sensitive to small deviations.

Assume historical in-control production periods with the same \( b_k \) value are available and “in-control” means the historical period has no assignable failures such as equipment failure. To construct EWMA-WSD chart, we first compute \( r_k(t) \), then \( \hat{\theta} \) which is the proportion of \( r_k(t) \) that is less than or equal to \( b_k \) in the training data. Then, we calculate the EWMA statistic, \( z_k(t) \), with certain smoothing parameter \( \lambda \), and denote the standard
deviation of $z_k(t)$ as $\hat{\sigma}_z$. The EWMA-WSD is defined as Equation 2.3 where $L$ is a user defined parameter

$$UCL_{EWMA-WSD} = b_k + L\sigma_z 2\hat{\theta} \tag{2.3a}$$

$$LCL_{EWMA-WSD} = b_k - L\sigma_z 2(1 - \hat{\theta}) \tag{2.3b}$$

When $r_k(t)$ is positively skewed and $\theta > 0.5$, the EWMA-WSD chart tends to have tighter lower control and wider upper control.

The WSD version of CUSUM chart (CUSUM-WSD) can be easily calculated from a traditional CUSUM chart as follows.

$$C_{WSD}^+(t) = \max[0, r(t) - (b_k + h\hat{\sigma}_r) + C_{WSD}^+(t-1)] = \frac{C^+(t)}{2\hat{\theta}} \tag{2.4a}$$

$$C_{WSD}^-(t) = \max[0, (b_k - h\hat{\sigma}_r) - r(t) + C_{WSD}^-(t-1)] = \frac{C^-(t)}{2(1 - \hat{\theta})} \tag{2.4b}$$

$\hat{\sigma}_r$ is the standard deviation of $r_k(t)$ in the training data, and $h$ is a user defined parameter. CUSUM-WSD charts issue an out-of-control alarm at the first $t$ for which $C_{WSD}^+(t) > H\sigma_r$ or $C_{WSD}^-(t) > H\sigma_r$ where $H$ is also an user defined parameter.

With high dynamic complexities present, it is possible that periods with baseline ship rates that are identical to $b_k$ cannot be found. In this case, we use a single parameter $\delta$ and simply search for baseline ship rates within $(b_k - \delta, b_k + \delta)$, and use production data from those similar periods for control limit estimation.

Assume one similar baseline ship rate is $b_k'$, and the ship rate computed from $b_k'$ is $r_k'(t)$. If we denote the standard deviation of $r_k'(t)$ as $\hat{\sigma}_{r'}$, and the proportion of $r_k'(t)$ being less than $b_k'$ as $\hat{\theta}_{r'}$, to estimate control limits for CUSUM-WSD chart, we replace $\hat{\sigma}_r$ with $\hat{\sigma}_{r'}$ and $\hat{\theta}$ with $\hat{\theta}_{r'}$ in Equation 2.4. To estimate control limits for EWMA-WSD chart, in addition to compute $\hat{\theta}_{r'}$, we also need to compute the EWMA statistic for $r_k'(t)$ and its standard deviation $\hat{\sigma}_{z'}$. Then we replace $\hat{\sigma}_z$ with $\hat{\sigma}_{z'}$ as well as $\hat{\theta}$ with $\hat{\theta}_{r'}$ in Equation 2.3.
To monitor a multi-product, multi-echelon SC, and provide details of where SC disruptions occur along with which SKUs are affected, we need to deploy multiple charts. Monitoring at the level of \( r_k(t) \) is essentially monitoring flows in the SC and can provide the visibility at the finest level.

When monitoring a small SC, the strategy of monitoring each \( r_k(t) \) does not produce a large number of control charts. However, for a SC with many SKUs processed, monitoring each \( r_k(t) \) can generate a large volume of control charts and inflate the risk of false alarms even if the false alarm rate for each individual chart is low.

For such SCs, one approach is to identify the "bottlenecks". Bottlenecks can be facilities or SKUs with a high probability of supply interruptions, long lead and reaction times, high equipment utilization, frequent engineering changes, etc [34, 51, 65, 73]. Once bottlenecks are identified, control charts can be applied to ship rates in SKUs associated with those bottlenecks to reduce the control chart use.

Another way to reduce the number of control charts is to monitor only aggregations of \( r_k(t) \). One possible aggregation is by shared equipment or shared parental SKU (Figure 2.6). When equipment loses capacity, the SKUs produced on that equipment face shortages. Also, shortage of a particular SKU impacts all the children SKUs derived from it. Prior knowledge of equipment assignments or the bill of materials (BOM) can be used for the grouping. In the semiconductor industry, those two groupings often obtain similar results because SKUs processed from the same parents often have similar technical features. Therefore, they are often processed by the same set of equipment.

Assume there are \( K \) SKUs in group \( j \) and their ship rates and baseline ship rates are \( r_k(t) \) and \( b_k(t) \) respectively. The aggregated ship rate for group \( v \) can be calculated by

\[
R_j(t) = \sum_{k=1}^{K} r_k(t) \tag{2.5}
\]
Figure 2.6: Grouping SKUs Based on Their Parental SKUs or Shared Equipments.

Often, $R_j(t)$ becomes a better monitoring choice because ship rates in those $K$ SKUs are negatively correlated due to the competition for limited manufacturing resources. Aggregation of $K$ ship rates will have greater signal-to-noise ratio when a underproduction occurs to all SKUs in the group. To estimate a control parameter for $R_j(t, w)$, we only need to find a period with summed baseline ship rate being close to the summed baseline ship rate in the target period regardless of the value of individual $b_k(t)$. When such a historical period cannot be found in group $j$, periods from other groups with similar aggregated baseline build plans can alternatively be used.

### 2.3 Experiments

Simulation experiments are used to illustrate that the control strategy is viable for a representative SC. We illustrate that control charts can be designed from historical data, false alarms can be controlled, and assignable causes can be detected with this approach. We show the effects of different window sizes and also consider different control charts applied to ship rates. Additionally, we consider the sensitivity of the control limit estimates
to the historical data selected. For confidentiality reasons, actual data cannot be reported, but our simulations are representative of actual SC performance.

Rather than assume distributions for statistics, a simulation of a multi-product, multi-echelon semiconductor SC is built as a data test bed. See Figure 2.7. Wafers are manufactured in Fabrication sites, then delivered to Material Warehouses sites ($M_1$-$M_2$) and processed into different types of semi-finished products. Semi-finished products are then shipped to assembly sites ($A_1$-$A_4$) where they are further assembled to different finished products, tested and delivered to Hubs or Warehouses. Inventories are held in all facilities, and in Figure 2.7 inventory of raw material, semi-finished products and finished products are marked as yellow, grey and purple, respectively.

![Simulation Model of a Multi-echelon, Multi-product SC.](image)

Figure 2.7: Simulation Model of a Multi-echelon, Multi-product SC.

The parameters in the simulation model are tuned so that simulation output is representative of a semiconductor SC. For example, the shipment size follows an exponential
distributions with a mean equals to 10 units, the process time of each unit follows a normal distribution and the transportation time between facilities follows a normal distribution. An advantage of using simulation is that changes can be imposed with exact knowledge of when and where they occur and which SKUs are affected. Therefore, detection time and other performance measures like false alarm rates can be calculated accurately.

2.3.1 Initial Analysis

For an initial analysis, we first examine ship rates under various baseline ship rates. We study the distributions of $r_k(t)$ obtained from a wide range of $b_k$ values (from 0.2 unit per hour to 80 units per hour) with $w = \gamma = 4$ hours. The skewness measure $\hat{\theta}$ and standard deviation $\hat{\sigma}_r$ of different $r_k(t)$ distributions are plotted in Figure 2.8. Generally, low $b_k$ values lead to more skewed $r_k(t)$, and as $b_k$ increases, the skewness decreases. In the experiment, $r_k(t)$ becomes normally distributed when $b_k > 40$ units per hour.

![Figure 2.8: Different Skewness in $r_k(t)$ from Different $b_k$ Values. Low $b_k$ Values Lead to More Skewed and Varied $r_k(t)$. Window Size $w = 4$.](image-url)
We then compare inventory level, deviation in inventory and ship rates with $b_k = 20$ unit per hour. We choose $b_k = 20$ unit per hour because $r_k(t)$ generated under this setting features moderate skewness of $\hat{\theta} = 0.6$ and can better test our method. Here $r_k(t)$ is computed from $w = \gamma = 4$ hours and inventory levels and deviations in inventory levels are also reported at the end of every four-hour period. The results are illustrated in Figure 2.9. When productions are in small batches, both inventory levels and deviation in inventory exhibit strong autocorrelation, and the autocorrelations in ship rates are much weaker and almost negligible.

Figure 2.9: Due to Small Batch Production, Inventory Levels and Inventory Deviations Exhibit Strong Autocorrelation. Autocorrelations in Ship Rates Are Weaker.
2.3.2 False Alarms

In the following experiments, EWMA, EWMA-WSD, CUSUM, CUSUM-WSD charts are evaluated. We set $\lambda = 0.1$ and $L = 2.81$ for EWMA and EWMA-WSD according to Lucas et al. [49], and $H = 5\hat{\sigma}_0$ and $k = 0.5\sigma_0$ for CUSUM and CUSUM-WSD according to Hawkins et al. [25], so that the in-control average run length, $ARL_0$, for all four charts are controlled to be 500 points. Here, $\hat{\sigma}_0$ denotes the in-control standard deviation of ship rates.

We use SKU FG4 in facility $A_1$ as the target and denote its ship rate as $r_0(t)$ and set its baseline ship rate $b_0$ to 20 units per hour. In the following experiments, we apply different control charts to monitor $r_0(t)$ and evaluate false alarms. We use three different window sizes, $w = 2, 4, 8$ hours and let $\gamma = w$. A simulation of 7200 hours is generate to estimate the control limits with different window sizes. Another 1000 replicates are generated to evaluate the in-control run lengths. Each replicate is simulated for 16000 hours, and long replicate length is used to guarantee signals always occur even for loose control limits. The average in-control run length ($ARL_0$) and standard error ($SE_0$) for each chart with different window sizes are reported in Table 2.1 and 2.2.

Table 2.1 shows $ARL_0$ of EWMA chart and EWMA-WSD chart are close to expected. When $L = 2.81$ and $\lambda = 0.1$, $ARL_0$ is expected to be around 500 [49], and $ARL_0$ of EWMA chart is between 570 to 750 and $ARL_0$ of EWMA-WSD is between 550 to 650.

Table 2.1: $ARL_0$ of EWMA-WSD and EWMA. $\lambda = 0.1$ and $L = 2.81$ Are Used.

<table>
<thead>
<tr>
<th>$w$</th>
<th>EWMA, $L=2.81$</th>
<th>EWMA-WSD, $L=2.81$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$ARL_0$</td>
<td>$SE_0$</td>
</tr>
<tr>
<td>2 h</td>
<td>611</td>
<td>16.8</td>
</tr>
<tr>
<td>4 h</td>
<td>728</td>
<td>21.2</td>
</tr>
<tr>
<td>8 h</td>
<td>575</td>
<td>16.2</td>
</tr>
</tbody>
</table>
Table 2.2 shows that the similar conclusion can be obtained from CUSUM and CUSUM-WSD. When $h = 0.5 \hat{\sigma}_0, H = 5 \hat{\sigma}_0$, $ARL_0$ of both CUSUM and CUSUM-WSD are between 500 to 650 which is close to the theoretical value. However, if taking a look at $ARL_0$ of positive and negative cumulative sum respectively, a very different result is presented. Due to the skewness in ship rate, $ARL_0$ for negative CUSUM and negative CUSUM-WSD ($ARL_0^-$) are much larger than that of positive CUSUM and CUSUM-WSD ($ARL_0^+$), especially when $w = 2$ and 4 hours. The results suggest, despite the existence of skewness, $ARL_0$ can still be approximately set from simple calculations for EWMA, EWMA-WSD, CUSUM and CUSUM-WSD.

Table 2.2: $ARL_0$ of CUSUM and CUSUM-WSD Charts with $h = 0.5 \hat{\sigma}_0$ and $H = 5 \hat{\sigma}_0$. $ARL_0$ in CUSUM and CUSUM-WSD Are Close to Expected.

<table>
<thead>
<tr>
<th>$w$</th>
<th>CUSUM, $h = 0.5 \hat{\sigma}_0, H = 5 \hat{\sigma}_0$</th>
<th>CUSUM-WSD, $h = 0.5 \hat{\sigma}_0, H = 5 \hat{\sigma}_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$ARL_0^+$ $SE_{ARL_0^+}$</td>
<td>$ARL_0^- SE_{ARL_0^-}$</td>
</tr>
<tr>
<td>2</td>
<td>535 15.7 2983 75.3</td>
<td>500 14.5</td>
</tr>
<tr>
<td>4</td>
<td>636 17.8 1590 41.4</td>
<td>548 15.7</td>
</tr>
<tr>
<td>8</td>
<td>527 14.2 700 15.5</td>
<td>464 13.5</td>
</tr>
</tbody>
</table>

2.3.3 Signal Detection

The objective of the second experiment is to demonstrate and compare the ability of different control charts to detect shifts. We choose the same target $r_0(t)$ and keep its baseline unchanged at $b_0 = 20$ units per hour. The control limits learned from the false alarm analysis is used for this experiment. A simulation of 1000 replicates is generated for testing. Within each replicate of the testing simulation, a small change is introduced at hour 801 and deviates $b_0$ from 20 unit per hour to 17 units per hour (15% decrease) to simulate an equipment failure. When $w = 4$ and $b_4 = 20$, the in-control standard deviation of $r_0(t)$, $\hat{\sigma}_0$ is about 10 (Figure 2.8), and the deviation is very small at only $0.3 \hat{\sigma}_0$. The same four control charts are used for detection. Parameter settings in those charts are the same as in the false alarm analysis, so that the $ARL_0$ for different charts can be found in Table 2.1
and Table 2.2. The mean and standard error of run length for different charts to detect the change, denoted as $ARL_1$ and $SE_1$, are reported in Table 2.3. Column $ND$ indicates the number of replicates that a control chart fails to detect the change within 800 hours, and if $ND > 0$, $ARL_1$ is underestimated.

Table 2.3: $ARL_1$ and $SE_1$ for Four Control Charts. Column $ND$ Indicates the Number of Replicates Chart Fails to Detect Change within 800 Hours.

<table>
<thead>
<tr>
<th>$w$</th>
<th>EWMA-WSD</th>
<th>EWMA</th>
<th>CUSUM-WSD</th>
<th>CUSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$ARL_1$</td>
<td>$SE_1$</td>
<td>$ND$</td>
<td>$ARL_1$</td>
</tr>
<tr>
<td>2</td>
<td>35.6</td>
<td>1.2</td>
<td>0</td>
<td>54.7</td>
</tr>
<tr>
<td>4</td>
<td>20.1</td>
<td>1.6</td>
<td>0</td>
<td>23.6</td>
</tr>
<tr>
<td>8</td>
<td>10.7</td>
<td>1.7</td>
<td>0</td>
<td>13.9</td>
</tr>
</tbody>
</table>

In Table 2.3, because of the skewness in ship rate, EWMA and CUSUM charts detect the small deviation slower than their WSD versions. The improvements in detection speed become more obvious as $w$ decreases from 8 hours to 2 hours. EWMA-WSD provides the most agile detection, the small deviation is detected in 70 hours for $w = 2$ hours. Considering $ARL_0$ for EWMA-WSD for $w = 2$ hours is around $554 \times 2 = 1108$ hours (47 days), we can further tighten the control limits to accelerate the speed of detection. The results suggest that small deviation in production can be quickly identified if the skewness is properly handled.

2.3.4 Estimation of Control Limits

In this experiment, we test estimating control limits from data when a similar but not identical baseline ship rate are implemented. We still choose the same target SKU and our goal is to estimate the control limits for $r_0(t)$ from data generated from $b_k$ that is slightly different from $b_0$. We only consider EWMA-WSD chart with $L = 2.81$, $\lambda = 0.1$ and $w = \gamma = 4$ hours in this experiment, and similar conclusions can be obtained for other charts. Control
limits learned from the previous experiments are used as the true values and are denoted as $UCL$ and $LCL$.

We let a similar build plan be $b_k = (1 + \delta)b_0$ and generate 15 $b_k$ values with different $\delta$ values ranging from -0.35 to 0.35. We then simulate each $b_k$ for 168 hours (a week) and repeated for 100 times. Simulation length is set to be a week, because in real SC, build plan is often updated every week. For each replicate, we compute $r_k(t)$ with $w = \gamma = 4$ and compute the lower control limits according to Equation 2.3. The lower control limits obtained from $r_k(t)$ are denoted as $\hat{LCL}$ and are plotted in Figure 2.10. We only present the estimations of $LCL$ because underproduction is our major concern. Similar result can be expected for $UCL$. The result in Figure 2.10 suggests that if one can not find historical data with identical baseline ship rates, SKUs with similar baseline ship rates ($b_k$ with $-0.1 < \delta < 0.1$ can also provide reasonably close estimations. For $b_k$ that is significantly less than

![Figure 2.10: Estimations of Control Limits for EWMA-WSD for $r_0(t)$ from Similar Baseline Build Plans with $w = \gamma = 4$ hours and $L = 2.81$.](image)

estimations of $LCL$ because underproduction is our major concern. Similar result can be expected for $UCL$. The result in Figure 2.10 suggests that if one can not find historical data with identical baseline ship rates, SKUs with similar baseline ship rates ($b_k$ with $-0.1 < \delta < 0.1$ can also provide reasonably close estimations. For $b_k$ that is significantly less than
such as \( \delta < -0.2 \), we tends to over-estimate the LCL. That is because \( r_k(t) \) from those \( b_k \) values features stronger skewness (Figure 2.8) and inflates \( \hat{\theta} \). Similarly, \( r_k(t) \) from \( b_k \) which is significantly greater than \( b_0 (\delta > 0.25) \) will underestimate \( \hat{\theta} \) and result a loose LCL. For Figure 2.8 we know that the skewness of \( r_k(t) \) are less different when \( b_k \) values are large. It is more likely to obtain a control limit estimate that is close to the true one from a similar SKU when the target SKU has high volume. In Figure 2.10, similar SKUs provide reasonably good control limits estimations for \( r_0(t) \), even though \( r_0(t) \) exhibits moderate skewness, and we believe that except for some extremely low volume SKUs, this estimation approach should be applicable in most cases.

### 2.4 Conclusions

We propose a SPC-based monitoring strategy for SCs featuring high levels of dynamic complexity by utilizing a control statistic, ship rate, which is sensitive to small deviations in plan executions, easy to access in ERP system, and features weak autocorrelation. To address the skewness caused by small batch production, we use weighted standard deviation EWMA and weighted standard deviation CUSUM charts to monitor the ship rates. We provide guidance for automatically updating control limits from historical production data when there are changes in production settings. A chart deploying strategy is provided to allow charts to be effectively deployed to bottlenecks of the system. This monitoring solution provides agile detection of small deviations in executing production plans in complex SC, and presents an opportunity for SPC tools to be applied in a new, rich domain of SC management.
CHAPTER 3

Monitor Production Execution in SC with High Dynamic and Detail Complexity

Internal SC nowadays can be large and complex systems that span multiple manufacturing and distribution sites [12]. Driven by life cycle shortening, lean production, product variety and customization levels increases, those SCs often feature high dynamic and detailed complexities [6]. Companies often put in place complex planning module to coordinate productions and logistics in different manufacturing sites [69]. Each facility receives build and delivery plans from the planning module for different SKUs, and it is critical to ensure those plans are well executed, because, with small safety stock held, even small underproductions can result in stock outs [69].

Snyder et al. [62] referred those stock outs caused by issues such as material shortage, equipment failures to as small scale SC disruptions. Compared to large SC disruptions (i.e. disruption caused by earthquake), small SC disruptions occur more frequently and can have ripple effects [62, 68]. The key to prevent small SC disruption is to agilely detect underproductions. It is not easy, because in complex SC, production plans are driven by the demand, they can be very different from one period to another and feature different normal variations. Besides, underproductions can be small and, with many SKU processed in one facility, occur to only a small subset SKUs (For example, SKUs processed by the failed equipment). Figure 1.2 illustrates the finished inventory of 16 SKUs in one simulated manufacturing site. A 15% underproduction is introduced to SKU S14 from time 380 (arrow pointed), and the underproduction is confounded with normal variation and changing production schedules.

In practice, detections of underproductions are often relied on people, such as factory managers aware work-in-process (WIP) piling up. Human can be biased and respond
to normal variations in the real world. Another solution for such complex system is to encode predefined deviations as events and monitored through a supply chain event management system [31]. As previous works [48, 61] pointed, SCEM is not sensitive to small deviation. Spearman et al. [63] proposed statistical throughput control (STC) to monitor production progress, however, their methods are not applicable when production is in small batches and the inventory diviation features strong autocorrelation. Besides, they provided no discussion on how to apply STC when many SKUs at different facilities need to be monitored simultaneously. Other solutions include monitoring based on radio frequency identification (RFID) technique [14, 54, 61], however, these monitoring solutions are designed for distribution networks and are not applicable when complex productions are involved.

Liu et al. [48] proposed a monitoring strategy based on traditional statistical process control. They compared different control statistics and proposed a control statistic, ship rate, that is both sensitive to small scale disruptions and robust to high dynamic complexity. They deploy multiple univariate control charts, each monitor production of one SKU. They discussed strategy of deploying control charts to bottlenecks and effectively reduce the number of charts used. However, their method focuses on dynamic complexity and the monitoring results depend on the choice of bottlenecks. Besides, when large number of SKUs need to be monitored, this strategy may produce high false alarm rate and searching appropriate training data to estimate control limits is computational expensive.

We propose a multivariate control strategy that is capable of detecting small underproduction in a subset SKU in a SC features detail and dynamic complexities. We first modify ship rate proposed by Liu et al. [48] and propose a simpler technique to efficiently remove skewness in the ship rate statistic. We then examine different visualization techniques and propose a heat map visualization to present information of multiple SKUs over
a period in one chart. The heat map visualization translates SC data such as inventory or ship rate into colors, and compared to traditional visualizations as line and scatter plots [9], heat map provides a visualization that is more compact and easier for trend detection and cross-product comparison.

We then examine different multivariate control charts and propose a multivariate control strategy called Heat Map Contrast (HMC). HMC converts the monitoring problem into a series of classification problems in which we contrast a heat map based on the most recent data against a in-control heat map at each time point. Classification results are monitored for out-of-control behavior. Autocorrelations in the ship rate data are carefully handled in the contrast through a classification algorithm modified to be robust to autocorrelations and avoid potential classification biasness. The classification algorithm is also modified to be robust to large number of noise variables and with feature selection algorithm embedded, deviated SKUs can be identified without requiring fault condition to be pre-specified.

Discussions of deployment strategy for a multi-echelon SC and sensitizing monitor are also provided. We discuss studying historical production data to create SKU aggregations which are more sensitive to certain critical failures and an example of grouping SKUs based on their shared parents is provided. Powered by heat map visualization, SC status and monitoring result can be better appreciated by SC experts. Data from a simulated multi-echelon SC system are used to demonstrate the detectability with a representative data from a semiconductor SC.

The remainder of this study is organized as follows. Section 3.1 provide a brief background on concepts of ship rate, visualization techniques and Real Time Contrast algorithm. Section 3.2 discusses the design of control statistic, multivariate time series contrast algorithm, followed by a discussion of sensitizing monitoring for equipment failure
and material shortage. Section 3.3 conducts two sets of experiments to demonstrate the viability of the strategy with a representative data from a semiconductor SC. Section 3.4 discussese future work and conclusions.

3.1 Background

This section provides a brief background on some methods utilized in our research, topics include multi-echelon SCs, ship rates, monitoring and visualization techniques for multivariate time series.

3.1.1 Complex Supply Chain

An internal SC can be complex (such as internal SC of semi-conductor manufacturer). It has multiple echelons, productions are carried out in multiple manufacturing facilities. Productions and logistics of different SKUs are coordinated by a complex planning module. According to Bozarth et al. [6], it often features both dynamic complexity and detail complexities. Dynamic complexity comes from the effort of aligning production to heterogeneity and dynamics in customer needs. As a result, productions are often in small batches and production schedule for the same SKU can be very different from one period to another. Each SKU builds its own inventory to anticipate a delivery that consumes much of the available stock, and the inventory building curve can be very different from one to another. Detail complexity is driven by the increasing number of final products and supported parts. Companies are eager to create more new SKUs to appeal to a larger, more diverse set of customers [2, 28, 71], and a large number of SKUs are produced, stored and moved between geographically separated facilities.
3.1.2 Ship Rates

Liu et al. [48] defined a shipment quantity as a new batch of SKUs enters finished inventory within a SC site. They studied shipments data in a semiconductor SC and found new shipments arrive frequently. The time interval between two consecutive shipments can be as short as one minute (or less). However, shipment quantity is not recorded periodically, and there can be large gaps without any values. Inspired by the flow of water in a pipe, they considers the rate of shipment arrivals which is referred to as ship rate. With a window of size \( c \) defined, the ship rate at time \( t \) is defined as the integral (sum) the the shipment quantities from time \( t - c \) to \( t \), divided by the window length \( c \). If we denote shipment quantity of SKU \( k \) in a facility at time \( u \) as \( h_k(u) \), its ship rate at time \( t \), \( r_k(t) \), can be calculated from Equation 3.1.

\[
r_k(t) = \frac{\int_{t-c}^{t} h_k(u) du}{c}
\]

(3.1)

A build plan is generated at the beginning of a production period, and indicates the quantities of different SKUs need to be produced in a facility during that production period. Build plan drives \( r_k(t) \). If the build plan of SKU \( k \) denotes \( Q \) units need to be completed during period \([t_1, t_2]\), we can obtain a baseline ship rate for SKU \( k \), \( b_k(t) \), during period \([t_1, t_2]\) as

\[
b_k(t) = \frac{Q}{t_2 - t_1}
\]

(3.2)

If the production is in-control, \( r_k(t) \) should fluctuate around \( b_k(t) \). When underproduction occurs, \( r_k(t) \) starts to deviate from \( b_k(t) \). Compared to statistics such as inventory level, \( r_k(t) \) features less autocorrelation and its normal variation can be differentiated from system adaptation; therefore it is robust SCs with high dynamic complexity.
3.1.3 Visualization of Multivariate Time Series

A compact and informative visualization of data merges human knowledge with modern computational power and garners increasing attention [23, 57, 64]. Build Visualization of status in a complex SC facilitates better understanding of the analytics. SC data, such as ship rate and inventory, are often time series. Visualization of multiple SKUs requires visualization of multivariate time series. We have searched different visualizations for multivariate time series, including line plots, stacked graph plots [9], spiral plots [74]. Peng et al. [55] proposed a MVTSPlot to represent a multivariate time series (Figure 3.1 is a heat map from [55]). Each row represents an individual variable, each column corresponds to a time point, and the value of time series is coded into different colors. MVTSPlot is applicable to high dimensional data, convenient for cross comparison and detecting common trends. Since MVTSPlot is essentially a heat map, in the rest of paper we refer to MVTSPlot as Heat Map.

![Figure 3.1: Visualization of Multivariate Time Series.](image-url)
Multivariate control charts, such as Hotelling $T^2$ chart [66], MEWMA chart [60], MCUSUM chart [16] and $U^2$ chart [58], trigger an alarm when the multivariate random variable deviates from its normal condition. MEWMA is considered to be sensitive to small deviation. When used to monitor a $K$ dimensional variable $X(t)$, $X(t)$ is first smoothed by $	ilde{X}(t) = \lambda X(t) + (1 - \lambda)\tilde{X}(t - 1)$, then the control statistic is calculated in Equation 3.1.4, and $\Sigma$ is the in-control covariance matrix of $\tilde{X}(t)$.

$$T^2(t) = \tilde{X}(t)'\Sigma \tilde{X}(t) \quad (3.3)$$

A multivariate charts requires estimation of the in-control covariance matrix, and to obtain a reasonable estimation, data from periods when the equipment was processing the same $K$ SKUs (no more and no less) are needed. In practice, finding sufficient training data to estimate the covariance matrix is challenging, because SKUs can be sent to different equipments depending on machine availabilities and build plans can vary from period to period. Furthermore, multivariate control charts suffer from the curse of dimensionality. According to Runger et al. [59] and Wang et al. [72], traditional multivariate control techniques are designed for low or medium dimension data and can loose their power to detect signals in individual or small subset of variables quickly when the dimension gets large. Later we illustrate multivariate controls are not effective when large number of SKUs are processed but only a small proportion associate with the failed equipment or materials suffers a mean shift.

Deng et al. [19] proposed a different multivariate monitoring scheme named Real Time Contrast (RTC). To monitor $X(t)$, RTC converts the monitoring problem into a series of classification problem. To see if time $t$ is out of control, RTC takes a small sliding
window and includes the most recent \( w \) data points. We denote data in the window as \( S_w = \{X(t-w+1), \cdots, X(t-1), X(t)\} \) and those prior to \( X(t-w+1) \) as \( S_0 \). In RTC \( S_0 \) grows in size as monitoring proceeds. A class label of “1” is assigned to \( S_w \) and “0” to \( S_0 \). A classifier is then built to differentiate \( S_0 \) from \( S_w \). The idea is that when there is no change, data from \( S_w \) and \( S_0 \) are essentially from the same distribution therefore the classifications are random guesses. When a change is introduced, \( S_w \) (class 1) begins to contains more points from a different distribution which will be discernible by a good classifier.

Although any classifier can be used to perform the classification, Deng et al. [19] preferred Random Forest (RF) because it handles data with complex structure, allows interactions between variables and provides class probability estimation besides classification error. RF builds a parallel ensemble of tree classifier [7, 8]. Because RTC uses a small \( w \), there is always a strong imbalance between class 0 and 1. Stratified sampling is used to adjust for the imbalance, and each tree classifier is built by taking the same number of bootstrap samples from \( S_0 \) and \( S_w \).

Multiple control statistics from RF are discussed and compared in Deng et al. [19], including classification error rate, generalized likelihood ratio, Out-of-Bag (OOB) probability estimation, etc. Among them, OOB probability estimation of class 0, \( \hat{p}_0(t) \), appears to be the most sensitive. In a RF with \( N \) tree classifiers, for each \( X(t) \), some trees are grown with it omitted. The set of data points omitted from a tree are its OOB samples. Let \( OOB(t) \) be the set of trees where \( X(t) \) is an OOB sample, and the OOB probability estimate for \( X(t) \) belonging to class 0 is the proportion of trees in \( OOB(t) \) that classify \( X(t) \) into class 0. If we denote the number of data points in \( S_0 \) at \( t \) as \( |S_0| \), \( \hat{p}_0(t) \) is the average OOB probability estimate of all \( X(t) \) in \( S_0 \) being correctly classified at \( t \). When there is no change in the distribution of \( X(t) \), RF is not suppose to differentiate \( S_0 \) from \( S_w \), and the \( \hat{p}_0(t) \) is expected to be 0.5 which means the classification is close to random guess. When
there is a change occur and $S_w$ starts to contain $X(t)$ from a different distribution, RF will catch the difference and correctly classify $S_0$ to class 0, then $\hat{p}_0(t)$ deviate from 0.5 towards 1.

After an out-of-control signal is detected, a fault diagnostic is needed to identify which variables are responsible for the deviation. Multivariate control chart does not come with fault diagnostics, additional diagnostic steps are needed. Jackson et al. [32] suggested a scheme of using principal component analysis to build low dimensional models and provide an orthogonal decomposition $T^2$ statistics. Manson et al. [50] proposed a decomposition by enumerating all possible combinations of variables which is very computational expensive. Li et al. [46] used a linear Gaussian Bayesian network to capture the causal relationships between variables and reduce the computational complexity. Liu [47] proposed a regression adjustment scheme to perform fault diagnostics. With the presence of complex manufacturing process, dynamic bill of material (BOM), safety stock and continuous system adaptation, the relationships between SKUs and failure are non-linear. To identify the right subset of SKUs from such a environment is also challenging.

Compared to charts like MEWMA, RTC has a clear advantage in fault diagnostic. RF has an embedded ability to select important variables during tree constructions. Fault diagnosis can be handled by scoring the importance of variables to the classifier at the time a signal is generated. The most important variables are considered the key contributors to the signal ([19]). We let $X_k(t)$ be the $k_{th}$ variable in $X(t)$ and its variable importance from the $n^{th}$ tree classifier be $v^n_k$, the variable importance of $X_k(t)$ from all $N$ tree classifiers, $v_k$, is calculated by

$$v_k = \frac{1}{N} \sum_{n=1}^{N} v^n_k$$  \hspace{1cm} (3.4)
3.2 Control System

The strategy of monitoring for small scale disruptions to multiple SKUs in one facility is first discussed, followed by a discussion of sensitizing the detection for equipment failure and material shortage, finally a discussion of how to deploy the monitoring across one SC and present the result is provided.

3.2.1 Monitoring Statistic

According to Liu et al. [48], monitoring ship rate \( r_k(t) \) requires updating control limits every time baseline \( b_k(t) \) changes. It is essentially equivalent with monitoring the deviation in \( r_k(t) \) from its \( b_k(t) \) with a fixed zero center line. Therefore, we denote the proportion deviation in \( r_k(t) \) as

\[
\tilde{s}_k(t) = \frac{r_k(t) - b_k(t)}{b_k(t)}
\] (3.5)

The proportion deviation \( \tilde{s}_k(t) \) is expected to be negatively skewed in many applications. For example, in semiconductor manufacturing, most shipments are small and large shipments are rare [48]. As a consequence, \( s_k(t) \) also features strong positive skewness and monitoring \( s_k(t) \) is insensitive to small deviation in production, especially for underproduction. Liu et al. [48] suggested monitoring the exponentially weighted moving average (EWMA) of \( r_k(t) \) accelerates detection of small overproduction and underproduction. In the light of this, we use a low pass filter (such as EWMA) to smooth \( s_k(t) \) and denote the smoothed \( s_k(t) \) as \( \tilde{s}_k(t) \). However, in many cases, the smoothing can not remove all the skewness, and remaining skewness in \( \tilde{s}_k(t) \) can still reduce model detectability ([48]).

One approach to remove remaining skewness in \( \tilde{s}_k(t) \) is to partition the skewed distribution into multiple zones and assign different zone scores to each zone. This approach is widely used in zone control charts [33]. In a zone control chart, a normally distributed
random variable is partitioned using its standard deviation (Figure 3.2). Using standard
deviation to partition ˜\(s_k(t)\) which is positively skewed would result zone scores that are
not sensitive to small negative deviations. One better zoning method for asymmetric dis-
tributions is to use the percentile. If \(\phi_q\) represents the \(q\%\) percentile of ˜\(s_k(t)\), an percentile
zoning example is illustrated in Figure 3.3 left. This approach results narrower zones for
negative ˜\(s_k(t)\) and wider zones for positive ˜\(s_k(t)\) and zone score obtained are less skewed
compared to ˜\(s_k(t)\). Percentile zoning criteria requires reference distributions and the identi-
fication of these distributions adds computational complexities, especially for large number
of SKUs. We therefore suggest a simple zoning approach called Simple Zones. In simple
zone, to determine zone scores for ˜\(s_k(t)\), we first take some in-control samples of ˜\(s_k(t)\), and
denote the sample minimum and maximum value as ˜\(s_{k,\text{min}}\) and ˜\(s_{k,\text{max}}\). We then partition
˜\(s_{k,\text{min}}\leq ˜s_k(t)\leq 0\) to \(M\) zones using zone width
\[
\tau_k^- = \frac{|\tilde{s}_{k,\text{min}}|}{M} \tag{3.6}
\]
Zone score \(-m\) will be assigned to ˜\(s_k(t)\) in range \((-m\tau_k^-, -(m-1)\tau_k^-)\] and an additional
zones with zone score \(-(M + 1)\) is added to contain future ˜\(s_k(t)\) that is less than ˜\(s_{k,\text{min}}\).
Zones for ˜\(s_k(t) > 0\) are defined in the similar way with zone width
\[
\tau_k^+ = \frac{\tilde{s}_{k,\text{max}}}{M} \tag{3.7}
\]
and zone score $M + 1$ is assigned to $\tilde{s}_k(t)$ that are greater than $\tilde{s}_{k,max}$. Simple Zones significantly reduce the computational complexity and are applicable to complex SC. Figure 3.3 right illustrates an example of using Simple Zones to score.

Figure 3.3: Zone Score from Skewed Distributions. Left: Percentile Zone. Right: Simple Zone.

Applying one single simple zone criteria to all SKUs might be inappropriate, because the distribution of $\tilde{s}_k(t)$ changes with production volume and $s_k(t)$ is more skewed when produced with a low volume [48]. Considering the total number of SKUs processed and their highly dynamic nature, determining one zoning criteria for each SKU is also not practical. A better zoning strategy would be to group SKUs based on their respective $b_k(t)$. With in a group, say group $g$, distributions of $s_k(t)$ for different $b_k(t)$ values are relatively similar. We then assign one pair of zone widths $\{\tau_g^-, \tau_g^+\}$ to group $g$. One example would be to define SKUs with $b_k(t)$ between 20 and 40 (unit per hour) into group $g$. To compute the zone widths, we first compute $\tilde{s}_k(t)$ from the historical in-control production data where $b_k(t)$ values are within the range, then let $\tilde{s}_{g,min}$ and $\tilde{s}_{g,max}$ to be the minimum and maximum of all those historical $\tilde{s}_k(t)$. Values of $\{\tau_g^-, \tau_g^+\}$ are computed from Equation 3.6 and 3.7 where we replace $\tilde{s}_k,min$ and $\tilde{s}_k,max$ with $\tilde{s}_{g,min}$ and $\tilde{s}_{g,max}$. This zone widths pair is
used to compute zone scores for all future SKU whose $b_k(t)$ value falls into the range, and we denote the simple zone score of SKU $k$ as $z_k(t)$. Because $z_k(t)$ is derived from $r_k(t)$, it inherits the merit of being sensitive to deviation in production. Narrower zones are applied for $\tilde{s}_k(t) \leq 0$ to make $z_k(t)$ less skewed compared to $\tilde{s}_k(t)$. More importantly, the technique to remove skewness is simple and applicable to the case of large number of SKUs.

3.2.2 Heat Map Contrast

Although changes in $b_k(t)$ cause changes in distribution of $\tilde{s}(t)$, different zone widths are used for different $b_k(t)$, therefore $z_k(t)$ is not depended on $b_k(t)$. Consequently, we can use RTC to directly monitor the vector of $z_k(t)$, $Z(t) = \{z_1(t), z_2(t), \ldots, z_K(t)\}$ from different build plans. However, there are several challenges. First, in order to be sensitive to small deviation, $s_k(t)$ is first smoothed then converted into $z_k(t)$, and smoothing introduces autocorrelations therefore $z_k(t)$ is autocorrelated. When applying RTC to autocorrelated data such as $z_k(t)$, $\hat{p}_0(t)$ is expected to deviate from 0.5 even when there is no change. That is because each tree classifier in RF is built by contrasting random samples from the reference period $S_0$ with random samples from the target period $S_w$. When data is autocorrelated, $S_w$ contains only data from a very short period (e.g. last several hours) and taking relatively large samples (sample size equals $w$) from such a short period preserves the auto-correlative structure. In contrast, $S_0$ contains data collected from a much larger period of time (e.g. past several months). When taking the same $w$ sample points from $S_0$, samples are more likely to be well-separated in time and, therefore, autocorrelation in the samples is much weaker (Figure 3.4 top). A decision tree can detect the subtle difference in autocorrelation and differentiate $S_0$ from $S_w$ as a result, and when many tree classifiers are ensembled, $\hat{p}_0(t)$ deviate from 0.5 towards 1. We refer the deviation of $\hat{p}_0(t)$ as bias. The presence of bias decreases the noise-to-signal ratio and deteriorate the detection capabilities.
Second, each time a small scale disruption occurs, only a small subset of SKUs are relevant. Although RF is in general a robust and accurate classifier, its performance could be degraded when it is applied to a high dimensional data with a number of “noise” variables ([4]). That is because, each tree classifier in RF only considers a small subset of variables at each split, when most variables are not informative to classification, trees can be built on completely irrelevant variables.

Third, in RTC, $S_0$ includes all data points previous to $S_w$, consequently it continuously increases in size as monitoring proceeds. This cause data maintenance issues and slow down the computation. Besides, if a failure occurs and persists, $S_0$ will start to contain more and more data from the out-of-control period. As a consequence, $\hat{p}_0(t)$ will decrease even through the failure never disappear.

To better monitor the vector of $z_k(t)$ in $K$ SKUs, we proposes a multivariate control approach based on RTC. We name the control approach Heat Map Contrast (HMC) because the control statistic is generated through contrasting two heat maps. We introduce HMC as a general monitoring methods. When using HMC to monitor a vector $X(t)$, we generate $S_w$ in the same way as in RTC, but use a fix reference period $S_0$ and do not let $S_0$ to grow as monitor proceeds (Figure 3.4 bottom).

To make HMC robust to autocorrelated $X(t)$, we modify the construction of each tree classifier in RF. Assume $S_0 = \{X(1), X(2), \ldots, X(t_0)\}$, for a specific tree classifier, instead of taking random samples from entire $S_0$ and assigning them to class 0, we randomly sample a starting time $t'$ from $1 \leq t' \leq t_0 - w + 1$, then select $w$ consecutive data points of $\{X(t'), X(t' + 1), \ldots, X(t' + w - 1)\}$. We then take $w$ random samples from $\{X(t'), X(t' + 1), \ldots, X(t' + w - 1)\}$ and $S_w$ respectively, and use these $2w$ data points to construct the tree classifier. When $S_w$ is in-control, samples from both reference period and target period feature the same autocorrelation and classification becomes a random guess.
The modified tree classifier is named Autocorrelation Corrected Tree (ACT). A parallel ensemble of ACTs is referred as Autocorrelation Corrected Random Forest (ACRF) and \( \hat{p}_0(t) \) from ACRF is expected to be 0.5 when \( X(t) \) is in-control.

\[
\hat{p}_0(t) \quad \text{from ACRF is expected to be 0.5 when } X(t) \text{ is in-control.}
\]

\[
\begin{align*}
S_0 &: \text{grows as } t_c \text{ increases} \\
S_w &: \text{last } w \text{ records}
\end{align*}
\]

\[
\begin{align*}
X_1 \\
\vdots \\
X_K
\end{align*}
\]

\[
\begin{align*}
1 & 2 \\
\vdots & \vdots \\
t_e & t_e - w \\
\text{w samples from } S_0 & \quad \text{w samples from } S_w
\end{align*}
\]

\[
\begin{align*}
X(t'), \\
\cdots, \\
X(t' + w - 1)
\end{align*}
\]

\[
\begin{align*}
\text{CLASS 0 0 0 0 0 0} \\
\text{CLASS 1 1 1 1 1 1}
\end{align*}
\]

Figure 3.4: Comparison Between Regular Tree Classifier and ACT. When ACT Is Used, Data of Class 0 and 1 Feature the Same Autocorrelation and Classification Is Unbiased.

In order to obtain better monitoring results when the number of variables is large and most of them are irrelevant to the change, we further modify the RF construction procedure. In a regular RF, each tree classifier considers only a small subset of variables at
each split, and we refer the probability of being considered for split to as weight. In regular RF, weight is the same for every variable. This equal weight strategy fails when the dimension is high and many noise variables exist. [4] proposed a tree-based ensemble with dynamic soft feature selection. They constructed a sequence of tree learners and minimized the sampling weights for noise variables while maintaining the sampling probabilities for relevant variables dominant. The main idea of their work is to select a small sample of features at every step of the ensemble construction. Following the same logic, instead of building one ACRF and having all variables equally likely to be selected for splitting, we build a sequence of ACRFs and gradually lower the probability of a variable being used for split if the variable is considered to be irrelevant. The later ACRF uses this process to focuses more on a small subset of variables that are informative to classification (Figure 3.5).

The ACRF sequence is built through iterations and superscription $j$ is the iteration index. For the first ACRF, prior knowledge can be used to determine the weights or equal weights can be applied when prior knowledge is not available. We let the weights of each variable in ACRF of iteration $j$ be depended on the variable importance learned in ACRF of iteration $j - 1$. If we let $\gamma_k$ to denoted the weight of $X_k$ at iteration $j + 1$, and $v_k$ to denote the variable importance of $X_k$ learned in iteration $j - 1$, then.

$$\gamma_k = \frac{v_k}{\sum_{r=1}^{K} v_r} \quad (3.8)$$

The formal description of constructing $J$ ACRF is described as below

1. Set $j = 1$ and $\gamma_k = \frac{1}{K}$ for each $X_k$.

2. Build ACRF with the probability of $X_k$ being used for splitting equals $\gamma_k$
3. Extract variable importance $v_k$ for each $X_k$

4. Update $\gamma_k$ by Equation 3.8

5. Set $j = j + 1$ and return to Step 2 if $j < J$

When HMC is applied to monitor $X(t)$, the control statistic is $\hat{p}_0(t)$. A $\hat{p}_0(t)$ being close to 0.5 means ACRF can not differentiate $S_w$ and $S_0$ at time $t$, and the system is in-control. If a $\hat{p}_0(t)$ deviate from 0.5 towards 1 means ACRF thinks $S_w$ and $S_0$ are different at $t$ and a change has occurred by time $t$. A threshold is used to trigger the signal, for example it can be $0.5 + 3\hat{\sigma}_{p_0}$ where $\hat{\sigma}_{p_0}$ is the in-control standard deviation of $\hat{p}_0(t)$. Whenever $\hat{p}_0(t)$ exceeds the threshold, an out-of-control signal is reported.
If a signal is triggered at $t$, we check the variable importances obtained at time $t$ to identify which SKUs are responsible for the out-of-control. A large variable importance means the variable is important for differentiating $S_0$ and $S_w$, therefore is considered to be responsible. We recommend to make the final decision based on $v_k$ obtained from ACRF in iteration $J$ which focuses most on the subset of $X_k$ that are relevant for the change. In Figure 3.6, we illustrate a series of $\hat{p}_0(t)$ and the corresponding heat map of variable importances.

![Graph](image)

Figure 3.6: Monitor $\hat{p}_0(t)$ with a Predefined Threshold. Variables Contribute to the Signal Are Identified in Variable Importance Heat Map.

### 3.2.3 Sensitize Monitor to Critical Failure

Knowing which variables will deviate when certain failure occurs helps create a control statistic that is more sensitive to this failures [58]. Prior knowledge from the bill of material and production schedules can be used to identify SKUs that will deviate together. One can
group SKUs based on the equipments used for processing, and SKUs in one group are all processed by one equipment. Those SKUs compete for the same production resource, their ship rate are negatively correlated and exhibit negative deviations when a failure occurs to that equipment.

Once SKUs are grouped, aggregated ship rate for SKU group \( g \) can be defined as \( r_g(t) = \sum_{k \in g} r_k(t) \). Similarly, we define aggregated baseline ship rate as \( b_g(t) = \sum_{k \in g} b_k(t) \). We can compute the aggregated proportion deviation \( s_g(t), z_g(t) \) in the same way as we do to the individual SKU. Zone score of the SKU aggregation, \( z_g(t) \), is expected to be more sensitive than individual SKU when the corresponding equipment fails and experiments are provided to illustrate it in later chapters.

We suggest monitoring \( z_k(t) \) in every individual SKU and \( z_g(t) \) from every aggregations simultaneously. When anticipated failure occurs (i.e. equipment failure), the aggregation becomes more sensitive and accelerates the detection. In the event of unanticipated failures that affects multiple SKUs in different groups, deviation can still be detected by \( z_k(t) \) of individual SKUs. Including aggregations will slightly increases the dimensionality, but ACRF largely focuses on variables that are responsible for the change, therefore the slight dimensionality increase is not a concern.

Groups of SKUs also help us answer another question: how to order variables in a heat map. To provide a context of fault diagnosis, we would like to have SKUs processed from the same parental SKUs be presented next to each other. When failure occurs to the shared parents, neighboring SKUs are exhibiting the same color change and provide a better visualization for human.
3.2.4 Deploy HMC to entire SC and Result Presentation

A schematic overview highlighting the HMC deployment strategy is shown in Figure 3.7. When productions in multiple facilities need to be monitored simultaneously, we recommend to deploy multiple HMCs. Each HMC monitors zone scores from every individual SKU and SKU aggregations in one facility. For each HMC, at the end of every $c$ hours period ($c$ is a user specified parameter), the most recent production data in its corresponding facility are extracted from the company’s data warehouse and contrasted against the reference data. These computations are carried out in an analytical module. The monitoring results from different HMCs are presented in one graphical user interface (GUI) for SC manager in a control room. In Figure 3.7, underproduction occurs in one facility and the

Figure 3.7: A Schematic Overview of the Deployment Strategy and Results Presentation.

$\hat{p}_0(t)$ from its corresponding HMC climbs above the threshold. SC manager can click the
signaled $\tilde{\rho}_0(t)$ curve to drill down to the problematic facility, check the variable importance
and identify which SKUs in that facility are contributing to the signal.

3.3 Experiments

Rather than assuming distributions for statistics, a simulation of a multi-product, multi-echelon semiconductor SC is built as a data test bed. The parameters in the simulation model are tuned so that simulation output is representative of a semiconductor SC. An advantage of using simulation is that changes can be imposed with exact knowledge of when and where they occur as well as which SKUs are affected. Therefore, performance measures can be calculated accurately.

3.3.1 Simulation Model

In the simulation model (Figure 3.8), one type of wafer box is manufactured in a Fab at a constant rate (one unit per hour). As soon as it is manufactured, the wafer box is delivered to a Material Preparation site ($M_1$); the delivery time follows a normal distribution with mean of 2 hours and standard deviation of 0.2 hours. In $M_1$, wafer boxes are first stored in a receiving inventory, then sent to a cutting tool where it is cut into hundreds of dies belonging to 16 different die types $P_1 - P_{16}$ (Table 3.1). The cutting tool processes one wafer box at a time, and the total number of dies obtained from one wafer box is randomly generated from a discrete uniform distribution with minimum of 1125 and maximum of 1135. When a wafer is cut, 57% of the dies are equally split into $P_1, P_3, P_5, P_6, P_9, P_{11}, P_{15}, P_{16}$ and the rest 43% are equally split into $P_2, P_4, P_7, P_8, P_{10}, P_{12}, P_{13}, P_{14}$. Dies are stored in a finished inventory in $M_1$. A delivery occurs every 12 hours moving all dies in the finished inventory to $A_1$. The transportation time from $M_1$ to $A_1$ follows a normal distribution with mean of 24 hours and standard deviation of 2.4 hours.
In $A_1$, 16 types of dies are further processed into 96 different finished products (SKU1-SKU96), and 16 tools are used in the process, one for each die type. Details of the tool assignments and process time can be found in Table 3.1. SKU1-96 are processed in batches and the batch sizes are independently generated from an exponential distribution with mean five.

Table 3.1: Equipment and Tool Information in the Simulation.

<table>
<thead>
<tr>
<th>Facility</th>
<th>Equipment</th>
<th>Input</th>
<th>Processing time (hour)</th>
<th>Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>cutting tool</td>
<td>wafer box</td>
<td>$N(1, 0.05)$ per wafer box</td>
<td>$P_1 - P_{16}$</td>
</tr>
<tr>
<td>$A_1$</td>
<td>1</td>
<td>P1</td>
<td>0.07 per batch</td>
<td>SKU1-6</td>
</tr>
<tr>
<td>$A_1$</td>
<td>2</td>
<td>P2</td>
<td>0.125 per batch</td>
<td>SKU7-12</td>
</tr>
<tr>
<td>$A_1$</td>
<td>3</td>
<td>P3</td>
<td>0.07 per batch</td>
<td>SKU13-18</td>
</tr>
<tr>
<td>$A_1$</td>
<td>4</td>
<td>P4</td>
<td>0.125 per batch</td>
<td>SKU19-24</td>
</tr>
<tr>
<td>$A_1$</td>
<td>5</td>
<td>P5</td>
<td>0.07 per batch</td>
<td>SKU25-30</td>
</tr>
<tr>
<td>$A_1$</td>
<td>6</td>
<td>P6</td>
<td>0.07 per batch</td>
<td>SKU31-36</td>
</tr>
<tr>
<td>$A_1$</td>
<td>7</td>
<td>P7</td>
<td>0.125 per batch</td>
<td>SKU37-42</td>
</tr>
<tr>
<td>$A_1$</td>
<td>8</td>
<td>P8</td>
<td>0.125 per batch</td>
<td>SKU43-48</td>
</tr>
<tr>
<td>$A_1$</td>
<td>9</td>
<td>P9</td>
<td>0.07 per batch</td>
<td>SKU49-54</td>
</tr>
<tr>
<td>$A_1$</td>
<td>10</td>
<td>P10</td>
<td>0.125 per batch</td>
<td>SKU55-60</td>
</tr>
<tr>
<td>$A_1$</td>
<td>11</td>
<td>P11</td>
<td>0.07 per batch</td>
<td>SKU61-66</td>
</tr>
<tr>
<td>$A_1$</td>
<td>12</td>
<td>P12</td>
<td>0.125 per batch</td>
<td>SKU67-72</td>
</tr>
<tr>
<td>$A_1$</td>
<td>13</td>
<td>P13</td>
<td>0.125 per batch</td>
<td>SKU73-78</td>
</tr>
<tr>
<td>$A_1$</td>
<td>14</td>
<td>P14</td>
<td>0.125 per batch</td>
<td>SKU79-84</td>
</tr>
<tr>
<td>$A_1$</td>
<td>15</td>
<td>P15</td>
<td>0.07 per batch</td>
<td>SKU85-90</td>
</tr>
<tr>
<td>$A_1$</td>
<td>16</td>
<td>P16</td>
<td>0.07 per batch</td>
<td>SKU91-96</td>
</tr>
</tbody>
</table>
We let different SKUs have different production rates and $b(t)$ of different SKUs are listed in Table 3.2. For simplicity, we further let $b(t)$ for each SKU to be constant, and those $b(t)$ are carefully chosen, so that they are expected to be fulfilled if no failures (such as tool failure or material shortage) occur. Similarly, finished SKUs are stored in a finished inventory in $A_1$ and later shipped to customers in bulks. The following experiments are conducted based on shipment data extracted from $A_1$. A shipment record is generated when a batch of finished products arrive at the finished inventory at $A_1$. The shipment

<table>
<thead>
<tr>
<th>SKU ID</th>
<th>$b_k$</th>
<th>SKU ID</th>
<th>$b_k$</th>
<th>SKU ID</th>
<th>$b_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>11</td>
<td>6</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>12</td>
<td>9</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>13</td>
<td>10</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>14</td>
<td>1</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>15</td>
<td>2</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>16</td>
<td>4</td>
<td>22</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
<td>17</td>
<td>17</td>
<td>23</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>37</td>
<td>18</td>
<td>24</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>38</td>
<td>19</td>
<td>25</td>
<td>25</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>39</td>
<td>20</td>
<td>26</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>21</td>
<td>27</td>
<td>27</td>
<td>21</td>
</tr>
<tr>
<td>12</td>
<td>41</td>
<td>22</td>
<td>28</td>
<td>28</td>
<td>22</td>
</tr>
<tr>
<td>13</td>
<td>42</td>
<td>23</td>
<td>29</td>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td>14</td>
<td>43</td>
<td>24</td>
<td>30</td>
<td>30</td>
<td>24</td>
</tr>
<tr>
<td>15</td>
<td>44</td>
<td>25</td>
<td>31</td>
<td>31</td>
<td>25</td>
</tr>
<tr>
<td>16</td>
<td>45</td>
<td>26</td>
<td>32</td>
<td>32</td>
<td>26</td>
</tr>
<tr>
<td>17</td>
<td>46</td>
<td>27</td>
<td>33</td>
<td>33</td>
<td>27</td>
</tr>
<tr>
<td>18</td>
<td>47</td>
<td>28</td>
<td>34</td>
<td>34</td>
<td>28</td>
</tr>
<tr>
<td>19</td>
<td>48</td>
<td>29</td>
<td>35</td>
<td>35</td>
<td>29</td>
</tr>
<tr>
<td>20</td>
<td>49</td>
<td>30</td>
<td>36</td>
<td>36</td>
<td>30</td>
</tr>
<tr>
<td>21</td>
<td>50</td>
<td>31</td>
<td>37</td>
<td>37</td>
<td>31</td>
</tr>
<tr>
<td>22</td>
<td>51</td>
<td>32</td>
<td>38</td>
<td>38</td>
<td>32</td>
</tr>
<tr>
<td>23</td>
<td>52</td>
<td>33</td>
<td>39</td>
<td>39</td>
<td>33</td>
</tr>
<tr>
<td>24</td>
<td>53</td>
<td>34</td>
<td>40</td>
<td>40</td>
<td>34</td>
</tr>
<tr>
<td>25</td>
<td>54</td>
<td>35</td>
<td>41</td>
<td>41</td>
<td>35</td>
</tr>
<tr>
<td>26</td>
<td>55</td>
<td>36</td>
<td>42</td>
<td>42</td>
<td>36</td>
</tr>
<tr>
<td>27</td>
<td>56</td>
<td>37</td>
<td>43</td>
<td>43</td>
<td>37</td>
</tr>
<tr>
<td>28</td>
<td>57</td>
<td>38</td>
<td>44</td>
<td>44</td>
<td>38</td>
</tr>
<tr>
<td>29</td>
<td>58</td>
<td>39</td>
<td>45</td>
<td>45</td>
<td>39</td>
</tr>
<tr>
<td>30</td>
<td>59</td>
<td>40</td>
<td>46</td>
<td>46</td>
<td>40</td>
</tr>
<tr>
<td>31</td>
<td>60</td>
<td>41</td>
<td>47</td>
<td>47</td>
<td>41</td>
</tr>
<tr>
<td>32</td>
<td>61</td>
<td>42</td>
<td>48</td>
<td>48</td>
<td>42</td>
</tr>
</tbody>
</table>

Table 3.2: Baseline Ship Rate $b_k(t)$ for SKU1-SKU96 at $A_1$ in the Simulation Model.
record contains facility name, SKU name, time arrived and shipment quantity. At the end of each simulation run, a data file containing all shipment records at $A_1$ is generated.

3.3.2 Computation of $s_k(t)$, $\tilde{s}_k(t)$ and $z_k(t)$

For each SKU $k$, we first use a window size $c = 8$ hours to compute $r_k(t)$, then use $b_k$ in Table 3.2 in Appendix 3.3.1 to compute its proportion deviation $s_k(t)$. We use a low pass filters of EWMA with $\lambda = 0.4$ to obtain $\tilde{s}_k(t)$. To compute $z_k(t)$, we first define SKU groups based on their respective $b_k$ values, then assign one pair of zone widths $\{\tau^-, \tau^+\}$ to each group and use it to compute $z_k(t)$ for all SKUs in that group.

To determine the groups, we study $\tilde{s}(t)$ distributions from a preliminary study. Liu et al. [48] used the same simulation model and simulated data from SKUs with a wide range of $b(t)$ values. We plot the distributions of $\tilde{s}(t)$ for each unique $b(t)$ values in the preliminary study (with the same $c = 8$ hours and low pass filter). By observing the shapes of those $\tilde{s}(t)$ distributions in the preliminary study, we define four SKU groups as $0 \leq b_k < 10$, $10 \leq b_k < 20$, $20 \leq b_k < 40$ and $40 \leq b_k < \infty$ (in unit/hour), so that the distributions of $\tilde{s}(t)$ from different SKUs in one group are relatively similar.

To determine the values of $\{\tau^-, \tau^+\}$ for a particular group, i.e $20 \leq b_k < 40$, we first collect $\tilde{s}(t)$ value in the preliminary study whose $b(t)$ falls in that range, then denote their minimum and maximum as $\tilde{s}_{\text{min}}$ and $\tilde{s}_{\text{max}}$. We further let $M = 3$ and compute $\{\tau^-, \tau^+\}$ from Equation 3.6 and 3.7. For each SKU $k$, we first find its corresponding group $g$. If $\{\tau_g^-, \tau_g^+\}$ are the zone widths of group $g$, conversion from $\tilde{s}_k(t)$ to $z_k(t)$ can be done according to Table 3.3. Note that changing low pass filter will change the distribution of $\tilde{s}_k(t)$. When a different low pass filter is used, we apply the new filter to the preliminary study data and adjust the category range and zone widths if needed.
Table 3.3: Zone Score Conversion. If SKU $k$ Belongs to Group $g$, Converting $\tilde{s}_k(t)$ to $z_k(t)$ Using the Zone Widths Pair $\{\tau_g^-, \tau_g^+\}$.

<table>
<thead>
<tr>
<th>$z_k(t)$</th>
<th>$\tilde{s}_k(t)$</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$[-1,-3\tau_g^-)$</td>
<td>$[-3\tau_g^-,-2\tau_g^-)$</td>
<td>$[-2\tau_g^-,-\tau_g^-)$</td>
<td>$[-\tau_g^-,0)$</td>
<td>$[0,\tau_g^-)$</td>
<td>$[\tau_g^-,2\tau_g^-)$</td>
<td>$[2\tau_g^-,3\tau_g^-)$</td>
<td>$[3\tau_g^-,\infty)$</td>
</tr>
</tbody>
</table>

3.3.3 SC Data Visualization

This experiment is to compare visualizations of $r_k(t)$, $s_k(t)$ and $z_k(t)$ for the same underproduction case. Using the $b_k$ in Table 3.2, we operate the simulation model for 400 hours (50 time points, each representing an eight hours period) with no failures and let tool 1 at $A_1$ fail from hour 401 to hour 480 (10 time points). As a consequence of the failure, SKU1-6 has 20% decrease in $r_k(t)$ from their respective $b_k$.

We compute $r_k(t)$, $s_k(t)$ and $z_k(t)$ for the 96 SKUs ($c = 8$ hours and low pass filter being EWMA $\lambda = 0.4$), and plot their respective heat maps in Figure 3.9. In the heat map, each row represents a SKU and each column represents a time point. An appropriate visualization would exhibit the same color change for each of row 1-6 after column 51, and no color change for row 7-12. In the heat map of $r_k(t)$, different colors across rows are mainly because the different $b_k$ and no color change corresponding to the underproduction can be found. In the heat map of $s_k(t)$, there are slight color change in row 1-6 after column 51. The change is not clear because of the skewness in $s_k(t)$. Only in the heat map of $z_k(t)$ do we clearly observe negative deviations, color turns from blue to brown in row 1-6 after column 51. Clearly, $z_k(t)$ is a better choice for visualization for underproduction compared to $r_k(t)$ and $s_k(t)$.
Figure 3.9: Heat Map of, $r_k(t)$ (Top), $s_k(t)$ and $z_k(t)$ (Bottom) of the Same Production Data. Only Heat Map of $z_k(t)$ Identifies the Negative Deviations.

3.3.4 Monitoring Underproductions

Two experiments are included in this subsection; the first one compares different monitoring strategies, and the second one tests the detection stability of HMCs.

3.3.4.1 Comparison of Different Control Strategies

This experiment is to compare the detection power of three different monitoring strategies: MEWMA, RTC, and HMC. We first operate the simulation model for 6720 simulation hours (840 data points) with no failure using $b_k$ in Table 3.2 Appendix 3.3.1 and generate a reference set $D_0$. To evaluate the performances of different monitoring strategies, we then design a test simulation. In the test simulation, we use the same $b_k$ (Table 3.2) and first
operate the simulation model for 400 simulation hours (time point 1 - 50) with no failure, then let tool 1 at $A_1$ fail from hour 401 to 672 (time point 51 - 84). Tool 1 failure causes $r_k(t)$ in SKU1-SKU6 to negatively deviate 15% from their respective $b_k$. Data generated from each run of the test simulation is referred to as a test set, and each test set contains 50 “no failure” time points and 34 “with failure” time points. The test simulation is repeated for 500 times to generate 500 test data sets.

To compare the performances of different control strategies, we use a measure called signal-to-noise ratio (SNR). The idea is that, when a multivariate control strategy is applied to a test set, the control statistics computed is expect have different mean values for $t = 1, \cdots, 50$ and $t = 51, \cdots, 84$ (Figure 3.10). For example, if HMC is applied to the vector of $\{z_1(t), z_2(t), \cdots, z_{96}(t)\}$ in a test set, $\hat{p}_0(t)$ is expected to fluctuate around 0 for $t = 1, \cdots, 50$, and shifts towards 1 for $t = 51, \cdots, 84$. The magnitude of the mean shift tells the effectiveness of the $\hat{p}_0(t)$ in differentiating periods with and without failures in that test set.

![Figure 3.10: Computation of Signal-to-noise Ratio.](image)
In the light of this, when a control strategy is applied to the $i^{th}$ test data and control statistic $C(t)$ is computed for $t = 1, \cdots, 84$. We denote the mean and standard deviation of $C(t)$ for $t = 1, \cdots, 50$ as $\hat{\mu}_0$ and $\hat{\sigma}_0$, mean and standard deviation for $t = 51, \cdots, 84$ as $\hat{\mu}_1$ and $\hat{\sigma}_1$. A signal-to-noise ratio of $C(t)$ on $i^{th}$ test set, $SNR_i$, is calculated as

$$SNR_i = \frac{\mu_1 - \mu_0}{\hat{\sigma}_0}$$

and a large $SNR_i$ means a large mean shift and small variance during the no failure period, therefore implies we can well separate period with and without failure by applying this monitoring strategy. For each monitoring strategy, we compute $SNR_i$ for $i = 1, \cdots, 500$.

The first monitoring strategy tested is MEWMA. MEWMA is applied to the $s_k(t)$ vector, $\{s_1(t), s_2(t), \cdots, s_{96}(t)\}$. We first estimate the in-control covariance matrix of the $s_k(t)$ vector from $D_0$, then compute $T^2(t)$ with $\lambda = 0.4$ for each time point $t$ in each test set. For the $i^{th}$ test set, $\mu_0$ and $\sigma_0$ are mean and standard deviation of $T^2(t)$ of $t = 1, \cdots, 50$, and $\mu_1$ and $\sigma_1$ are mean and standard deviation of $T^2(t)$ from $t = 51, \cdots, 84$. We compute $SNR$ value for each test set.

To enhance the performance, we create SKU aggregations based on the tool assignments. Two different aggregations assumptions are considered. The first assumption is that we know correct tool assignments (correct aggregation), and the second assumption is that we are given partially correct tool assignments (incorrect aggregation). Details of SKU aggregations are listed in Table 3.4.

Each SKU aggregation is treated as a new SKU and is handled in the same manner as the individual ones. If we use subscript $g$ to indicate aggregation IDs, aggregation ship rate $r_g(t)$ and baseline ship rate $b_g(t)$ are sums of $r(t)$ and $b(t)$ from all the member SKUs at time $t$. For example, if aggregation $g$ contains SKU1-SKU6, for each $t$, $r_g(t) = \sum_{k=1}^{6} r_k(t)$ and $b_g(t) = \sum_{k=1}^{6} b_k(t)$ where $k$ is the member SKU ID. We test MEWMA with the same
Table 3.4: Different Aggregations for SKU1-SKU96 at $A_1$.

<table>
<thead>
<tr>
<th>Aggregation ID</th>
<th>Correct Aggregations SKU ID</th>
<th>Incorrect Aggregations SKU ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-6</td>
<td>1,3,5,7,9,11</td>
</tr>
<tr>
<td>2</td>
<td>7-12</td>
<td>2,4,6,8,10,12</td>
</tr>
<tr>
<td>3</td>
<td>13-18</td>
<td>13,15,17,19,21,23</td>
</tr>
<tr>
<td>4</td>
<td>19-24</td>
<td>14,16,18,20,22,24</td>
</tr>
<tr>
<td>5</td>
<td>25-30</td>
<td>25,27,29,31,33,35</td>
</tr>
<tr>
<td>6</td>
<td>31-36</td>
<td>26,28,30,32,34,36</td>
</tr>
<tr>
<td>7</td>
<td>37-42</td>
<td>37,39,41,43,45,47</td>
</tr>
<tr>
<td>8</td>
<td>43-48</td>
<td>38,40,42,44,46,48</td>
</tr>
<tr>
<td>9</td>
<td>49-54</td>
<td>49,51,53,55,57,59</td>
</tr>
<tr>
<td>10</td>
<td>55-60</td>
<td>50,52,54,56,58,60</td>
</tr>
<tr>
<td>11</td>
<td>61-66</td>
<td>61,63,65,67,69,71</td>
</tr>
<tr>
<td>12</td>
<td>67-72</td>
<td>62,64,66,68,70,72</td>
</tr>
<tr>
<td>13</td>
<td>73-78</td>
<td>73,75,77,79,81,83</td>
</tr>
<tr>
<td>14</td>
<td>79-84</td>
<td>74,76,78,80,82,84</td>
</tr>
<tr>
<td>15</td>
<td>85-90</td>
<td>85,87,89,91,93,95</td>
</tr>
<tr>
<td>16</td>
<td>91-96</td>
<td>86,88,90,92,94,96</td>
</tr>
</tbody>
</table>

$\lambda$ values under both aggregation assumptions, and apply MEWMA to the vector of $s_k(t)$ from both 96 individual SKUs and 16 aggregations.

We also tested MEWMA with two other $\lambda$ values, $\lambda = 0.2$ and 1, in the same way. When $\lambda = 1$ is used, no smoothing is applied. To better understand variation among individual SKUs, we apply Principal Component Analysis [36] to the in-control covariance matrix of $s_k(t)$ to see if the system variations can be explained by a small number of principal components. We find the first principal component only explained 4% of the total variance and it requires around 30 principal components to explain 70% of the total variance.

The second strategy is RTC. RTC is applied to the $z_k(t)$ vector, \{\, z_1(t), z_2(t), \cdots, z_{96}(t) \}. We compute $z_k(t)$ vector values for $D_0$, and make it the in-control reference $S_0$. For the $i^{th}$ test set, we start from $t = 10$, and at time $t$, we let $S_{iw}$ contains 10 most recent data points ($z_k(t)$ vector from $t-9$ to $t$) and contrast it against 10 data points randomly sampled from $D_0$. We obtain 75 $\hat{\rho}_0(t)$ values for $t = 10, 11, \cdots, 84$ and no $\hat{\rho}_0(t)$ is computed for
When computing $SNR_t$, we let $\hat{\mu}_0$ and $\hat{\sigma}_0$ to be the mean and standard deviation of $\hat{p}_0(t)$ for $t = 11, \ldots, 50$, and $\hat{\mu}_1$ and $\hat{\sigma}_1$ to be those from $\hat{p}_0(t)$ for $t = 51, \ldots, 84$.

The same aggregations are considered to enhance the performance of RTC, and RTC is applied to the vector of $z(t)$ of 96 individual SKUs and 16 aggregations when aggregations are used. We also tested two other low pass filters (EWMA with $\lambda = 0.2$ and 1) in the same way.

The third control strategy tested is HMC. The testing procedures are the same as RTC and the only differences is $\hat{p}_0(t)$ is obtained from ACRF instead of RF. For the HMC, we set $J = 5$ so that for each $t$, five ACRFs are built and the first four ACRFs are built to obtain the weights while $\hat{p}_0(t)$ from the fifth ACRF is used to determine if failures have occurred by that $t$.

All experimental results are listed in Figure 3.11. Each box plot summarizes 500 $SNR$ obtained from applying one control strategy to all test sets under selected $\lambda$ values and aggregation assumption. In Figure 3.11, MEWMA provides the lowest $SNR$. Although correct aggregations are better predictors, including them further increases the dimensionality and does not improve $SNR$ for MEWMA. HMC and RTC provide higher $SNRs$ compared to MEWMA. Compared to RTC, HMC provides greater but more variated $SNRs$. When $\lambda = 0.2$, RTC and HMC provide similar results. When increase $\lambda$ to 0.4 and 1, we observe a larger increase of $SNR$ in HMC than in RTC. Unlike MEWMA, including aggregations helps both HMC and RTC, especially when correct aggregations are included. Compared to RTC, HMC better differentiates out-of-control period from the in-control period.

In Figure 3.12, each line represents a series of $T^2(t)$ for $t = 30, \ldots, 80$ obtained from applying MEWMA to one test data set. A total of 500 series are plotted. We do not observe the change at time point 51. No fault diagnostics are provided in MEWMA.
Figure 3.11: SNR for Different Monitoring Strategies under Two Aggregation Assumptions and Three Low Pass Filters. HMC Provides the largest SNRs.

Figure 3.12: $T^2$ from MEWMA with $\lambda = 0.4$. SNR Is Very Low.

Figure 3.13 plots 500 $\hat{p}_0(t)$ series from applying RTC to each test sets. RTC outperforms MEWMA and mean shifts in $\hat{p}_0(t)$ are observed after time point 50. Furthermore, with even partially correct aggregations included, $\hat{p}_0(t)$ can be improved. Figure 3.14 shows the averaged variable importance over 500 test data sets. The individual deviated SKUs are identified regardless of whether aggregations are used or not. When correct ag-
gregations are used, the aggregation of SKU1-6 is considered to be the most important to the change. When incorrect aggregations are used, two partially correct aggregations, each contains 3 deviated SKUs, are identified. Besides, $\hat{p}_0(t)$ for the in-control period is around 0.7 which is greater than the expected in-control mean of 0.5 ([19]). The deviation is caused by the autocorrelations in $z(t)$ and the reasons are explained in section 3.2.2.

![Figure 3.13: Plot of $\hat{p}_0(t)$ from RTC with $\lambda = 0.4$. RTC Outperforms MEWMA.](image)

![Figure 3.14: Variable Importance Plot from RTC with $\lambda = 0.4$. Deviated SKUs Are Identified.](image)

Figure 3.13 plots 500 $\hat{p}_0(t)$ series from applying HMC to test data sets. HMC also outperformed MEWMA. Compared to RTC, $\hat{p}_0(t)$ obtained from HMC was closer...
to 0.5 when SC is in-control period but features greater variance. HMC focuses more on important variables, $SNR$ improves especially when correct aggregations are used. Figure 3.16 shows the averaged variable importance over 500 test data sets. The results are similar to RTC. The only difference is when correct aggregations are used, aggregation of SKU1-6 dominates, and individual SKUs become less important. As long as the aggregation of SKU1-6 is picked up, we are not concerned.

Figure 3.15: Plot of $\hat{p}_0(t)$ from HMC with $\lambda = 0.4$. Compared to RTC, HTC with $\lambda = 0.4$ provides larger $SNRs$.

Figure 3.16: Variable Importance of HMC with $\lambda = 0.4$. Correct SKU Aggregation is more sensitive than individual SKUs.
Figure 3.17 shows $\hat{p}_0(t)$ series from HMC if no low pass filtering of EWMA is applied ($\lambda = 1$). With no smoothing, $\tilde{s}_k(t)$ is the same as $s_k(t)$. Compared to results where $\lambda = 0.4$ is applied, applying no smoothing to $s_k(t)$ results in much larger variance in $\hat{p}_0(t)$ for the out-of-control period. Therefore, although smoothing introduces certain autocorrelations, it actually removes some skewness and improves the monitoring.

![Plot of $\hat{p}_0(t)$ from HMC with $\lambda = 1$. Smoothing Removes the Skewness and Improves Model Detectability.](image)

3.3.4.2 Detecting Underproduction Scenarios with HMC

In this experiment, we test the detection power of HMC on five other failure cases. Those failure cases cause different numbers of SKUs in $A_1$ to be under-produced with different magnitudes (Table 3.5). We only test one low pass filter EWMA with $\lambda = 0.4$ which delivers the best results from experiment 3.3.4.1. To simulate data for each failure case, we operate the simulation model in the same way as in the test simulation of experiment 3.3.4.1 but with different numbers and severities of tool failures (Table 3.5). For each failure case, we also generate 500 test sets.

We apply HMC in the same way as in experiment 3.3.4.1, and for each failure case, we calculate 500 $\text{SNR}$ and have them summarized into box plots in Figure 3.18. Besides
Table 3.5: Five Failure Cases Are Tested in Experiment 3.3.4.2. Each Failure Case Causes Different Number of SKUs to Deviate with Different Magnitudes.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Failure Case</th>
<th>SKU Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tool 1 fail at simulation hour 401</td>
<td>10% decrease in $r_k(t)$ for $k=1-6$</td>
</tr>
<tr>
<td>2</td>
<td>tool 1 fail at simulation hour 401</td>
<td>20% decrease in $r(t)$ for $k=1-6$</td>
</tr>
<tr>
<td>3</td>
<td>tool 1 and nine fail at simulation hour 401</td>
<td>10% decrease in $r(t)$ for $k=1-6, 48-54$</td>
</tr>
<tr>
<td>4</td>
<td>tool 1 and nine fail at simulation hour 401</td>
<td>15% decrease in $r(t)$ for $k=1-6, 48-54$</td>
</tr>
<tr>
<td>5</td>
<td>tool 1 and nine fail at simulation hour 401</td>
<td>20% decrease in $r(t)$ for $k=1-6, 48-54$</td>
</tr>
</tbody>
</table>

Figure 3.18: Box Plots of $SNR$ from HMC with $\lambda = 0.4$ for Different Decreases.

The five new failure case, we also include failure case in experiment 3.3.4.1 in Figure 3.18 so that we can better study the stability of HMC. The results suggest that including aggregations always improves $SNR$, and correct aggregations help the most. The detectability decreases as less number of SKUs or a smaller magnitude of deviations are involved, which is expected. What we have observed is consistent with results from experiment 3.3.4.1. HMC can detect all failure cases tested. Even with a 10% decrease in six SKUs, HMC still provides reasonable $SNR$.

To conclude experiment 3.3.4.1 and 3.3.4.2, compared to RTC, HMC is designed to better handle autocorrelations in SC and delivers better results. We tested HMC on a relatively large number of SKUs (96 SKUs) and those SKUs are simulated from a wide range of $b(t)$, which illustrate HMC’s viability for monitoring multiple SKUs. If more SKUs
need to be monitored, HMC is expected to be less impacted by the dimensionality increase compared to RTC because the ACRF will decrease the weights for SKUs that are irrelevant to the change. Correct aggregations significantly increase HMC’s detectability, and even partially correct aggregations are helpful. This finding is very important because in a complex production system, SKUs can be sent to different equipment for processing due to reasons such as the equipment maintenance, faster processing speed and etc, therefore it is difficult to always obtain correct aggregations, and partially correct aggregations are common. In the experiment, we set the in-control period in each test simulation to be 400 hours because it allows us to better estimate $\mu_0$. If we increase or decrease the length of in-control period, experimental results will not be changed. We use data from a preliminary study to obtain zone widths, and in real production system, historical production data can be used to serve this purpose.

### 3.4 Conclusion

As the number of SKUs increases, the risk of having both in-control and deviated SKUs in one group also increases. With in-control SKUs included, the deviation in the aggregation is diluted and the mean shift in the aggregation is smaller than that in each individual SKU. In such cases, HMC is still able to capture the deviation because our monitoring approach is designed to be sensitive to small deviations. The experiments result also supports this idea, because when $z_k(t)$ in both individual and incorrect aggregations are used for contrasting, incorrect aggregations are also identified.

In conclusion, a control strategy to monitor multi-echelon SC system with high level dynamic and detail complexities is presented. The scheme is designed to be sensitive to small scale disruption when large number of SKUs are processed. A multivariate control algorithm, Heat Map Contrast, is proposed based on RTC algorithm. HMC converts mon-
itoring problem into a series of classification problems. It is robust to high dimensional SC data and capable of detecting small deviations occurring in only a small subset SKUs. A fault diagnostic function is built inside HMC to identify deviated SKUs every time a signal is detected. To make the monitoring strategy applicable to high dimensional data, zone score is applied to ship rate to remove the skewness. We also provide discussion of sensitizing the detection of certain critical failure and visualization of high dimensional SC data. Beside monitoring, an informative visualization called heat map is proposed for high dimensional SC data. An empirical study using simulated data from the semiconductor industry was conducted to validate the strategy. Different control strategies under various scenarios were tested and results validated the viability of our model.
CHAPTER 4

Proactive Monitor of Demand-side Deviations

Point forecasts of demand are a fundamental component of the planning process in modern supply chains. They are typically generated at regular time intervals and they influence complicated supply chain decisions ranging from procurement to logistics. The accuracy of these forecasts can therefore have a large impact on the operating costs of the entire supply chain. Two components of the forecast influence its accuracy: the point estimate itself, and the dispersion of possible demand relative to the point estimate. The latter is referred to as forecast uncertainty [21] and is the focus of this paper.

Our motivation for studying forecast uncertainty is its close relationship with unknown future forecast errors. Higher levels of forecast uncertainty are associated with greater dispersion in future possible demands, which makes it more likely that a large forecast error will occur in the future. Forecast uncertainty is also directly linked to cost-impacting supply chain decisions such as choosing safety stock levels, since maintaining a constant service level while facing increasing forecast uncertainty requires a corresponding increase in safety stock.

Demand forecast uncertainty is dynamically driven by factors such as market conditions and competitor behavior [43]. In order to respond to these dynamic changes, forecast uncertainty should be monitored continuously so that it can be used to drive appropriate operational changes in the supply chain. For example, an increase in forecast uncertainty for a particular group of products may warrant an increase safety stock levels to reduce the likelihood of experiencing a stockout in the near future. Similarly, cost-saving reductions in safety stock levels may be appropriate when forecast uncertainty has reduced relative to the past.
Existing analytical methods for understanding forecast uncertainty are limited in that they are primarily reactionary to large forecast errors that have already occurred and assume the magnitude of forecast uncertainty is the same regardless of market variability [21, 52]. Other methods in econometrics studies link past forecast errors with future market conditions but require the use of multiple forecasts arising from different sources [5, 17, 20] which are often not available in supply chain forecasting applications.

There is a need for proactive analytical methods which can monitor forecast uncertainty in real-time. Such a method cannot rely solely on realized forecast errors since the objective is to adjust supply chain operational decisions to reflect the possibility of large errors before they occur. Since most practical product forecasting applications involve multiple products, a strategy is needed for monitoring the forecast uncertainty across multiple products simultaneously.

In this paper, we present methods for proactive monitoring of forecast uncertainty when only a single point forecast is available. Our methods are motivated by the rich information available from past forecast updates, which is depicted in Figure 4.1. In this figure, the data available from a single forecasting system with a 3-period horizon is depicted where $t$ represents the current time. At time $t$, we have at our disposal all of the actual realized demands in the previous periods as well as their progression of forecast updates over the 3-period horizon. We also have forecasts for the future demands to occur at periods $t+1, t+2, \ldots, t+H$ where $H$ denotes the length of the forecasting horizon. This information can be used in conjunction with past forecast updates to quantify the uncertainty in forecasts for future time periods. We propose two statistics extracted from the data depicted in Figure 4.1 to achieve this and test their viability using two datasets comprised of real forecasting data. Our forecast uncertainty estimation is completely data driven and requires no domain knowledge of the forecast model. We also discuss how this method can
be extended to monitoring forecasts for multiple products simultaneously and future possible enhancements such as false alarm pruning and extraction of more complex patterns from the forecasting data.

Figure 4.1: Rolling Horizon Forecast with Prediction Horizon Equals to Three.

The remaining of this study is organized as follows. Section 4.1 provides a background review on relevant works. Section 4.2 discusses the details of our forecast uncertainty estimation method. Section 4.3 tests the viability of our method with two real data sets. Section 4.4 discusses potential enhancements and Section 4.5 concludes our contributions.

4.1 Background

This section provides an overview of research relates to our work. We first discuss existing methods for estimating forecast uncertainty, then discuss multivariate control strategies that can be used to monitor forecast uncertainties in multiple SKUs.
4.1.1 Estimation of Forecast Uncertainty

Estimating the future margin of error is itself a forecasting problem [70]. One way to estimate forecast uncertainty is through studying historical forecast errors. Measures of past forecast performance such as Mean Square Forecast Error and Mean Absolute Error are used to estimate future forecast uncertainties [39, 52, 75]. Those estimates omit the dynamic nature of forecast uncertainty and assume the same magnitude regardless of the economic outlook, which can result in large deviations of these estimates from the true forecast uncertainty [15].

More advanced estimating methods are available in economical forecast uncertainty estimation. One estimation assumes various individual forecasters are available such as Survey of Professional Forecasters [21]. They address the question of how the forecast uncertainty can be determined from forecasts provided from individual forecasters. For example, the Federal Reserve uses the interquantile range of forecasts from individual forecasters to quantify the future forecast uncertainty [21]. Other important contributions include [5, 15, 17, 18, 20, 24, 41, 42, 76]. This method is conditional on the economic outlook and better captures the dynamic nature of forecast uncertainty. However, multiple forecasters are needed and for typical product demand forecasting applications, it is more common that a given product has only one point forecast available [11].

Another method for estimating future forecast uncertainty is skewing and rescaling a past forecast performance measure based on the outlook of risk factors. For example, when forecasting global growth, the International Monetary Fund uses financial condition, oil price risk and inflation risk to rescale and skew the past forecast errors and construct asymmetric fan chart [37]. However, this method is not applicable for general demand forecasting situations because when different products over a wide range of niche markets
are being forecasted, the demand for each product will likely be sensitive to different risk factors. A change in one risk factor (e.g. a competitor launches a new product) may have different impact on different products depending on the market in which they compete.

Heath et al. [26] proposed the martingale model of forecast evolution for characterizing forecast changes over time. The model describes the evolution of what the authors refer to as forecast update vectors, and was proposed for the purpose of simulating forecast updates over time. However, their purpose is completely different from ours. They studied how current forecast error observed impacts the forecast updates for future period. They are more interested in understanding the forecaster’s behaviors, such as forecaster of future time tends increase if current demand is underestimated, so that rolling horizon forecasts can be better simulated. They assume forecast uncertainty is constant throughout the process, because their study is restricted to demand from stationary process. The assumption can limit its usage, because many forecasts are non-stationary [77] and forecast uncertainty is also dynamic [15]. Our study is different from theirs, we do not restrict ourselves to stationary demand, and relate the forecast update to forecast uncertainty in the future.

4.1.2 Multivariate Control

Multivariate control charts, such as Hotelling $T^2$ chart [66], MEWMA chart [60], MCUSUM chart [16] and $U^2$ chart [58] are widely used for such multivariate control tasks. However, they can lose their power to detect signals in individual or small subset of variables quickly when the dimension gets large [59, 72] and they do not come with fault diagnostics. Real Time Contrast [19] and Heat Map Contrast convert the monitoring problem into a series of classification problems. They better handle high dimensional data and provide fault diagnostics when a signal is detected. To evaluate if current time $t$ is out-of-control, they both take a small sliding window and includes the most recent $w$ data points, and use a

75
Random Forest model [7, 8] to differentiate them from a in-control reference period. They report a statistic called OOB probability estimation of class 0, \( \hat{p}_0(t) \), to signal if there is a difference between the reference data and recent data. When there is no difference, \( \hat{p}_0(t) \) is expected to be 0.5, while there is a change \( \hat{p}_0(t) \) is expected to deviate from 0.5 towards 1. Fault diagnosis can be handled by scoring the importance of variables to the Random Forest at the time a signal is generated. The most important variables are considered the key contributors to the signal.

Compared to RTC, HMC is designed to better handle data features strong autocorrelation, besides it builds more than one random forest at each time \( t \) to gradually increase the focus on important variables which make it a better choice when the data dimensionality is high and signals are only found in a small fraction of variables.

4.2 Method

In this section, we propose a new way to quantify forecast uncertainties which require no domain knowledge of the market variability or forecast model itself.

4.2.1 Forecast Grid

In many industries such as semiconductor manufacturing, forecasts are generated weekly or monthly over a rolling horizon. For a individual products or stock keeping units (SKUs), forecasts are generated for the next \( H \) periods at each time point. Here each time point can be a week or a month depending on the forecast frequency, and \( H \) is generally refereed as the forecast horizon.

For each \( t \), rich information on both realized demand as well as the outlook for several future time points are available. Figure 4.2 illustrates all the information available at \( t \)
(with \( H = 4 \)). We refer this table as the forecast grid at \( t \) and denote it as \( G_t \). Diagonally, \( \{y_1, y_2, \cdots, y_{t-1}, y_t\} \) indicates the realized demand, and \( \{\hat{y}_{1,2}, \hat{y}_{2,3}, \cdots, \hat{y}_{t-2,t-1}, \hat{y}_{t-1,t}\} \) is the one-period-ahead forecast. We can understand one-period ahead forecast accuracy by contrasting the one period ahead forecasts with the realized demand \( \{y_2 - \hat{y}_{1,2}, y_3 - \hat{y}_{2,3}, \cdots, y_{t-1} - \hat{y}_{t-2,t-1}\} \). We can study \( h \) period ahead forecast accuracy for \( h = 2, \cdots, H \) in the same way.

\[
\begin{array}{cccccccc}
& t - 3 & t - 2 & t - 1 & t & t + 1 & t + 2 & t + 3 & t + 4 \\
\hline
\text{actual} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\hat{y}_{t-4,t-3} & \hat{y}_{t-4,t-2} & \hat{y}_{t-4,t-1} & \hat{y}_{t-4,t} & \hat{y}_{t-3,t} & \hat{y}_{t-3,t+1} & \hat{y}_{t-2,t+2} & \hat{y}_{t-2,t+1} & \hat{y}_{t-2,t} \\
\hat{y}_{t-3,t} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} & \hat{y}_{t-3,t+1} & \hat{y}_{t-2,t+2} & \hat{y}_{t-2,t+1} & \hat{y}_{t-2,t} & \hat{y}_{t-2,t+3} & \hat{y}_{t-2,t+2} \\
\hat{y}_{t-2,t} & \hat{y}_{t-2,t-1} & \hat{y}_{t-1,t} & \hat{y}_{t-1,t+1} & \hat{y}_{t-1,t+2} & \hat{y}_{t-1,t+3} & \hat{y}_{t-1,t+2} & \hat{y}_{t-1,t+3} & \hat{y}_{t-1,t+4} \\
\hat{y}_{t-1,t} & \hat{y}_{t-1,t-1} & \hat{y}_{t} & \hat{y}_{t+1} & \hat{y}_{t+2} & \hat{y}_{t+3} & \hat{y}_{t+2} & \hat{y}_{t+3} & \hat{y}_{t+4} \\
\hat{y}_{t} & \hat{y}_{t-2} & \hat{y}_{t-1} & \hat{y}_{t} & \hat{y}_{t+1} & \hat{y}_{t+2} & \hat{y}_{t+3} & \hat{y}_{t+4} & \hat{y}_{t+5} \\
\hat{y}_{t+1} & \hat{y}_{t+2} & \hat{y}_{t+3} & \hat{y}_{t+4} & \hat{y}_{t+5} & \hat{y}_{t+6} & \hat{y}_{t+6} & \hat{y}_{t+6} & \hat{y}_{t+7} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\hat{y}_{t+H} & \hat{y}_{t+H+1} & \hat{y}_{t+H+2} & \hat{y}_{t+H+3} & \hat{y}_{t+H+4} & \hat{y}_{t+H+5} & \hat{y}_{t+H+6} & \hat{y}_{t+H+7} & \hat{y}_{t+H+8} \\
\end{array}
\]

Figure 4.2: Forecasting Grid \( G_t \): Historical Forecasts and Realized Demand Information at \( t \).

Horizontally, each row contains the vision of future demand across different time periods. For example, \( \{y_t, \hat{y}_{t+1}, \hat{y}_{t+2}, \cdots, \hat{y}_{t+H}\} \) is the vision of demand from \( t + 1 \) to \( t + H \) at \( t \). If we jointly consider realized demands to \( t \), \( \{y_1, y_2, \cdots, y_{t-1}, y_t\} \) and the vision at \( t \), \( \{\hat{y}_{t+1}, \hat{y}_{t+2}, \cdots, \hat{y}_{t+H}\} \), we can better understand the life cycle stage of the SKU. Vertically, each column contains we can the forecast evolutions for a specific time period \( t \). Studying the realized demand at \( t \), we can understand how much forecast uncertainty at \( t \) was resolved through updates. For unrealized \( t \), the forecasts updates also provide important information of the forecast uncertainty at \( t \). Even more information is available if we have \( G_t \) of multiple SKUs, especially SKUs that are competing in the same niche market. Information regarding their competition or if their forecasts are influenced by the same market events can all be studied through mining the forecast grids.
Most companies do not adequately utilize the rich information contained in $G_t$. Instead, they use only the diagonal information corresponding to realized demand for the purpose of analyzing historical demand changes, and one period ahead forecast accuracy to evaluate model adequacy. At each time point $t$, future demands of $t + 1$ to $t + H$ have already been forecasted, and $t + 1$ to $t + H - 1$ have been forecasted for multiple periods. These forecasts contain very useful information regarding the future trend and visions at different time periods and provides useful insight into forecast uncertainty in near future.

4.2.2 Predictive Monitor of Forecast Uncertainty

For a product, its demand during some periods may be inherently more difficult to forecast than others [21]. Difficult periods feature higher forecast uncertainty and forecaster is more likely to make mistakes. These difficult periods often contain changes which can not be well apprehend in advance (e.g. step change, slops change). For example, forecasting demand when there is a new advertisement launch is more difficult because it is hard to anticipate how customers will respond to the advertisement.

In SPF, multiple $\hat{y}_{t-1,t}$ are made by individual forecasters at $t - 1$, the forecast uncertainty of $t$ is identified through the disagreements among those forecasts (forecast dispersion, Plot b in Figure 4.3). When a company has only one forecaster and the forecasting is done over a rolling horizon, $y_t$ is also forecasted for multiple times before it is realized. These forecasts are made by the same forecaster, but with different information. As the forecaster approaches $t$, more information becomes available and the forecast of $y_t$ gets updated (Plot a in Figure 4.3). Similarly, the difference between consecutive forecasts (forecast updates) can be used to understand the forecast uncertainty for future time periods.
Figure 4.3: One Forecaster with Rolling Horizon and Survey of Professional Forecasters.

When a forecaster faces a future change in demand that is not well understood, the initial forecast will likely be less accurate and new information realized later may contradict the forecaster’s previous vision, resulting in large forecast updates. Figure 4.4 (a) illustrates a case that, at $t - 1$, the forecaster anticipated the demand increase at $t + 2$, but information released at $t$ let him believe that his previous judgment regarding the timing of the change was incorrect, the change is postponed till $t + 3$. Figure 4.4 (b) illustrates the forecaster adjusted the size of the change as he was approaching to the change point.

When forecasting a period with a stable $y_t$, or which contains a well-understood change, the initial forecast is accurate and the forecast updates tend to be small because there is little new information released between time $t - 1$ and $t$. Figure 4.4 (c) shows a case of forecasting stable demand. Figure 4.4 (d) shows a case where the demand features a step increase, but the change is well apprehended in advance. The information received at $t$ did not change the forecaster’s view, resulting in no large forecast updates.

To measure the forecast updates, we contrast the forecasts at time $t$ against those at time $t - 1$, and denote the difference as the vision change vector at time period $t$, $V_t$, as
Figure 4.4: Four Different Changes Are Illustrated. (a) (b) Illustrate Not-well-apprehended Demand Changes; (c) Illustrates No Change in Demand, (d) Illustrates a Well-apprehended Demand Change.

\[
V_t = \begin{pmatrix}
\hat{y}_{t,t+1} - \hat{y}_{t-1,t+1} \\
\hat{y}_{t,t+2} - \hat{y}_{t-1,t+2} \\
\vdots \\
\hat{y}_{t,t+H-1} - \hat{y}_{t-1,t+H-1}
\end{pmatrix} = \begin{pmatrix}
u_{t,t+1} \\
u_{t,t+2} \\
\vdots \\
u_{t,t+H-1}
\end{pmatrix} \quad (4.1)
\]

When some \(u_{t,t+h}\) in \(V_t\) are significantly different than zero, the forecaster is anticipating a change to which he does not have sufficient knowledge and therefore the forecast uncertainty should be increased. The vision change considers several future time periods. It is because demand is strongly autocorrelated, and when the forecaster changes his opinion of a future change, he would update forecasts in more than one time period. For example, in
Figure 4.3 (a), at time $t$, the forecaster realized a coming change starting at $t + 1$, he not only increased $\hat{y}_{t,t+1}$ but also increased $\hat{y}_{t,t+2}$ and $\hat{y}_{t,t+3}$.

### 4.2.3 Statistics

The first statistic proposed is $m_t$, which is defined as

$$m_t = \max_{i=1,\ldots,H-1} |u_{t,t+i}|$$  \hspace{1cm} (4.2)

The statistic $m_t$ summarizes $V_t$ into the largest forecast update (in terms of magnitude) it contains, and it does not consider which time point the large forecast update is observed. The change can occur earlier or later than expected, so for $H$ that is relative small ($H \leq 5$) we use $m_t$ to estimate prediction uncertainty at $t + 1, t + 2, \ldots, t + H - 1$. For large $H$, complex feature extraction methods can be used and a discussion is provided in Section 4.4.2. Figure 4.5 illustrates that, at $t - 3$, we summarize $V_{t-3}$ into $m_{t-3}$ (top), then let $m_{t-3}$ be the estimates of forecast uncertainty at $t - 2, t - 1$ and $t$ (bottom). We then repeat the process for $t - 2, t - 1$ and $t$.

For each $t$, we have multiple estimates of its forecast uncertainty. We introduce $\tilde{m}_t$ and $\hat{m}_t$ takes the maximum among all these estimates, and is defined as

$$\hat{m}_t = \max_{i=0,\ldots,H-1} m_{t-i}$$  \hspace{1cm} (4.3)

For example, in Figure 4.5, at time $t - 1$, we have three estimates for the forecast uncertainty at $t$ ($m_{t-3}, m_{t-2}$ and $m_{t-1}$), and $\hat{m}_{t-1}$ is the maximum of these three. Similarly, $\tilde{m}_{t-1}$ is the maximum of $m_{t-2}, m_{t-1}$ and $m_t$.

When using $\hat{m}_t$ to estimate the forecast uncertainty in $t + 1$, we fully consider the uncertainties in the timing of the change. For example, in Figure 4.5, $\hat{m}_t$ is large no only
when large $u_{t-2,t+1}$, $u_{t-1,t+1}$ or $u_{t,t+1}$ are observed. When large forecast updates are observed for time at $t - 2$ or $t - 1$ (large $u_{t-2,t-1}$, $u_{t-2,t}$ or $u_{t-1,t}$), we also have large $\tilde{m}_t$. It is because, although the change was anticipated to occur at time points immediate before $t + 1$, it can be postponed to $t + 1$. Similarly, when a change is expected to occur immediately after $t + 1$, it can occur earlier. Large $u_{t-1,t+2}$, $u_{t,t+2}$ or $u_{t,t+3}$ will also lead to large $\tilde{m}_t$.

The second statistic proposed is $a_t$ and $a_t$ is the Hotelling $T^2$ for $V_t$ which is the most widely used statistics for multivariate control [66]. It is defined as

$$a_t = V_t/\Sigma^{-1}V_t$$

Figure 4.5: Different Ways to Summarize $V_t$. 

when large $u_{t-2,t+1}$, $u_{t-1,t+1}$ or $u_{t,t+1}$ are observed. When large forecast updates are observed for time at $t - 2$ or $t - 1$ (large $u_{t-2,t-1}$, $u_{t-2,t}$ or $u_{t-1,t}$), we also have large $\tilde{m}_t$. It is because, although the change was anticipated to occur at time points immediate before $t + 1$, it can be postponed to $t + 1$. Similarly, when a change is expected to occur immediately after $t + 1$, it can occur earlier. Large $u_{t-1,t+2}$, $u_{t,t+2}$ or $u_{t,t+3}$ will also lead to large $\tilde{m}_t$.

The second statistic proposed is $a_t$ and $a_t$ is the Hotelling $T^2$ for $V_t$ which is the most widely used statistics for multivariate control [66]. It is defined as

$$a_t = V_t/\Sigma^{-1}V_t$$

Figure 4.5: Different Ways to Summarize $V_t$. 

when large $u_{t-2,t+1}$, $u_{t-1,t+1}$ or $u_{t,t+1}$ are observed. When large forecast updates are observed for time at $t - 2$ or $t - 1$ (large $u_{t-2,t-1}$, $u_{t-2,t}$ or $u_{t-1,t}$), we also have large $\tilde{m}_t$. It is because, although the change was anticipated to occur at time points immediate before $t + 1$, it can be postponed to $t + 1$. Similarly, when a change is expected to occur immediately after $t + 1$, it can occur earlier. Large $u_{t-1,t+2}$, $u_{t,t+2}$ or $u_{t,t+3}$ will also lead to large $\tilde{m}_t$.

The second statistic proposed is $a_t$ and $a_t$ is the Hotelling $T^2$ for $V_t$ which is the most widely used statistics for multivariate control [66]. It is defined as

$$a_t = V_t/\Sigma^{-1}V_t$$

Figure 4.5: Different Ways to Summarize $V_t$. 

when large $u_{t-2,t+1}$, $u_{t-1,t+1}$ or $u_{t,t+1}$ are observed. When large forecast updates are observed for time at $t - 2$ or $t - 1$ (large $u_{t-2,t-1}$, $u_{t-2,t}$ or $u_{t-1,t}$), we also have large $\tilde{m}_t$. It is because, although the change was anticipated to occur at time points immediate before $t + 1$, it can be postponed to $t + 1$. Similarly, when a change is expected to occur immediately after $t + 1$, it can occur earlier. Large $u_{t-1,t+2}$, $u_{t,t+2}$ or $u_{t,t+3}$ will also lead to large $\tilde{m}_t$.

The second statistic proposed is $a_t$ and $a_t$ is the Hotelling $T^2$ for $V_t$ which is the most widely used statistics for multivariate control [66]. It is defined as

$$a_t = V_t/\Sigma^{-1}V_t$$

Figure 4.5: Different Ways to Summarize $V_t$. 

when large $u_{t-2,t+1}$, $u_{t-1,t+1}$ or $u_{t,t+1}$ are observed. When large forecast updates are observed for time at $t - 2$ or $t - 1$ (large $u_{t-2,t-1}$, $u_{t-2,t}$ or $u_{t-1,t}$), we also have large $\tilde{m}_t$. It is because, although the change was anticipated to occur at time points immediate before $t + 1$, it can be postponed to $t + 1$. Similarly, when a change is expected to occur immediately after $t + 1$, it can occur earlier. Large $u_{t-1,t+2}$, $u_{t,t+2}$ or $u_{t,t+3}$ will also lead to large $\tilde{m}_t$.

The second statistic proposed is $a_t$ and $a_t$ is the Hotelling $T^2$ for $V_t$ which is the most widely used statistics for multivariate control [66]. It is defined as

$$a_t = V_t/\Sigma^{-1}V_t$$

Figure 4.5: Different Ways to Summarize $V_t$. 

when large $u_{t-2,t+1}$, $u_{t-1,t+1}$ or $u_{t,t+1}$ are observed. When large forecast updates are observed for time at $t - 2$ or $t - 1$ (large $u_{t-2,t-1}$, $u_{t-2,t}$ or $u_{t-1,t}$), we also have large $\tilde{m}_t$. It is because, although the change was anticipated to occur at time points immediate before $t + 1$, it can be postponed to $t + 1$. Similarly, when a change is expected to occur immediately after $t + 1$, it can occur earlier. Large $u_{t-1,t+2}$, $u_{t,t+2}$ or $u_{t,t+3}$ will also lead to large $\tilde{m}_t$.

The second statistic proposed is $a_t$ and $a_t$ is the Hotelling $T^2$ for $V_t$ which is the most widely used statistics for multivariate control [66]. It is defined as

$$a_t = V_t/\Sigma^{-1}V_t$$

Figure 4.5: Different Ways to Summarize $V_t$. 

when large $u_{t-2,t+1}$, $u_{t-1,t+1}$ or $u_{t,t+1}$ are observed. When large forecast updates are observed for time at $t - 2$ or $t - 1$ (large $u_{t-2,t-1}$, $u_{t-2,t}$ or $u_{t-1,t}$), we also have large $\tilde{m}_t$. It is because, although the change was anticipated to occur at time points immediate before $t + 1$, it can be postponed to $t + 1$. Similarly, when a change is expected to occur immediately after $t + 1$, it can occur earlier. Large $u_{t-1,t+2}$, $u_{t,t+2}$ or $u_{t,t+3}$ will also lead to large $\tilde{m}_t$.
where $\Sigma$ is the covariance matrix of $V_t$ computed from a reference period. A reference period is a period where the same forecasting model is used and provide desired forecast accuracy. We then swap $m_t$ in Equation 4.3 and define $\tilde{a}_t$ as

$$\tilde{a}_t = \max_{i=0,\ldots,H-1} a_{t-i}$$

(4.5)

### 4.3 Empirical Results

In this section, two real datasets are used to test the viability of our method.

#### 4.3.1 Nominal GDP Forecast

The first data set we studied is SPF forecasts of US Nominal GDP (NGDP) conducted by Federal Reserve Bank of Philadelphia. At each quarter $t$, Fed of Philadelphia asks multiple forecasters to provide their forecasts of the next five quarters $h$. We denote the forecast made by an individual forecaster at time $t$ as $\{\hat{z}_{t,t+1}, \hat{z}_{t,t+2}, \cdots, \hat{z}_{t,t+5}\}$. Fed of Philadelphia summarizes individual forecasts and publish their own forecasts of NGDP, denoted as $\{\hat{y}_{t,t+1}, \hat{y}_{t,t+2}, \cdots, \hat{y}_{t,t+5}\}$.

For $t$, Fed of Philadelphia collects all of the one-period ahead forecasts $\hat{z}_{t-1,t}$ from the different forecasters and calculates $\hat{z}_{t-1,t}(75\%)$ and $\hat{z}_{t-1,t}(25\%)$, which are the 75 and 25 percentile of those forecasts. The difference between 75 and 25 percentile is then used to quantify the the forecast uncertainty at $t$, which is expressed as

$$s_t = \hat{z}_{t-1,t}(75\%) - \hat{z}_{t-1,t}(25\%)$$

The measure $s_t$ is considered as a common benchmark for the uncertainty of the GDP forecast [43].
Assume we only have the final NGDP forecasts from the Fed (the $\hat{y}_{t,t+h}$) and we compute our statistics discussed previously from the forecast grid for each quarter from 1992 Q1 to 2012 Q4. When $a_t$ is computed, the covariance matrix $\Sigma$ is computed from 1992 Q1 to 1994 Q4. We chose this period as the reference period because the forecast uncertainties during this period is relatively low and stable [43].

We first run an experiment to see which statistic can best describe the variation in $s_{t+1}$. This data is organized in Figure 4.6. Notice that for each row in the figure, we used

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
<th>$x_8$</th>
<th>$x_9$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_1$</td>
<td>$u_{1.2}$</td>
<td>$u_{1.3}$</td>
<td>$u_{1.4}$</td>
<td>$u_{1.5}$</td>
<td>$m_1$</td>
<td>$\tilde{m}_1$</td>
<td>$a_1$</td>
<td>$\tilde{a}_1$</td>
<td>$s_2$</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$u_{2.3}$</td>
<td>$u_{2.4}$</td>
<td>$u_{2.5}$</td>
<td>$u_{2.6}$</td>
<td>$m_2$</td>
<td>$\tilde{m}_2$</td>
<td>$a_2$</td>
<td>$\tilde{a}_2$</td>
<td>$s_3$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$e_{t-2}$</td>
<td>$u_{t-2,t-1}$</td>
<td>$u_{t-2,t}$</td>
<td>$u_{t-2,t+1}$</td>
<td>$u_{t-2,t+2}$</td>
<td>$\tilde{m}_{t-2}$</td>
<td>$\tilde{m}_{t-2}$</td>
<td>$a_{t-2}$</td>
<td>$\tilde{a}_{t-2}$</td>
<td>$s_{t-1}$</td>
</tr>
<tr>
<td>$e_{t-1}$</td>
<td>$u_{t-1,t}$</td>
<td>$u_{t-1,t+1}$</td>
<td>$u_{t-1,t+2}$</td>
<td>$u_{t-1,t+3}$</td>
<td>$m_{t-1}$</td>
<td>$\tilde{m}_{t-1}$</td>
<td>$a_{t-1}$</td>
<td>$\tilde{a}_{t-1}$</td>
<td>$s_t$</td>
</tr>
</tbody>
</table>

Figure 4.6: Training Data of US Nominal GDP for Random Forest.

information of different statistics at $t$ to predict the forecast uncertainty at $t + 1$, $s_{t+1}$. A random Forest [7] is used to identify important predictors that can explain the variations in $s_t$. The random forest model explains 50.2% of the variations in $s_t$, and the variable importance chart is plotted in Figure 4.7. Figure 4.7 indicates that $\tilde{a}_{t-1}$ and $\tilde{m}_{t-1}$ provides the best leading indication of $s_t$. The previous forecast error $e_{t-1}$ is the least important variable. Individual forecast updates ($u_{t-1,t+h}$) are considered more important than $e_{t-1}$, but less important than $\tilde{a}_{t-1}$, $\tilde{m}_{t-1}$, $a_{t-1}$ and $m_{t-1}$.

To better illustrate our results, we let $\hat{q}_t$ to be the estimated forecast uncertainty of GDP at time $t$ and first let $\hat{q}_t = \tilde{m}_{t-1}$ and plot $\hat{q}_t$ and $s_t$ in one chart (Figure 4.8 top), the correlation coefficient for $\hat{q}_t$ and $s_t$ is 0.62. We then let $\hat{q}_t = \tilde{a}_{t-1}$ and repeat (Figure 4.8 middle), the correlation coefficient for this pair is 0.59. Both $\tilde{m}_{t-1}$ and $\tilde{a}_{t-1}$ captures
the forecast uncertainty increase at 2009. For comparison, we also compute the moving
average of mean absolute error (window size $w = 8$), defined as

$$MAE_t = \frac{\sum_{i=0}^{7} |e_{t-i}|}{8}$$

(4.6)

We choose the $w = 8$ because it contains two years of forecast errors. Figure 4.8 bottom
illustrates that if we let $\hat{q}_t = MAE_{t-1}$, the variability in forecast errors is not well explained
and the correlation coefficient is only 0.26, suggesting that the previous forecast error is
not a good indicator of future forecast uncertainty.

4.3.2 Commodity Price Forecasts

We study the World Bank’s historical forecasts of 25 commodities from 2002 to 2013.
We used this time range because forecasts data prior to 2002 was unavailable. The 25
commodities belong to five different categories: Energy, Fat and Oil, Grain, Fertilizer and
Figure 4.8: Forecast Uncertainty Estimation from Three Different Statistics. Top: $\tilde{m}_{t-1}$, Middle: $\tilde{a}_{t-1}$, and Bottom: $MAE_{t-1}$.

Mineral. The World Bank forecasted each commodity twice a year (in January in June) and the actual annual price of the previous year is announced annually in January.

For year $t$, the forecast grid for a particular commodity is defined as in Figure 4.9 and $\hat{y}_{t-0.5,t}$ is defined as the forecast of year $t$ made in June of the previous year (six months
before). We let $e_t = y_t - \hat{y}_{t-0.5,t}$ and because different prediction horizons were used in
\[
\begin{pmatrix}
\cdots & \cdots & \cdots & \cdots \\
0_{t-4} & 0_{t-3} & \hat{y}_{t-3,t-2} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} \\
0_{t-4} & 0_{t-3} & \hat{y}_{t-3,t-2} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} \\
0_{t-4} & 0_{t-3} & \hat{y}_{t-3,t-2} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} \\
0_{t-4} & 0_{t-3} & \hat{y}_{t-3,t-2} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} \\
0_{t-4} & 0_{t-3} & \hat{y}_{t-3,t-2} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} \\
0_{t-4} & 0_{t-3} & \hat{y}_{t-3,t-2} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} \\
0_{t-4} & 0_{t-3} & \hat{y}_{t-3,t-2} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} \\
0_{t-4} & 0_{t-3} & \hat{y}_{t-3,t-2} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} \\
0_{t-4} & 0_{t-3} & \hat{y}_{t-3,t-2} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} \\
0_{t-4} & 0_{t-3} & \hat{y}_{t-3,t-2} & \hat{y}_{t-3,t-1} & \hat{y}_{t-3,t} \\
\end{pmatrix}
\]

Figure 4.9: Raw Data of Commodity Price Forecasts from the World Bank.

different years, we use $H = 3$ since it is the minimum horizon. The vision change of year $t$, $V_t$, is calculated by contrasting the forecasts made at January with those made in June of the same year.
\[
V_t = \begin{pmatrix}
\hat{y}_{t+0.5,t+1} - \hat{y}_{t,t+1} \\
\hat{y}_{t+0.5,t+2} - \hat{y}_{t,t+2} \\
\hat{y}_{t+0.5,t+3} - \hat{y}_{t,t+3}
\end{pmatrix}
\tag{4.7}
\]

Since we only have $11 V_t$ vectors for each commodity and do not have enough data points to estimate the covariance matrix $\Sigma$ of $V_t$, we only compute $\tilde{m}_t$. The World Bank does not conduct SPF, so there is no forecast uncertainty index to benchmark our statistic. Instead, we use $\tilde{m}_{t-1}$ to estimate forecast uncertainty at $t$. That is, stick to the old notation and use $\hat{q}_t$ to denote the estimated forecast uncertainty of a commodity price at time $t$, we then let $\hat{q}_t = \tilde{m}_{t-1}$. We study how $|e_t|$ values are associated with $\hat{q}_t$ values.

Figure 4.10 shows the plots of actual price ($y_t$, black solid line), one period ahead forecasts ($\hat{y}_{t-1,t}$, circled blue dash line), and the estimate forecast uncertainty ($\hat{q}_t = \tilde{m}_{t-1}$, brown dotted line) for each individual commodity. Among all the 25 commodities, 15 (60%) reached their largest $|e_t|$ values when their $\hat{q}_t$ values are at the highest level.
Figure 4.10: Plot of Actual Price (Black), One-period-ahead Forecasts (Blue), and $\hat{q}_t$ (Brown).
For different commodities, \( \hat{q}_t \) and \( e_t \) are different in ranges. To better compare the results obtained from the different commodities, we standardized \( \hat{q}_t \) and \( e_t \). For a particular commodity, we let \( \hat{\mu}_q \) and \( \hat{\sigma}_q \) be the mean and standard deviation of \( \hat{q}_t \) during year 2002 to 2005, and \( \hat{\mu}_e \) and \( \hat{\sigma}_e \) be the mean and standard deviation \( e_t \) during year 2002 to 2005. We choose this 2002 to 2005, because the \( e_t \) for this period are relatively small. We rescale \( e_t \) and \( \hat{q}_t \) for each commodity through

\[
e'_t = \frac{e_t - \hat{\mu}_e}{\hat{\sigma}_e} \quad \text{and} \quad \hat{q}'_t = \frac{\hat{q}_t - \hat{\mu}_q}{\hat{\sigma}_q}
\]

(4.8)

For all 25 commodities, we have 275 \((25 \times 11)\) \( e'_t \) and \( \hat{q}'_t \) pairs. Figure 4.11 shows the scatter plots of all 25 commodities as well as each of the five groups. The horizontal red lines indicate the \( e'_t = -3 \) and 3 and vertical red lines are \( \hat{q}'_t = -3 \) and 3. In scatter plots, having linear relationship between \( e'_t \) and \( \hat{q}'_t \) isn’t our primary concern since we are interested in the co-occurrence of their large values. We expect for time points where \( |\hat{q}'_t| \leq 3 \), they have more \( |e'_t| \leq 3 \), and for time points where \( |\hat{q}'_t| > 3 \), they have more \( |e'_t| > 3 \). Points are less vertically spread out when their \( |\hat{q}'_t| \leq 3 \). We have 137 pairs have both \( |\hat{q}'_t| \leq 3 \) and \( |e'_t| \leq 3 \), and only 13 pairs have \( |\hat{q}'_t| \leq 3 \) and \( |e'_t| > 3 \). For pairs whose \( |\hat{q}'_t| > 3 \), 57 pairs have \( |e'_t| \leq 3 \) and 68 pairs have \( |e'_t| > 3 \). It is hard to see in the plot that contains all commodities (row one, column one in Figure 4.11), because even after the standardization (equation 4.8), the ranges of \( \hat{q}'_t \) and \( e'_t \) for different commodity are very different. For example, the largest \( \hat{q}'_t \) for Phosphate Rock is around 120, and for Soybean, it is around 5. When look at the individual commodity group, such as Fertilizer or Mineral, the difference becomes much clear.

Figure 4.12 shows the number of data points with \( |e'_t| \leq 3 \) (white bar) and \( |e'_t| > 3 \) (gray bar) when their \( \hat{q}'_t \) is set to be no greater than 3 and greater than 3. It is very clear that for \( \hat{q}'_t \leq 3 \), most \( |e'_t| \) values are small (less than 3), and for \( \hat{q}'_t > 3 \), at least half \( |e'_t| \) are
Figure 4.11: Scatter Plots of $e_t'$ and $\hat{q}_t'$. Points Are Less Vertically Spread Out When $|\hat{q}_t'| \leq 3$. 
greater than 3. This is true for commodity groups such as Fat and Oil, Grain, Fertilizer and Minerals. This pattern is also observed if we combine all the commodities.

![Bar charts for different commodities showing counts for $|q'_t| \leq 3$ and $|q'_t| > 3$.](image)

Figure 4.12: Plot of $|e'_t| \leq 3$ (White) and $|e'_t| > 3$ (Gray) for $|q'_t| \leq 3$ and $|q'_t| > 3$.

We notice that large $|e'_t|$ is also observed when $q'_t$ is relatively small. This is because large forecast errors can be caused by “we don’t know that we don’t know” [13, 21]. For example, a change could be attributed to a factor that is not considered important by the forecaster and is therefore not considered in the forecasting model. Uncertainty sources from what forecaster is not aware of are very difficult to model. The $q'_t$ only measures uncertainties from source the forecaster knows. We found that when $q'_t$ is large, we have higher probability of large $|e'_t|$.
4.4 Enhancement

In this section, we first discuss how demand forecast in multiple SKUs should be monitored simultaneously, then discuss extracting complex features with even greater predictive capabilities from the Forecast Grid.

4.4.1 Monitor Vision Change in Multiple SKUs with Heat Map Contrast

A company often offers multiple SKUs spanning several different niche markets, and it is common for a change to only impact some SKUs with respect to their forecast uncertainties. Let the total number of SKUs be \( K \) and \( \tilde{m}^k_t \) be \( \tilde{m}_t \) of SKU \( k \). The goal is to monitor the vector of \( \tilde{m}^k_t \)

\[
\tilde{M}_t = \{\tilde{m}^1_t, \tilde{m}^2_t, \cdots, \tilde{m}^K_t\}
\]

and identify which SKUs are exhibiting large vision change when a signal is detected. Here we use \( \tilde{m}_t \) as an example, but \( \tilde{a}_t \) can be utilized in the same way.

We employ the HMC algorithm for this task since the \( \tilde{M}_t \) contain strong autocorrelation. At each time point \( t \), the control statistic \( \hat{p}_0(t) \) is reported. Normally a threshold is used to trigger a signal, such as \( 0.5 + 3\hat{\sigma}_{p_0} \) where \( \hat{\sigma}_{p_0} \) is the in-control standard deviation of \( \hat{p}_0(t) \). Whenever \( \hat{p}_0(t) \) exceeds the threshold, an out-of-control signal is reported. In Figure 4.13, we illustrate a series of \( \hat{p}_0(t) \) from monitoring \( M_t \) and the corresponding heat map of variable importances.

If a signal is triggered at \( t \), we check the variable importances obtained at time \( t \) to identify which SKUs are responsible for the out-of-control behavior. A large variable importance means the variable is important for differentiating the most recent \( w \) points from the reference period, therefore indicating that the variable is primarily causing the signal.
If SKU $k$ is responsible for the signal, it means the forecaster is experiencing forecast uncertainty that is much larger than what was experienced during the reference period. This suggests that there is a greater chance of having a large forecast error sometime in the near future $(t + 1, t + 2, \ldots, t + H - 1)$. We would therefore recommend increasing the safety stock level of SKU $k$ in future periods and that the forecaster check the adequacy of the forecast model.

4.4.2 Feature Extraction via Deep Learning

The vision change statistic is only one example of a feature with predictive capabilities which can be extracted from previous forecast errors and updates. We believe that the previously presented forecast grid $G_t$ contains a wealth of information that can be used
to predict future forecast errors through both linear and nonlinear combinations of its elements. Features with even greater predictive capabilities than the vision change statistic could potentially be found through such combinations.

One method of extracting features representing nonlinear relationships in $G_t$ is through auto-associative neural networks, which have been proposed as a form of nonlinear principal component analysis (NLPCA) by [40]. It has recently become feasible to train neural networks with deeper architectures than originally proposed for NLPCA through unsupervised pre-training techniques such as contrastive divergence [27]. These deep neural networks are particularly useful for problems with large feature spaces, such as the product- and time-dense information in our forecast grid $G_t$. NLPCA with multiple hidden layers could therefore be an effective method of extracting features characterizing the nonlinear relationships between previous forecast updates and forecast errors.

4.4.3 False Alarm Pruning

Since similar products within a common product family or segment are often affected by similar market forces, false alarms to looming forecast errors could potentially be pruned by considering the state of the forecasting process across other products within a given group. For example, if the objective is to predict only large systematic changes in the market which impact the accuracy of future forecasts, it may be prudent to only raise an alarm when the forecast updates for multiple products within a family or segment each produce alarms. This would allow many false alarms to be ignored while providing greater confidence in the prediction of future forecast errors when multiple products within a group reach a consensus.
4.5 Conclusion

We have presented a new approach for quantifying forecast uncertainty by utilizing past forecast updates and have proposed two statistics that serve as proactive indicators of future forecast errors. The statistics have shown promising results when applied to two datasets consisting of real forecast updates. Unlike existing methods for analyzing forecast uncertainty, our approach allows operational adjustments to be made prior to the occurrence of forecast errors rather than reacting after they have occurred and can be implemented even when only a single point forecast is available for a given product. Our method does not require any domain knowledges of the forecast model and companies can esimate the forecast uncertainty for forecasts purchased from third party institutions.

There is a rich amount of information available in the progression of forecast updates as depicted in Figure 4.1, and an even greater amount of data is available when considering the progression of forecast updates for multiple products simultaneously. In future work, we would like to explore how the forecast updates across products in similar product segments or families can be leveraged for producing better predictive capabilities of future forecast errors and false alarm pruning. We also intend to explore alternative statistics that can be extracted from the forecast update grid by applying nonlinear feature extraction methods such as NLPCA.
CHAPTER 5
Conclusions and Future Work

The growing complexity within a SC increases its vulnerabilities to small deviations in production execution and demand forecasts which can trigger serious SC disruptions. Our research studies the important problem of monitoring complex SCs for deviations that can potentially have a profound impact on the SC operation. Compared to large scale SC disruptions, small scale SC disruptions occur more frequently and are harder to detect. In this study, we identify two major sources of system deviations and propose monitoring strategies that address these events.

The first source that leads to small scale SC disruptions are deviations in plan execution caused by equipment failures or material shortages. In a complex SC, various products are offered, and productions of each individual finished product involve many supporting parts requiring the cooperation of multiple manufacturing sites. A planning system generates build and delivery plans to coordinate materials movements between manufacturing sites, however, even small deviations in executing the production plans can lead to failures in delivery. Considering this interconnectedness, delivery failures in one manufacturing site can further impact its downstream sites and cause serious disruption within the system. Traditional solutions of monitoring plan executions either rely on line operators or line leaders to notice these deviations, or encode predefined deviations as events and use a SCEM to monitor the executions. Monitoring from such a perspective can be both overreacting to normal variations and insensitive to true disruptions that are present in real-world systems.

In this study, we approach the task of monitoring plan execution in two steps. In Chapter 2, we propose a SPC-based monitoring strategy for SCs with only high levels of
dynamic complexity by utilizing a control statistic, ship rate, which is sensitive to small
deviations in plan executions, easy to access in ERP system, and features weak autocor-
relation. To address the skewness caused by small batch production, we use weighted
standard deviation EWMA and weighted standard deviation CUSUM charts to monitor the
ship rates. We provide guidances to automatically update control limits from historical
production data when there are changes in production settings. A chart deploying strategy
is provided to allow charts to be effectively deployed to bottle-necks of the system. This
monitoring solution provides agile detection to small deviations in executing production
plans in complex SC and presents an opportunity for SPC tools to be applied in a new, rich
domain of SC management.

In Chapter 3, we take one step forward and monitor SCs with both high levels
of dynamic complexity and detail complexity. As detail complexity increases, challenges
arise, because the number of SKUs that need monitor in each manufacturing site becomes
large, and equipment failure or material shortage only impact a small subset of these SKUs.
When the number of SKUs are large, deploying multiple univariate charts, one for each
SKU, will result in high false alarm rates. Monitoring multiple SKUs with one multivariate
control chart lose sensitivity, because multivariate control charts are designed for low or
medium dimension data can lose their power to detect signals in individual or small subset
of variables quickly. In this work, we first propose a new method to remove the skewness
in the ship rate and derive a new statistic called a zone score. Compared to the ship rates,
the zone score requires less computation and is more applicable when the number of SKUs
to monitor is large. we present a multivariate control strategy called HMC to monitor zone
scores in multiple SKUs. HMC converts monitoring into a series of classification problems.
It carefully handles autocorrelations in SC data and is designed to be less impacted when the
number of noise variables is large. An embedded fault diagnostic function allows HMC to
accurately identify deviated SKUs every time a signal is detected. We also discuss creating SKU aggregations that are more sensitive to certain failures, and compare HMC with other multivariate monitoring strategies under various failure cases and settings. The results validate the viability of our model, and we provide a visualization for multidimensional SC data which allows for a more compact representation to facilitate trend detection and cross-product comparison.

In Chapter 4, we address the situation when demand deviates from the anticipated range, which is another source of deviations leading to SC disruptions. Build and delivery plans are generated based on the point estimation of future demand, and each demand forecast is associated with some forecast uncertainty which indicates how actual demand can deviate from the point estimation and plays a key role in determining safety stock levels. When actual demands feature higher variation than forecaster’s anticipation, it is more likely to result large forecast errors that can not be handled by safety stocks.

Current methods to estimate forecast uncertainty are primarily reactionary to large forecast errors that have already occurred. Some estimation methods proposed in econometrics link past forecast errors to future market conditions, but they require multiple forecasters or experts’ manual adjustments, which is not applicable to a demand forecast with a large number of SKUs. We present a new way to quantify forecast uncertainties when rolling horizon forecasts are available. We utilize forecast updates and propose two statistics that serve as proactive indicators of future forecast errors. The statistics requires no prior knowledge of the forecasting model itself and have shown promising results when applied to two datasets consisting of real forecast updates. Additionally, we discuss how demand forecast uncertainties in multiple SKUs can be monitored simultaneously with HMC. In this chapter we illustrate that there is a rich amount of information available in the progression of forecast updates which has not be fully studied, and we illustrate one
example of a feature with predicative capabilities which can be extracted from it. This indicates the potential for the extraction of more sophisticated features.

5.1 Future Work

In Chapter 2, we show that grouping SKUs based on processing equipments can accelerates the detections of equipment failures. When an equipment fails, all SKUs processed by that equipment deviate, and the aggregated ship rate becomes a better statistic. Following the same logic, if SKUs that are processed from the same parental SKU are grouped, the aggregated ship rate is expected to be more sensitive to material shortages than individual ship rate.

However, grouping SKUs based on their shared parents becomes more challenging in a semiconductor SC. This is because the BOMs in semiconductor SCs have unique fan-out-fan-in structure. Each die unit (raw material) can potentially be processed into a large number tests units (semi-finished goods), then processed into multiple finished products. For a particular finished product, different parental SKUs can be used at different production periods, and the decision of which parental SKUs are used is dynamically determined by the planning system, that is SKU A can share the same parents with SKU B in one production task but with SKU C in another one. Considering the number of SKUs involved, the many-to-many relation between parental and children SKUs and high dynamics presented in production, an interesting and useful extension of current work is to design an algorithm that is capable of quantifying the likelihoods of SKUs to deviate together when material shortage occurs, automatically creating SKU groups based on these likelihoods and updating grouping when there are changes in production settings.

Also more works can be done with the demand forecast uncertainty estimations. The vision change statistic is only one example of a feature with predicative capabilities
which can be extracted from previous forecast errors and forecast updates. We believe that
the forecast grid discussed in Chapter 4 contains a wealth of information especially when
the prediction horizon is large. Features with even greater predictive capabilities than the
vision change statistic could potentially be found through linear and nonlinear combina-
tions of the elements in the forecast grid. One method of extracting features representing
nonlinear relationships is through deep neural networks. These deep neural networks are
particularly useful for problems with large feature spaces, such as the product- and time-
dense information in our forecast grid.

We also find similar products within a common product family or segment are of-
ten affected by similar market forces, and another extension can be to prune false alarms
by considering the state of the forecasting process across other products within a given
group. For example, if the objective is to predict only large systematic changes in the mar-
et which impact the accuracy of future forecasts, it may be prudent to only raise an alarm
when the forecast updates for multiple products within a family or segment each produce
alarms. This would allow many false alarms to be ignored while providing greater con-
fidence in the prediction of future forecast errors when multiple products within a group
reach a consensus.
REFERENCES


