Editors: **Kang Zhao,** University of Iowa; **Yao Xie,** Georgia Institute of Technology; and **Kwok-Leung Tsui,** City University of Hong Kong

System Informatics: From Methodology to Applications

Kang Zhao, *University of Iowa*
Yao Xie, *Georgia Institute of Technology*
Kwok-Leung Tsui, *City University of Hong Kong*

System informatics analyzes data collected from complex science and engineering systems in different domains, such as manufacturing, energy, logistics, and healthcare.[1–4] In an attempt to improve and optimize the design, performance, and control of these systems, system informatics research leverages methods from various research areas—including statistics, data mining, machine learning, automation and control, and simulation—and offers great opportunities for interdisciplinary collaboration.

With the rapid development of information technologies, the big data collected by ubiquitous sensors has posed new challenges for system informatics research. This installment of Trends & Controversies introduces novel methods and interesting applications of system informatics research in the big data era. We hope the work highlighted here encourages a further dissemination of ideas and collaborative opportunities in this important domain.

On the methodology side, "Projection-Based Process Monitoring and Empirical Divergence" proposes a framework of projection-based methods for real-time online process monitoring by contrasting newly observed data with a reference dataset. "One-Class Classification Methods for Process Monitoring and Diagnosis" discusses how a data analytics algorithm can be used as a control chart for improving process capability through reliable online monitoring and diagnosis.

On the application side, "IoT-Enabled System Informatics for Service Decision Making" reviews current trends and future opportunities for IoT, with a special focus on issues related to the big data collected by multiple sensors. "Quantifying the Risk Level of Functional Chips in DRAM Wafers" not only identifies research challenges and opportunities for decision making with massive data in the process of semiconductor manufacturing, but also quantifies the risk level of functional chips in DRAM wafers. Finally, "Flight Operations Monitoring through Cluster Analysis: A Case Study," describes a new method called cluster-based anomaly detection to help airline safety experts monitor daily flights and detect anomalies.

We thank all the authors for their contributions to this special issue. We also thank *IEEE Intelligent Systems* and its editor in chief, Daniel Zeng, for the opportunity to highlight the state of the art in this emerging area.

References

1. K.-L. Tsui et al., "Recent Research and Developments in Temporal and Spatiotemporal Surveillance for Public Health," *IEEE Trans. Reliability*, vol. 60, no. 1, 2011, pp. 49–58.
2. A. Kusiak and W. Li, "The Prediction and Diagnosis of Wind Turbine Faults," *Renewable Energy*, vol. 36, no. 1, 2011, pp. 16–23.
3. K. Zhao et al., "Analyzing the Resilience of Complex Supply Network Topologies Against Random and Targeted Disruptions," *IEEE Systems J.*, vol. 5, 2011, pp. 28–39.
4. J. Shi and S. Zhou, "Quality Control and Improvement for Multistage Systems: A Survey," *IIE Trans.*, vol. 41, no. 9, 2009, pp. 744–753.

Kang Zhao is an assistant professor in the Department of Management Sciences at the University of Iowa. Contact him at kang-zhao@uiowa.edu.

Yao Xie is an assistant professor in the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. Contact her at yao.xie@isye.gatech.edu.

1541-1672/15/$31.00 © 2015 IEEE
Published by the IEEE Computer Society

Kwok-Leung Tsui is a professor in and head of the Department of Systems Engineering and Engineering Management at City University of Hong Kong. Contact him at kltsui@cityu.edu.hk.

Projection-Based Process Monitoring and Empirical Divergence

Qingming Wei, Wenpo Huang, Wei Jiang, and Yanting Li,
Shanghai Jiao Tong University

Process quality is critical to modern complex systems in manufacturing and service operations. Statistical process control (SPC) is a typical statistical method for maintaining process quality at a satisfactory level. It has been successfully applied in manufacturing system monitoring, healthcare surveillance, and hotspot detection.

Conventional SPC methods are often model-based, and process distributions are assumed known or can be estimated before monitoring. However, these assumptions are very restrictive in many applications. For example, data collected via RFID in fresh-food delivery systems usually contains both continuous data (such as temperature) and attribute data (such as truck ID, driver ID, and so on). Data dimensionality can be quite high—often, much larger than the number of samples. Consequently, it is hard to make reasonable assumptions of process distribution or to estimate parametric models.

Another challenge is that conventional SPC methods often require practitioners to have knowledge of potential process shifts. The monitoring procedure is then designed to be sensitive to certain shifts. Because is is increasingly common to collect data in real time with the help of distributed sensing and high-speed wireless communication technologies, determination of process shift magnitudes or directions becomes a challenging task in real-time monitoring. How to quickly detect a process change, especially when we have little knowledge about the process, is especially difficult.

Here, we propose a general framework of projection-based methods for monitoring process conditions in real time that contrasts newly observed data with reference data. Projection methods can help reduce the dimensionality of multivariate process data; accordingly, our method searches the optimal projection direction to maximize the divergence between the projected reference data and new observations. Because the degree of process deviation from in-control state can be measured with the divergence between in-control distribution and that of newly observed data, we propose using this measure as the charting statistic when monitoring a complex process in real time.

Monitoring a Complex System Based on Divergence

The data from complex system $\mathbf{x} \in R^p$ is assumed to follow a change-point model as follows:

$$\mathbf{x}_i \sim \begin{cases} f_0(\mathbf{x}), & 1 \le i \le \tau \\ f_1(\mathbf{x}), & i > \tau \end{cases},$$

where f_0 and f_1 are the in-control (IC) and out-of-control (OOC) distributions of $\mathbf{x}$, respectively, that are usually unknown in real-time monitoring; τ is the change-point time. In complex systems, multivariate observations $\mathbf{x}_i$ often contain both continuous and categorical variables, such as image pixels, environment conditions, and locations, and the dimension p can be very high.

Real-Time Contrast Procedure

In traditional SPC, a large collection of historical data $S_0 = \{\mathbf{x}_{-n+1}, \ldots, \mathbf{x}_{-1}, \mathbf{x}_0\}$ is usually required to determine the IC condition. IC and OOC data could be mixed or well separated in the historical dataset—if data are mixed, phase 1 methods must be applied to differentiate them. Without loss of generality, we assume that the historical dataset only contains n identically and independently distributed (i.i.d.) IC observations with probability density function $f_0(\mathbf{x})$.

The control limits—that is, the decision boundaries—of conventional control charts such as the support vector data description (SVDD)[1] are determined based on the IC reference dataset, without using information from real-time observations. However, the decision boundary is trained on the most resent observations and should be more sensitive to process shifts: recent observations contain more information about current process conditions. The idea of real-time contrast (RTC) compares the most resent observations with the reference dataset once a piece of new observation arrives. The advantage of this method is that it doesn't require any prespecified knowledge about multivariate process shifts.

In this study, we use only the most recent m observations as representatives of the real-time process condition. Whenever a piece of newly observed data arrives, the oldest one is excluded—that is, a sliding window is imposed on the data stream. Here, observations in the sliding window at time t are denoted as $S_t = \{x_{t-m+1}, \ldots, x_t\}$. In the following, we describe charting statistics based on the RTC between the reference data S_0 and the data in sliding window S_t.

RTC Monitoring Based on the Kullback-Leibler Divergence

SPC aims to detect process changes as soon as possible. When the process

has the parametric form $f(\mathbf{x}, \theta)$, the IC parameter θ_0 and OOC parameters θ_1 are often estimated from process data. Then charting statistics such as cumulative sum (CUSUM) can be built based on the log-likelihood ratio (LLR) – $\log[f(\mathbf{x}, \theta_0) = f(\mathbf{x}, \theta_1)]$. The LLR is usually small when the process is IC and large otherwise.

Use of the LLR method has several prerequisites. First, the underlying process must follow a certain parametric form. Second, the distribution parameters must be known in advance or accurately estimated. Unfortunately, these requirements aren't always met in complex systems. Therefore, monitoring the difference between two contrasting sets S_0 and S_t is a better alternative. The Kullback-Leibler (KL) divergence is commonly used to measure the difference between distributions of two random variables.[2] Given two density functions $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$, the KL divergence is defined by

$$D(f_0 \| f_1) = \int_{\mathbb{R}^p} f_0(\mathbf{x}) \log \frac{f_0(\mathbf{x})}{f_1(\mathbf{x})} \mathbf{dx}.$$

The divergence is zero when two densities come from the same distribution and larger than zero when they're from different distributions. Therefore, monitoring the process change can be converted to monitoring the divergence between the reference data and the sliding window data.

The KL divergence calculation requires estimation of the densities f_0 and f_1. The number of observations required to accurately estimate the process distribution increases exponentially with the process dimension. When the process dimension is high, the number of observations in datasets S_0 and S_t should be large. However, the wide sliding window could deteriorate the monitoring procedure's sensitivity. As an alternative, we can project the reference and sliding window data onto a lower-dimensional space. Assuming the projection direction is ω, the KL divergence between two projected datasets $S_0^\omega = \{\omega^T \mathbf{x}_{-i+1}\}_{i=1}^n$ and $S_t^\omega = \{\omega^T \mathbf{x}_{t-j+1}\}_{j=1}^m$ can be approximated by

$$\hat{d}_t(\omega) = \hat{D}\left(f(\cdot, S_0^\omega) \| f(\cdot, S_t^\omega)\right) = \frac{1}{n} \sum_{\mathbf{x} \in S_0} \log \frac{\hat{f}\left(\omega^T \mathbf{x}; S_0^\omega\right)}{\hat{f}\left(\omega^T \mathbf{x}; S_t^\omega\right)}, \quad (1)$$

where $\hat{f}(\cdot, S_0^\omega)$ and $\hat{f}(\cdot, S_t^\omega)$ are the density estimators of S_0^ω and S_t^ω. The optimal projection direction is the one that best separates the two projected datasets—that is, $\omega_t^* = \arg\max_\omega \hat{d}_t(\omega)$. The maximal divergence $\hat{d}_t(\omega_t^*)$ is then used as the monitoring statistic at time t.

Estimating the Density Function and Calculating the Optimal Projection

Conventionally, histograms estimate the density of univariate projected data. The density function in Equation 1 can be replaced by the proportion of observations falling in each bin. A similar idea can be found in the kth nearest neighbor (kNN)-based approach in the multivariate situation. The difficulties lie in selecting the number and width of bins, which are often subjective. Chances are good that no observations fall in certain bins, which can be problematic when calculating the empirical KL divergence in Equation 1.

The kernel approach is a nonparametric approach widely used to estimate distribution density.[3] For example, the kernel density estimator (KDE) based on dataset S_t^ω is given by

$$\hat{f}\left(y; S_t^w\right) = \frac{1}{m} \sum_{\mathbf{x} \in S_t} K_{H_t}\left(y - \omega^T \mathbf{x}\right), \quad (2)$$

where $K_{H_t} = K(y / H_t) / H_t$ is the scaled kernel function, and H_t is the bandwidth whose selection is based on projected data S_t^ω through a rule-of-thumb approach.[4] We should point out that the kernel-based density estimation has at least two advantages when compared with the histogram or kNN approach: it avoids empty bins in the histogram method, and the KDE in Equation 2 is continuous in the projection direction. Thus, we can efficiently search for the optimal direction via gradient methods. We use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method based on gradient information in this work.

Performance Evaluation

We tested our method by comparing its detection ability on a real dataset from a white wine production process. The dataset contains 4,898 observations that are publicly available in the University of California's (UCI's) "Wine Quality" dataset (http://archive.ics.uci.edu/ml/datasets/Wine+Quality). The data were collected from May 2004 to February 2007, using protected designation of origin samples that were tested at the official certification entity, an interprofessional organization with the goal of improving the quality and marketing of Portuguese Vinho Verde wine.

Each observation had 11 measurements (based on physicochemical tests) including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, PH, sulphates, and alcohol. A categorical variable that indicates wine quality between 0 (very bad) and 10 (very good) was also provided for sensory analysis, the goal of which was to monitor wine quality based on physicochemical tests. A more detailed discussion about this example and dataset appears elsewhere.[5]

Because the distribution of observations is unknown, we compared

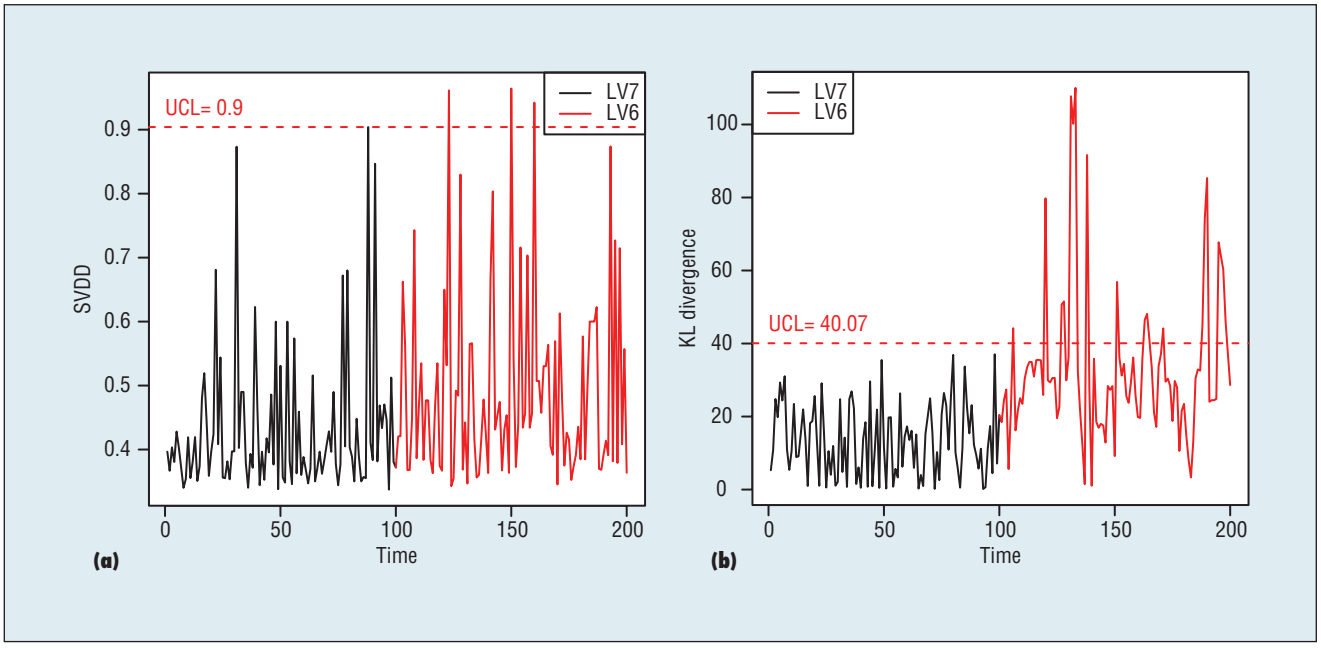

Figure 1. The monitoring statistics of the (a) support vector data description (SVDD) and (b) Kullback-Leibler (KL) divergence methods when observations in S_t come from standard quality level seven (LV7) and LV6. The KL divergence method not only produces a much earlier signal but also produces more signals than the SVDD method when the data change from LV7 to LV6.

our KL divergence method with the SVDD method, which is also distribution-free. Following settings from previous work,[5,6] we chose the index "seven" (LV7) as the standard quality level. To approximate the control limits of both methods, we bootstrapped 5,000 observations from the reference dataset and calculated the kernel distances or KL divergence between the bootstrapped and the reference data for both approaches. The control limit was approximated by the 99.5th percentile, or the 25th largest distance, of all 5,000 distances. Based on our results, the SVDD and KL divergence methods' approximated control limits were 0.9 and 40.07, respectively.

In real-time monitoring, we artificially assumed that 100 observations from LV7 were sequentially followed by the observations categorized as LV6. Figure 1a plots the kernel distances of the SVDD method along with the approximated control limit; Figure 1b shows our proposed method's KL divergences. We can see the KL divergences of LV6 data are substantially larger than LV7 data. The KL divergence method not only produces a much earlier signal but also produces more signals than the SVDD method when the data change from LV7 to LV6.

The proposed approach for more efficient detection of process changes in a complex system involves two steps: find the optimal projection direction that maximizes the projected dataset's KL divergence and then monitor the process through the RTC procedure. Our proposed projection-based approach can easily be extended to detect variance or other process changes via the kernel method, which maps the data to a higher dimensional feature space.

Acknowledgments

We thank Kang Zhao and the referees for their valuable comments that led to improving this article substantially. This work was supported by the National Natural Science Foundation of China (grants 71172131, 71272003, 71325003, 71502106, 71531010), China Postdoctoral Science Foundation (grant 2014M560337), and the Program of Shanghai Subject Chief Scientist (grant 15XD1502000).

References

1. R. Sun and F. Tsung, "A Kernel-Distance-Based Multivariate Control Chart Using Support Vector Methods," *Int'l J. Production Research*, vol. 41, no. 13, 2003, pp. 2975–2989.
2. F. Perez-Cruz, "Kullback-Leibler Divergence Estimation of Continuous Distributions," *Proc. Int'l Symp. Information Theory*, 2008, pp. 1666–1670.
3. S.J. Sheather and M.C. Jones, "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *J. Royal Statistical Soc. Series B (Methodological)*, 1991, pp. 683–690.
4. B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, CRC Press, 1986.
5. P. Cortez et al., "Modeling Wine Preferences by Data Mining from Physicochemical Properties," *Decision Support Systems*, vol. 47, no. 4, 2009, pp. 547–553.

6. C. Zou, W. Jiang, and F. Tsung, "A Lasso-Based Diagnostic Framework for Multivariate Statistical Process Control," *Technometrics*, vol. 53, 2012, pp. 297–309.

Qingming Wei is a PhD candidate in the Antai College of Economics and Management at Shanghai Jiao Tong University. Contact him at weiqm1999@163.com.

Wenpo Huang is a postdoctoral fellow in the Antai College of Economics and Management at Shanghai Jiao Tong University. Contact him at bobhuang09@gmail.com.

Wei Jiang works in the Antai College of Economics and Management at Shanghai Jiao Tong University.

Yanting Li works in the Antai College of Economics and Management at Shanghai Jiao Tong University.

One-Class Classification Methods for Process Monitoring and Diagnosis

Sugon Cho and Seoung Bum Kim,
Korea University

Process monitoring and diagnosis are widely recognized as important techniques for improving quality and detecting abnormal behavior.[1] In manufacturing systems, statistical process control (SPC) methods have improved process capability through reliable online monitoring and diagnosis. An important method in SPC is a *control chart*, which monitors a process's performance over time. Although traditional control charts are effective in situations that involve generating a small volume of independent data, these charts are incapable of handling the large streams of complex data frequently found in modern systems.

Data analytics algorithms can effectively analyze large amounts of data. They use one-class classification (OCC) methods that share a common goal with control charts: both methods assume that the majority class is the only population, and they can both be used to measure the degree of abnormality in new observations. For control chart problems, the number of in-control observations greatly exceeds the number of out-of-control observations—thus, in-control observations are typically used to construct these control charts. OCC methods generate a closed control boundary around a single class of observations of interest and use the boundary to determine whether a future observation belongs to the majority class. Most OCC methods don't require distributional assumptions and effectively accommodate any data format.

As the limitations of traditional control chart techniques become increasingly obvious in the face of more complex systems, OCC methods have the potential to resolve many challenging problems in modern manufacturing and service systems. However, despite this potential, few studies have been conducted to bridge the gap between OCC methods and traditional control charts.

OCC and Control Charts

In general, control charts are constructed in two phases. Phase 1 analysis extracts the in-control data from unknown historical data and uses them to establish the control limits for future monitoring; these limits are then used in Phase 2 analysis to monitor the process. These two phases are analogous to the two phases (training and testing) of classification modeling in data analytics: the training phase uses a training dataset to construct the models that generate a decision boundary, and the testing phase assigns the existing class (category) to an unknown future observation based on the decision boundary determined from the training phase. Based on the availability of class labels in the training dataset, classification models can be divided into either supervised or semisupervised learning. Supervised classification constructs a model by using class labels; semisupervised classification creates a model by using partial information from class labels. OCC, which corresponds to an example of semisupervised learning, uses observations from only one class (primarily the majority class) to construct the decision boundary and uses the decision boundary to determine whether a future observation belongs to the majority class. Generally, for control chart problems, the number of in-control observations greatly exceeds the number of out-of-control observations; thus the majority class is in control.

Table 1 shows the relationship between OCC methods and control charts in terms of their key components. OCC methods calculate a score that quantifies how much an observation deviates from the center of the majority class (the "novelty score"). These scores can be considered the equivalent of monitoring (charting) statistics in control charts. OCC creates a closed decision boundary that encompasses the dataset—this decision boundary resembles the control limit in control charts.

Table 1. Relationship between one-class classification (OCC) and control charts.

Key component	OCC	Control charts
Degree of abnormality	Novelty score	Monitoring statistic
Threshold that determines the abnormality's significance	Decision boundary	Control limit

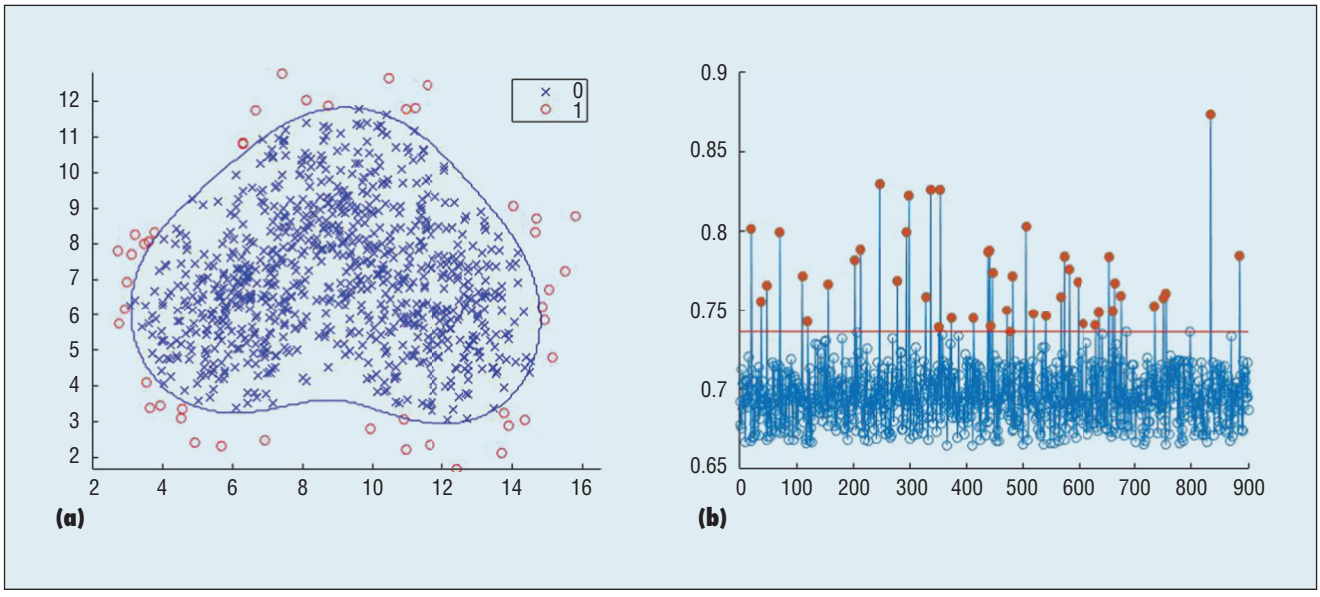

Figure 2. Decision boundary: (a) support vector data description (SVDD) algorithm and (b) the corresponding SVDD-based control chart. The boundary adapts well to the shape of the data.

Recent Developments

A support vector data description (SVDD)[2] is one of the most popular OCC methods. SVDD's goal is to identify a hypersphere that can describe the p-variate training data well, $\{x_i \in \Re^p, i = 1,2,...,n\}$. To achieve this goal, the SVDD algorithm solves the following optimization problem:

$$\min_{R,a,\xi} R^2 + C\sum_{i=1}^{N} \xi_i \text{ subject to}$$
$$\left\|\Phi(x_i) - a\right\|^2 \le R^2 + \xi_i, \text{ for } i = 1,2,...,n, \quad (1)$$

where $\xi_i \ge 0$, $i = 1, 2, \ldots, n$ is a set of slack variables that allows x to be outside the hypersphere. Here, C (> 0) is a regularization parameter that compromises between hypersphere volume and error tolerance. By allowing errors, we can avoid the overfitted hypersphere, and $\Phi(\bullet)$ is a kernel function that maps the original data into a higher dimensional space. The solutions of Equation 1 are the center a and the radius R that characterize the hypersphere; then we can declare observations to be abnormal if $\|\Phi(x) - a\|^2 > R$. Note that this primal formulation of SVDD seems to have nothing in common with the original support vector machine (SVM), but its dual form closely resembles the SVM dual problem.[3] This is why we call it SVDD.

The SVDD algorithm can be converted into a control chart quite easily. SVDD-based control charts' monitoring statistics and the control limit are, respectively, $\|\Phi(x_i) - a\|^2$, $i = 1, 2, \ldots, n$, and R. Figure 2a shows the scatter plot of observations in the 2D space with the boundary obtained from the SVDD algorithm; the boundary adapts well to the shape of the data. Figure 2b shows the resulting control chart representation.

The monitoring statistics that describe the distance from $\Phi(x_i)$ to the center a is straightforward. However, the control limit, R, isn't obvious because R doesn't involve a false alarm rate (that is, type I error rate = α) in its calculation. This is a clear limitation in process monitoring because we can't use the control limit to control the false alarm rate. One study[4] proposed using support vectors to construct the control limits based on the boundary kernel distance. However, any limits constructed this way still couldn't control the false alarm rate. One idea to address this limitation is to use the percentile values of the monitoring statistics estimated by bootstrapping.[5] Another idea is to estimate the distribution of the monitoring statistics using nonparametric estimation methods such as the kernel density estimation (KDE) technique. Although both bootstrapping and KDE deliver control limits that can control false alarm rates, determining the necessary parameters for both methods is complicated and requires a high computational load. Thus, developing an efficient way to determine the control limits in SVDD-based control charts is an open research question.

In addition to SVDD-based control charts, several OCC-based control charts have been developed that use novelty scores as monitoring statistics. One work[6] proposed a hybrid novelty score-based control chart whose monitoring statistics are computed based on the distance to local observations as well as the distance to the convex hull constructed by its neighbors; that study's authors[7] also recently compared eight novelty scores in terms of their control chart performance (such

as average run length). Walid Gani and Mohamed Limam[8] compared the SVDD-based control chart's performance and *k*-nearest neighbor-based control charts proposed elsewhere.[6] Observations in the low-density regions are more likely to be out of control because they're remote from their neighbors. However, the original SVDD doesn't consider data density in constructing its boundary. To address this issue, a density-focused SVDD method that considers both the data's shape and their dense regions has been proposed,[9] as has an improved design of SVDD-based charts.[10]

We hope this article boosts awareness within both the data mining and SPC communities of the relevance of their discipline to an aspect of quality control. We also hope this article stimulates further investigation into the development of better procedures for OCC modeling in system monitoring and diagnosis.

References

1. D.C. Montgomery, *Introduction to Statistical Quality Control*, 6th ed., Wiley, 2009.
2. D.M.J. Tax, "One-Class Classification: Concept-Learning in the Absence of Counter-Examples," PhD dissertation, Delf University of Technology, 2001.
3. D.M.J. Tax and R.P.W. Duin, "Support Vector Data Description," *Machine Learning*, vol. 54, 2004, pp. 45–66.
4. R. Sun and F. Tsung, "A Kernel-Distanced-Based Multivariate Control Chart Using Support Vector Methods," *Int'l J. Production Research*, vol. 41, 2003, pp. 2975–2989.
5. T. Sukchotrat, S.B. Kim, and F. Tsung, "One-Class Classification-Based Control Charts for Multivariate Process Monitoring," *IIE Trans.*, vol. 42, 2010, pp. 107–120.
6. G. Tuerhong et al., "Hybrid Novelty Score-Based Multivariate Control Charts," *Comm. Statistics–Simulations and Computation*, vol. 43, 2014, pp. 115–131.
7. G. Tuerhong and S.B. Kim, "Comparison of Novelty Score-Based Multivariate Control Charts," *Comm. Statistics–Simulations and Computation*, vol. 44, 2015, pp. 1126–1143.
8. W. Gani and M. Limam, "Performance Evaluation of One-Class Classification-Based Control Charts through an Industrial Application," *Quality and Reliability Engineering Int'l*, vol. 29, 2013, pp. 841–854.
9. P. Phaladiganon, S.B. Kim, and V.C.P. Chen, "A Density-Focused Support Vector Data Description Method," *Quality and Reliability Engineering Int'l*, vol. 30, 2014, pp. 879–890.
10. X. Ning and F. Tsung, "Improved Design of Kernel Distance-Based Charts Using Support Vector Methods," *IIE Trans.*, vol. 45, 2013, pp. 464–476.

Sugon Cho is a PhD candidate in the School of Industrial Management Engineering at Korea University. Contact him at sugoncho@gmail.com.

Seoung Bum Kim is a professor in the School of Industrial Management Engineering at Korea University. Contact him at sbkim1@korea.ac.kr.

IoT-Enabled System Informatics for Service Decision Making

Kaibo Liu, *University of Wisconsin-Madison*
Jianjun Shi, *Georgia Institute of Technology*

The Internet of Things (IoT) refers to the interconnection of embedded computing devices within the Internet infrastructure.[1] IoT has created a data-rich environment in which multiple sensors continuously monitor a unit's health status, and multiple units simultaneously transfer these data through the communication network to the processing center for analysis. This has provided an unprecedented opportunity for improving service decision making, which could lead to closer monitoring of a unit's health status, quicker fault diagnosis, more accurate forecasts of a unit's remaining lifetime, and proactive maintenance and control decisions that are better aligned to a unit's future conditions and performance.

However, the existing literature is limited to satisfying the unique needs and challenges of IoT-enabled aftersales service and support. Advanced system informatics methodologies that leverage diverse "gene" pool information could move us from being data-rich to service decision-smart. As an initial effort, we present concepts, current trends and achievements, and research challenges and future opportunities regarding this topic.

Trends and Achievements

A unit's service life cycle refers to the time period during which the unit is in service (from the beginning to the end of its operational life). We can classify the research that analyzes the service life cycle for decision making into two broad categories. *Reliability-based analysis* often treats failure as a random process and employs time-based parametric distributions to model uncertainty in failure times for a population. A *condition-based monitoring and maintenance strategy* considers the degradation evolution across the life cycle for each individual unit.

While the existing literature is rich, one common limitation is that these

methods focus on analyzing single degradation signals. Thus, they are effective only under the assumption that single sensor data can capture the underlying degradation mechanism. Before IoT, this assumption was likely to be satisfied: practitioners could manually choose the most appropriate sensor data or key performance indicator (KPI) to monitor based on empirical or domain knowledge (such as using vibration signals to monitor a rotation bearing). However, as engineering systems become more complex and multiple sensors simultaneously monitor the same unit, existing methods fail to address these new challenges.

Because different sensor data often contain partial and correlated information about the same unit, data fusion methods attempt to both address multiple sensors and improve analytics results. Specifically, data fusion methods fit into two broad categories based on the fusion technique's implementation level at the data or decision level. For service decision making, the existing literature heavily relies on the decision-level fusion approach, which integrates the results (such as voting) created by separately analyzing each individual sensor data. However, such an approach ignores the dependency in multiple sensor data and treats each sensor signal as equally important, thus it often leads to biased results. Moreover, decision-level fusion requires repeated computations based on each individual sensor's data, which is not scalable to IoT applications.

Currently, a more effective and recognized trend is to consider data-level fusion that directly combines multiple sensor data or extracted features to construct a health index (HI) for accurately characterizing a unit's health condition. Such a 1D HI is essential in the after-sales service and support industry because it provides a fundamental understanding of how a unit's health status progresses over time. Furthermore, the HI facilitates data visualization and health comparisons in a dashboard, providing a scientific way to support maintenance scheduling, sparse part logistics, and inventory control based on multiple units' health status in real time. Specifically, let $\boldsymbol{x}_{(i,.,t)} = [x_{(i,1,t)} \ldots, x_{(i,s,t)}] \in R^{(1\times s)}$ be the s sensor data for unit i at observation time t. Then, HI $h_{(i,t)}$ can be constructed via a fusion model, $h_{(i,t)} = f(\boldsymbol{x}_{(i,.,t)})$.

Researchers have made various efforts along this direction in the literature, although not always explicitly using the term HI. For example, multivariate statistical process control (SPC) methods combine multiple sensor data into an HI (called *monitoring statistics* in the SPC field) for change detection and fault diagnosis. However, these methods cannot effectively present an HI that continuously and accurately measures a unit's health status across the life cycle. Although few purely data-driven methods attempt to tackle this issue, they often act like a black box that only provides a final predicted result without any insights into the underlying degradation process, meaning they are not suitable for service decision making, either. In practice, the HI is constructed by combining sensor data (or KPIs) based on simplified physical laws and empirical knowledge. However, such approaches rely on special personnel skills and many years' experience, and they are limited to systems with simple structures and few sensor data.

In theory, the main challenge lies in the fact that HI $h_{(i,t)}$ is not observable, so the problem substantially differs from conventional regression-based issues (that is, assuming the response variable is known). Consequently, in isolation, neither conventional physical-based modeling nor purely data-driven empirical techniques can effectively address the challenge. Developing advanced system informatics that seamlessly integrates domain knowledge and data analytics models is quite essential.

Since the first introduction of system informatics, researchers have made some initial achievements tailored to the needs of service decision making. One work,[2] for example, considered the degradation process's inherent characteristics and identified two essential properties that the HI should possess for successful service decision making: maximizing the HI's monotonic property over the service life cycle and minimizing variance in the failure threshold. An optimization formulation that optimizes these two properties was developed to construct a composite HI via a weighted average of multiple sensor data. Considering that the constructed HI might not be suitable for the selected degradation model, researchers[3] further proposed a semiparametric data fusion model that aims to minimize degradation model uncertainty while constructing the HI. In this way, degradation modeling and the fusion procedure were solved in a unified manner. Case studies performed on aircraft gas turbine engines showed that the developed HI provides a much better characterization of a unit's condition than any original sensor data, leading to superior prediction for the remaining lifetime.

Research Challenges and Future Opportunities

Thanks to IoT technology, historical offline records of a large number of units' service life cycles have become available. In addition, multiple sensor data from in-service units can

be collected in real time, providing a tremendous opportunity to make optimal service decisions. Although system informatics topics have been investigated in several research areas, including multistage manufacturing modeling, sensor allocation and network design, and SPC, few efforts have focused on service decision making, especially in exploring IoT-enabled opportunities and challenges. Recently, initial achievements in the HI approach have demonstrated a promising path.

One unique advantage is that the constructed HI can be regarded as another point of single-sensor data, something to readily integrate with the existing literature on service decision making. However, the current state of the art is limited to linear fusion models and assumes that the unit resulted from one single failure mode under a single environmental condition. Apparently, practical scenarios are much more complex and dynamic, leading us to believe that the ultimate solution relies on a holistic, system-level data fusion and analytics approach through a seamless integration of theories, tools, and techniques from multiple disciplines such as engineering domain knowledge, statistics, data mining, operation research, and decision control. Indeed, this has been an ever-growing trend in many other research fields as well.[4] In particular, we believe the following challenges and problems will be investigated soon, and corresponding research efforts will grow significantly in the near future.

The term *big data* refers to the issues that result when dealing with high volumes of, high velocity in, and a high variety of information. Among the three characteristics, high variety presents a central challenge for service decision making because each sensor's data can convey a different message when used to assess, diagnose, and predict a unit's condition due to different data characteristics, signal-to-noise ratios, and complex dependent relationships. On the other hand, because each sensor's data can exhibit different measurement units and signal scales, they also impose a severe challenge to the development of system informatics techniques. How to overcome this big data challenge by effectively leveraging the diverse "gene" pool created by multiple sensors is a question that remains to be resolved.

The second challenging issue is how to simultaneously deal with multiple sensor data and multiple failure modes. In practice, different failure modes can have distinct influences on a unit's lifecycle path. Therefore, an essential task here is to continuously and accurately estimate failure along the life cycle. The central challenge lies in that some sensors might be sensitive to certain types of failure modes by showing a strong degradation pattern, and other sensors might not. Although many data fusion studies have been done for online fault diagnosis, they either employ a simple voting scheme that combines results created by separately analyzing each individual sensor's data or they fail to address the specific needs of service decision making. Particularly, two unique requirements must be satisfied when developing the system informatics technique: the fault diagnosis must be estimated and updated continuously along the life cycle, and the diagnostic result must be more accurate as the unit approaches its end of life, to ensure better service planning and avoid unexpected failure. How to tackle these unique challenges isn't well addressed in the existing literature.

Multiple and time-varying environmental conditions have been a challenge for service decision making, with environmental conditions having a big impact on a unit's service life cycle, which in turn influences the decision-making process. As shown in several applications, units are often operated under complex and dynamic conditions, some of which might not even be known or predictable. Some studies consider the effects of environmental conditions based on a single sensor's data, but there's still a lack of effective system informatics methods that simultaneously overcome the challenges from both multiple sensors and multiple conditions.

Finally, the ultimate goal of after-sales service and support is to improve customer satisfaction, overall cost, productivity, and efficiency at the enterprise level, all of which require making an effective maintenance and control decision that leverages health assessments and predictions from multiple units in real time. Existing methods often optimize service decisions based on an individual unit's condition, and thus they can lead to a loss of overall throughput, efficiency, and cost without considering the enterprise's resource constraints and availability. How to solve this new challenge is still lacking in the current literature.

Almost 25 billion devices are connected worldwide, with IoT pushing the fourth industrial revolution. A recent report from the McKinsey Global Institute stated that IoT could unleash up to $6.2 trillion in new global economic value annually by 2025, and that 80 to 100 percent of manufacturers will be using IoT applications by then.[5] In particular, after-sales service and support is likely to generate new revenue greater than $300 billion by 2020.

With data availability and computational power reaching an unprecedented level in recent years, the question of how to exploit these emerging opportunities to improve service decision making will take the center stage of engineering research in the near future. We believe that system informatics research concentration will play an important role and will stimulate numerous opportunities for both academia and industry.

References

1. J. Holler et al., *From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence*, Academic Press, 2014.
2. K. Liu, N. Gebraeel, and J. Shi, "A Data-Level Fusion Model for Developing Composite Health Indices for Degradation Modeling and Prognostic Analysis," *IEEE Trans. Automation Science and Eng.*, vol. 10, 2013, pp. 652–664.
3. K. Liu and S. Huang, "Integration of Data Fusion Methodology with Degradation Modeling Process to Improve Prognostics," to be published in *IEEE Trans. Automation Science and Eng.*, 2015.
4. J. Shi, "Data Fusion for In-Process Quality Improvement," *Proc. 5th Manufacturing Eng. Soc. Int'l Conf.*, 2013; http://mesic2013.unizar.es/proceedings/documents/259.pdf.
5. J. Manyika et al., "Disruptive Technologies: Advances that Will Transform Life, Business, and the Global Economy," McKinsey Global Inst., 2013; www.mckinsey.com/insights/business_technology/disruptive_technologies.

Kaibo Liu is an assistant professor in the Department of Industrial and Systems Engineering at the University of Wisconsin-Madison. Contact him at kliu8@wisc.edu.

Jianjun Shi is the Carolyn J. Stewart Chair and Professor in the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. Contact him at jianjun.shi@isye.gatech.edu.

Quantifying the Risk Level of Functional Chips in DRAM Wafers

Young-Seon Jeong, *Chonnam National University, Gwangju, Republic of Korea*
Byunghoon Kim, *Korea Institute of Science and Technology Information*
Seung Hoon Tong and In-Kap Chang, *Samsung Electronics*
Myong K. Jeong, *Rutgers University*

The semiconductor industry relies on informatics tools to extract useful information such as process conditions and yield rate predictions from multiple data sources.[1,2] In semiconductor manufacturing, various online sensor signals for monitoring process conditions and detecting defects are collected while semiconductor wafers are fabricated. Something as basic as 1-Gbyte dynamic random access memory (DRAM) has approximately 8.5 billion cells, with each cell's electrical device sorting (EDS) test results stored for later analysis.

Perfecting the automatic identification of defective chips on DRAM wafers based on big data analytics tools is one of the semiconductor industry's key challenges. Even though various studies have analyzed semiconductor wafer data, they've been limited to simple flash memory wafers from functional testing results. The DRAM wafer has a complex data structure, with multiple spatial maps and 2D fail bit maps for each chip. Several research challenges and opportunities related to massive DRAM wafers data include

- detection of spatial abnormalities in DRAM wafers with multiple spatial maps;
- defect pattern classification both at the chip level based on fail bit maps and at the wafer level based on multiple spatial maps;
- big data analytics with features with uncertain values; and
- quantification of the risk level for functional chips.

In addition, each integrated circuit (IC) involves multiple stages during fabrication. Output variables from a preceding stage in a multistage manufacturing process (MMP) sometimes function as input variables for the current stage. Also, there are lots of missing values in data due to sampling and other technical issues such as dysfunctional equipment or damaged sensor devices. Thus, additional challenges and opportunities in this area include

- inputting a substantial amount of missing values;
- modeling the interrelationship among the multiple stages under the presence of massive missing data; and
- quantifying the contribution of each individual stage to the variability of quality characteristics in the final stage.

Here, we introduce the important topic of quantifying the risk level of functional chips based on wafer testing results.

Quantifying the Risk Level

In the DRAM wafer process, each chip goes through a series of fail bit tests (FBTs), which are commonly used as a diagnosis tool for testing memory devices.[3,4] Based on multiple FBT results, each chip on a DRAM wafer is determined to be functional or defective. Figure 3 shows the graphical representation of the DRAM wafer testing process and corresponding test results.

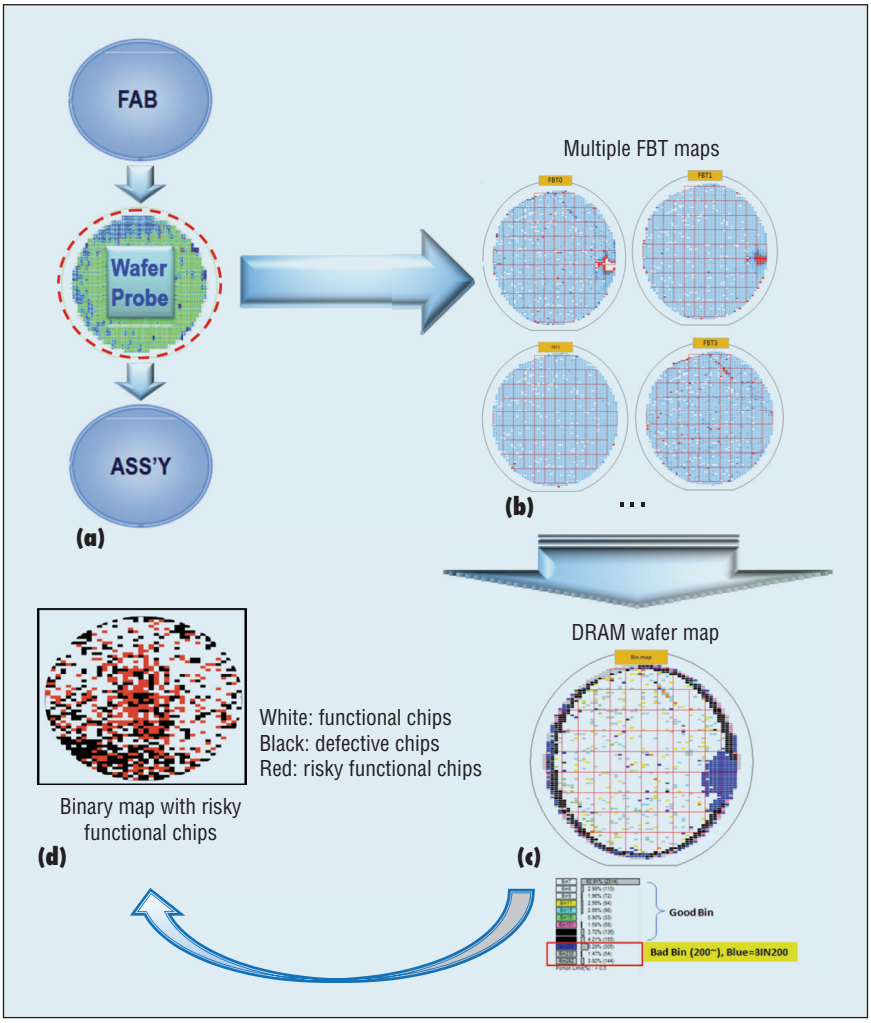

Figure 3. Graphical representation of the DRAM wafer testing process: (a) conceptual wafer manufacturing process; (b) multiple spatial fail bit test (FBT) maps generated from FBT tests; (c) binary wafer map; and (d) binary wafer map with the level of risk for functional chips. Risky functional chips can be classified as functional in the inspection process but could fail in the near future.

Multiple spatial FBT maps are generated based on FBT results in the final test process; a binary DRAM wafer map is built by overlapping those FBT maps. In practice, however, even though the chip is considered functional, it's important to quantify the probability of its being defective based on FBT results; we define these chips to be risky functional chips (marked as red in Figure 3d). Note that these risky functional chips were classified as functional in the inspection process but could fail in the near future. Thus, even though a functional chip passes all FBTs, it's important to quantify the risk of being defective so that engineers can take appropriate actions and ensure customer satisfaction.

This article introduces a novel procedure for quantifying the risk level of functional chips along with accurately identifying defective chips on DRAM wafer maps. To screen for risky functional chips, we propose using a robust relevance vector machine (RRVM) that can give the probability that a functional chip could potentially be defective.

Methodology

The discriminant analysis classifier and RRVM for classification quantify a functional chip's risk level.

Discriminant Analysis

Suppose we have n FBT results of a chip, $x = [x_1, x_2, \ldots, x_n]$, and the chip is either defective or functional. We can calculate the chip's distance to a decision boundary in the linear discriminant analysis (LDA) as follows[5]:

$$L(x) = x^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_0)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0), + \log \hat{\pi}_1 - \log \hat{\pi}_0,$$

where $\hat{\pi}_k$ denotes the fraction of class k in the data, $\hat{\Sigma} = \frac{\Sigma_{k=1}^{2} \Sigma_{g_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^t}{N-2}$ denotes the estimated covariance matrix of FBTs, N represents the total number of chips, $\hat{\mu}_k = \frac{\Sigma_{g_i} = k^{x_i}}{N_k}$ denotes the estimated mean of FBTs in class k, g_i represents class label of the ith chip, $i = 1, 2, 3, \ldots, N$, and N_k denotes the total number of chips in class k. In addition, the distance between a chip and a quadratic decision boundary can be obtained in the following quadratic equation[5]:

$$Q(x) = \frac{1}{2}\{-(x - \hat{\mu}_1)^T \hat{\Sigma}_1^{-1}(x - \hat{\mu}_1) + (x - \hat{\mu}_0)^T \hat{\Sigma}_0^{-1}(x - \hat{\mu}_0) + C$$

where

$$C = \frac{1}{2}\{\log|\hat{\Sigma}_0| - \log|\hat{\Sigma}_1|\} + \log \hat{\pi}_0 - \log \hat{\pi}_1.$$

Here, $\hat{\Sigma}_0 = \frac{\Sigma_{g_i=0}(x_i - \mu_0)(x_i - \mu_0)^T}{N_0 - 2}$ and, $\hat{\Sigma}_1 = \frac{\Sigma_{g_i=1}(x_i - \mu_1)(x_i - \mu_1)^T}{N_1 - 2}$ where N_0 and N_1 represent the total number of chips in each class, respectively.

RRVM

Relevance vector machine (RVM) is a Bayesian inference-based machine learning method for regression and classification problems.[6] RVM can provide probabilistic outputs for the given inputs while achieving a sparse prediction model. To build a predictive model that is robust to noises in the process, a robust relevance vector machine (RRVM) for classification was proposed recently.[7]

Consider a dataset of N FBT results target pairs $\{\mathbf{x}_i, t_i\}_{i=1}^{N}$, where $\mathbf{x}_i$ represents a d-dimensional input vector and t_i represents binary class labels: $t_i = 1$ if a chip is defective, and $t_i = 0$ otherwise. We can obtain a nonlinear decision boundary as a linear combination of M nonlinear basis functions (such as a Gaussian kernel function) as follows:

$$f(\phi(\mathbf{x})) = \boldsymbol{\beta}^T \phi(\mathbf{x}) = \beta_0 + \sum_{i=1}^{M} \beta_i \phi_i(\mathbf{x}),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots \beta_M)^T$ is a vector of model coefficients and $\phi(\mathbf{x}) = (1, \phi_1(x), \ldots, \phi_M(\mathbf{x}))^T$ is a vector of basis functions. Assuming independently distributed data, the conditional distribution for t under the framework of a standard logistic regression can be written as

$$p(\mathbf{t} \mid \boldsymbol{\beta}) = \prod_{i=1}^{N} p(t_i \mid \boldsymbol{\beta}) = \prod_{i=1}^{N} \sigma(\boldsymbol{\beta}^T \phi(\mathbf{x}_i))^{t_i} \{1 - \sigma(\boldsymbol{\beta}^T \phi(\mathbf{x}_i))\}^{1-t_i}.$$

Here, σ_u is the logistic function defined as $\sigma(u) = 1/(1 + e^{-u})$. To obtain robust model coefficients $\boldsymbol{\beta}$, a weighting strategy can be employed to maximize the likelihood function of the standard logistic regression model as follows

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} w_i l\{(2t_i - 1) f(\phi(\mathbf{x}_i))\},$$

where w_i is a weight associated with the ith observation, and $l(u) = \ln(1 + e^{-u})$ denotes the logistic loss function. We approximate the posterior distribution over model coefficients $\boldsymbol{\beta}$ using a variational inference method under the following prior distributions over the model parameters:

$$p(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) = \prod_{i=0}^{N} N(\beta_i \mid 0, \alpha_i^{-1})$$

$$p(\boldsymbol{\alpha} \mid a, b) = \prod_{i=0}^{N} Gamma(\alpha_i \mid a, b)$$

$$p(\mathbf{w} \mid c, d) = \prod_{i=0}^{N} Gamma(w_i \mid c, d)$$

Then, the best candidate model can be obtained by minimizing the following variational lower bound based on the iterative algorithm:

$$\tilde{L}[Q(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{w})] = \mathrm{E}[\ln h(\boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\xi})] + \mathrm{E}[\ln p(\boldsymbol{\beta} \mid \boldsymbol{\alpha})] + \mathrm{E}[\ln p(\boldsymbol{\alpha} \mid a, b)] + \mathrm{E}[\ln p(\mathbf{w} \mid c, d)] - \mathrm{E}[\ln Q_{\boldsymbol{\beta}}(\boldsymbol{\beta})] - \mathrm{E}[\ln Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})] - \mathrm{E}[\ln Q_{\mathbf{w}}(\mathbf{w})]$$

Performance Evaluation

A company in the semiconductor industry collected 62 DRAM wafers with 1,461 chips per wafer for experiments. Table 2 shows the data structure, where the second column (i, j) indicates each chip's location on a given DRAM wafer, and the other columns show the number of failed unit cells on each chip from each FBT. The last column shows whether each chip is defective (value = 1) or functional (value = 0). The dataset was randomly split into a training set of 42 wafers and a test set of 20 wafers.

Table 2. Data structure of FBT results from a DRAM wafer.

Wafer index	(*i*, *j*)	FBT 00	FBT 01		FBT *N*-2	FBT *N*-1	Test result
188G1W02	14, 54	6828717	454942	…	172	8073979	1
188G1W02	14, 55	3387648	448173	…	163	4650002	1
…	…	…	…	…	…	…	
188G1W10	52, 59	24	128	…	0	220	0
188G1W10	52, 60	1	107	…	0	192	0

Table 3. Accuracies of classification methods (%).

Method	Actual class	Predicted class: Defective	Predicted class: Functional	Accuracy
Linear discriminant analysis	Defective	79.1	21.0	83.9
	Functional	10.6	89.4	
Quadratic discriminant analysis	Defective	78.1	21.9	83.8
	Functional	9.9	90.1	
Robust relevance vector machine	Defective	94.2	5.7	93.5
	Functional	7.5	92.3	

Table 3 shows RRVM and other procedures' performance in terms of classification accuracies. The RRVM approach shows superior performance compared to LDA and QDA. Even though this is classified as functional, we need to quantify the risk level of being defective. Figure 4 visualizes the level of risk for each functional chip on a wafer map using the proposed RRVM approach. Here, a black unit block represents a defective chip, whereas white ones indicate functional chips. Each unit block's color represents the estimated risk level for the corresponding functional chip. A red chip represents a functional chip with a higher probability of failure in the near future. We can see that a functional chip close to

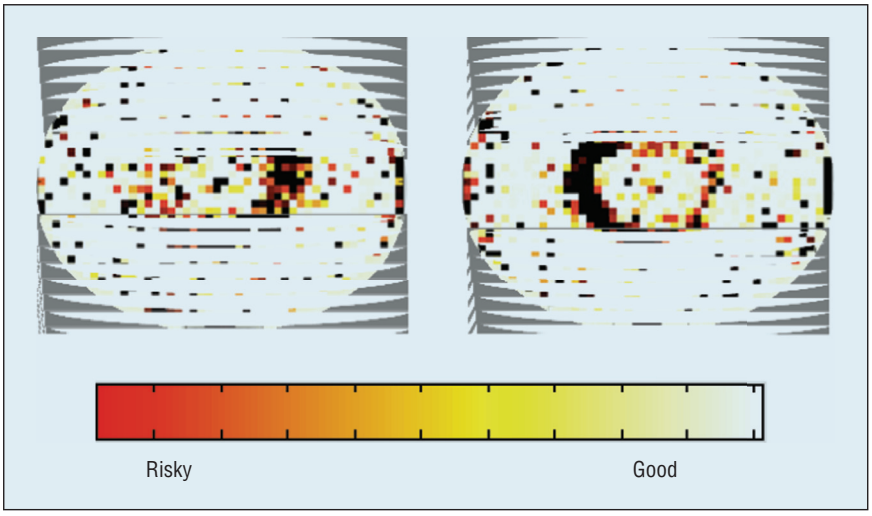

Figure 4. Visualization of risky functional chips. A red chip represents a functional chip with a higher probability of failure in the near future. We can see that a functional chip close to defective chips tends to have a higher risk level.

defective chips tends to have a higher risk level.

Our proposed Bayesian inference-based machine learning method effectively addresses the problem of quantifying risk levels for functional chips in semiconductor industry. Further investigations will aim to develop advanced models to deal with risky functional chips and confirm the effectiveness of informatics tools by experiments with massive real-life DRAM wafers.

Acknowledgments

Part of this work was supported by US National Science Foundation grant 1233800 and Chonnam National University internal grant 2014-0542.

References

1. J. Lee, B. Bagheri, and H. Kao, "Recent Advances and Trends of Cyber-Physical Systems and Big Data Analytics in Industrial Informatics," *Proc. Int'l Conf. Industrial Informatics*, 2014, DOI: 10.13140/2.1.1464.1920.
2. A. Bleakie and D. Djurdjanovic, "Feature Extraction, Condition Monitoring, and Fault Modeling in Semiconductor Manufacturing Systems," *Computers in Industry*, vol. 64, 2013, pp. 203–213.
3. B. Kim et al., "A Regularized Singular Value Decomposition-Based Approach for Failure Pattern Classification on Fail Bit Map in a DRAM Wafer," *IEEE Trans. Semiconductor Manufacturing*, vol. 28, no. 1, 2015, pp. 41–49.
4. B. Kim et al., "Step-Down Spatial Randomness Test for Detecting Abnormalities in DRAM Wafers with Multiple Spatial Maps," to be published in *IEEE Trans. Semiconductor Manufacturing*, 2015.
5. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.
6. M.E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Machine Learning Research*, vol. 1, 2001, pp. 211–244.
7. S. Hwang and M.K. Jeong, "Robust Relevance Vector Machine for Classification with Variational Inference," to be published in *Annals Operations Research*, 2015.

Young-Seon Jeong is an assistant professor in the Department of Industrial Engineering at Chonnam National University, Republic of Korea. Contact him at youngjeong@jnu.ac.kr.

Byunghoon Kim is a senior researcher in the Department of Small- and Medium-Sized Enterprises Innovation at the Korea Institute of Science and Technology Information (KISTI), Republic of Korea. Contact him at bhkim@kisti.re.kr.

Seung Hoon Tong is a master engineer (executive) in the Department of Memory Business at Samsung Electronics, Republic of Korea. Contact him at shtong@samsung.com.

In-Kap Chang is a principle research engineer in the Department of Memory Business at Samsung Electronics, Republic of Korea. Contact him at inkap.chang@samsung.com.

Myong K. Jeong is an associate professor in the Department of Industrial and Systems Engineering at Rutgers University. Contact him at mjeong@rci.rutgers.edu.

Flight Operations Monitoring through Cluster Analysis: A Case Study

Florent Charruaud and Lishuai Li, *City University of Hong Kong*

Although aviation safety has improved significantly over the past few decades, its track record is far from perfect in light of recent mishaps that have made worldwide headlines. Despite the industry's continuing efforts to monitor and analyze flight operations, modern aircraft systems have become immensely complex, to a degree that traditional analytical methods are reaching their limits. We're sitting on piles of underutilized data that could be otherwise used to improve aviation safety.

Digital flight data are a good example. Recorded continuously from engine start to engine shutdown on each flight, the data contain rich

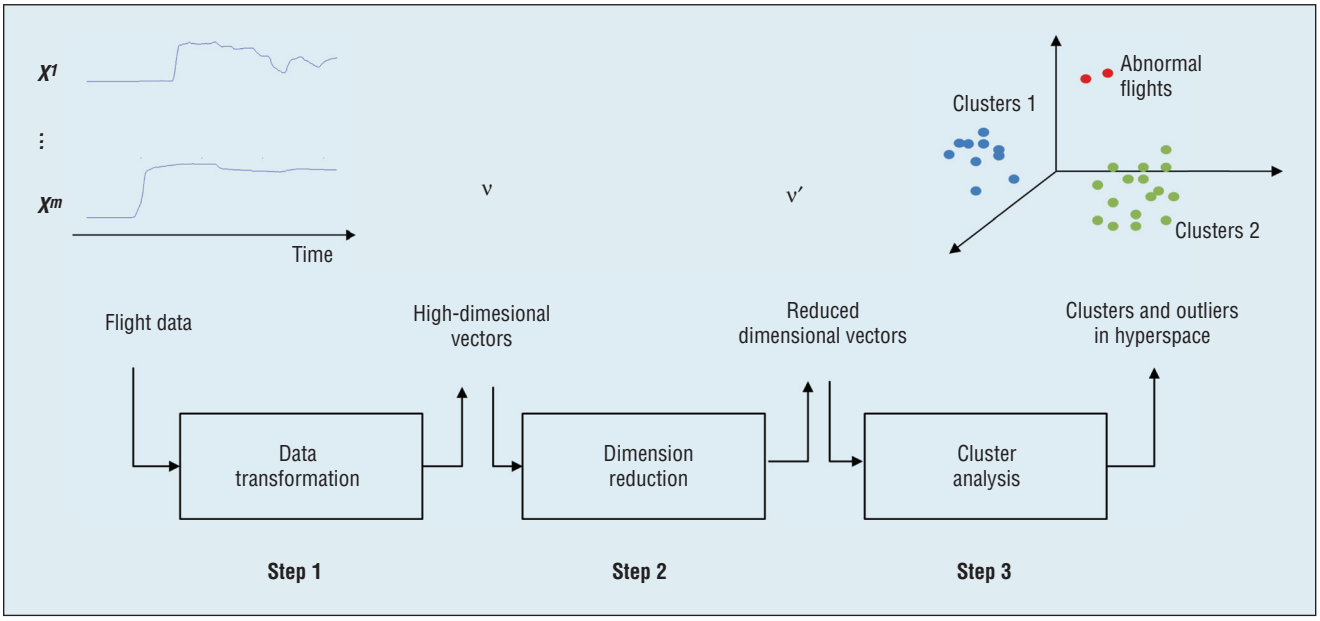

Figure 5. Key steps in ClusterAD-Flight. The first step is to transform flight data into high-dimensional data vectors that capture each flight's multivariate and temporal characteristics. In the second step, the dimensions of these vectors are reduced to address issues related to data sparseness and multicollinearity. The third step applies cluster analysis to the reduced vectors.

information about aircraft systems, pilot operations, and flight conditions. A typical tool the airline industry uses is called exceedance detection (ED), in which flight parameters are compared with predefined thresholds to raise the red flag and identify safety risks, such as equipment problems, environmental hazards, and pilot errors. Needless to say, ED's effectiveness is limited to known safety issues only—for emerging issues that are unknown, the tool can't respond until thresholds are eventually updated, possibly triggered by another accident.

Recent advances in data mining have attracted new interest in monitoring and analyzing flight data. Compared with existing options, data-driven approaches can proactively detect anomalies based on data patterns rather than predefined thresholds. The Inductive Monitoring System (IMS),[1] for example, uses supervised learning in which typical system behaviors are derived from a training dataset and then compared with operational data to detect abnormal behaviors. However, IMS doesn't account for temporal patterns (event sequence and timing), which are critical features of dynamic systems such as aircraft. The sequence miner algorithm[2] focuses on discrete flight parameters to monitor pilot operations, such as cockpit switch flips, yet a majority of flight data come from continuous parameters, such as altitude, airspeed, and engine temperature. Multiple kernel anomaly detection (MKAD)[3] applies a one-class support vector machine (SVM) for anomaly detection, but it assumes one type of data pattern for normal operations, which isn't always valid in real operations because standards can vary according to flight conditions.

We propose a new method[4,5] to help airline safety experts monitor daily flights and detect anomalies. Without knowing the norm standard a priori, this method, referred to as cluster-based anomaly detection to detect abnormal flights (ClusterAD-Flight), applies clustering techniques on flight data to identify standard operations and anomalies. We also provide a case study to demonstrate ClusterAD-Flight in action with real data.

Flight Operations Monitoring via ClusterAD-Flight

ClusterAD-Flight is based on cluster analysis, a widely used data mining technique to identify common patterns. We assume that most flights exhibit common patterns under routine operations; a few outliers that deviate from those common patterns are of interest to airline safety management. Our method includes three key steps, as Figure 5 shows. The first step is to transform flight data into high-dimensional data vectors that capture each flight's multivariate and temporal characteristics. In the second step, the dimensions of these vectors are reduced to address issues related to data sparseness and multicollinearity. The third step applies cluster analysis to the reduced vectors. Groups of proximate vectors are clusters, or the common patterns; stand-alone vectors are anomalies, also called outliers.[4,5]

We developed a process of applying ClusterAD-Flight for flight operations monitoring (see Figure 6). First, raw flight data from different

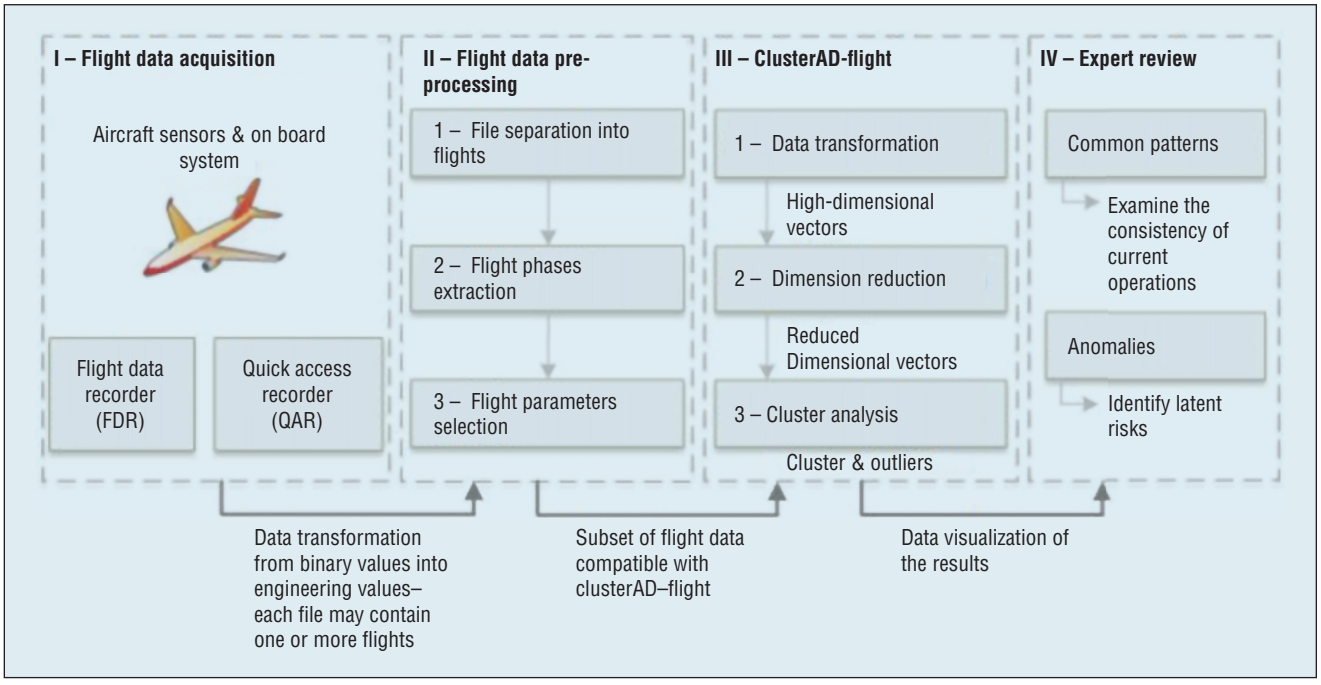

Figure 6. Process of using ClusterAD-Flight for flight operations monitoring. First, raw flight data from different recording devices are translated from binary files into engineering values. Second, a preprocessing step segments data by flight and phase, and selects subsets of interest for analysis. Afterward, ClusterAD-Flight identifies common patterns and anomalies. Finally, a group of domain experts review the results.

Table 4. Key characteristics of three clusters identified by ClusterAD-Flight.

	No. flights	Flight length	Gross weight	Power setting	Flap angle
Cluster 1 (red)	483	Long-haul	Heavy	High	15
Cluster 2 (green)	198	Short-haul	Light	Low	5
Cluster 3 (blue)	181	Long-haul	Heavy	High	5

recording devices, such as Quick Access Recorder (QAR), and Digital Flight Data Recorder (DFDR), are translated from binary files into engineering values using available flight data processing software tools. Second, a preprocessing step segments data by flight and phase, and selects subsets of interest for analysis. Afterward, ClusterAD-Flight identifies common patterns and anomalies. Finally, a group of domain experts (similar to some airlines' monthly safety review boards, which include safety experts, line pilots, flight data analysts, and pilot training experts) review the results to check if common patterns are consistent with standard procedures and identify risks from anomalies, if any, and inform operations before accidents occur.

Case Study

We tested ClusterAD-Flight on operational data from an international airline. Our objective was to illustrate its capability in recognizing common patterns in flight operations and detecting anomalies with unique data patterns. A highlight of this case study is that we're evaluating a method that doesn't rely on pre-existing criteria but is capable of detecting unknown issues. There's no "gold standard" that defines which flights are abnormal in definitive terms, nor is there labeled data for benchmarking. Thus, unlike standard evaluation metrics such as detection accuracy rates or receiver operating characteristic (ROC) curves, we used qualitative analysis based on input from airline experts in this study.

As a proof-of-concept, we focused on the takeoff operation of the B777-300ER fleet at the Hong Kong International Airport (HKG) over a period of one month. We incorporated 17 flight parameters from 957 flights in the analysis, including but not limited to engine parameters, aircraft position, speeds, accelerations, attitudes, control surface positions, and wind information.

Common Patterns

Cluster-AD Flight identified three distinct clusters, which airline experts later confirmed as common patterns of standard takeoff operations at HKG over that period. Table 4 summarizes key characteristics of the three clusters; Figure 7 shows important flight parameters. All the short-haul

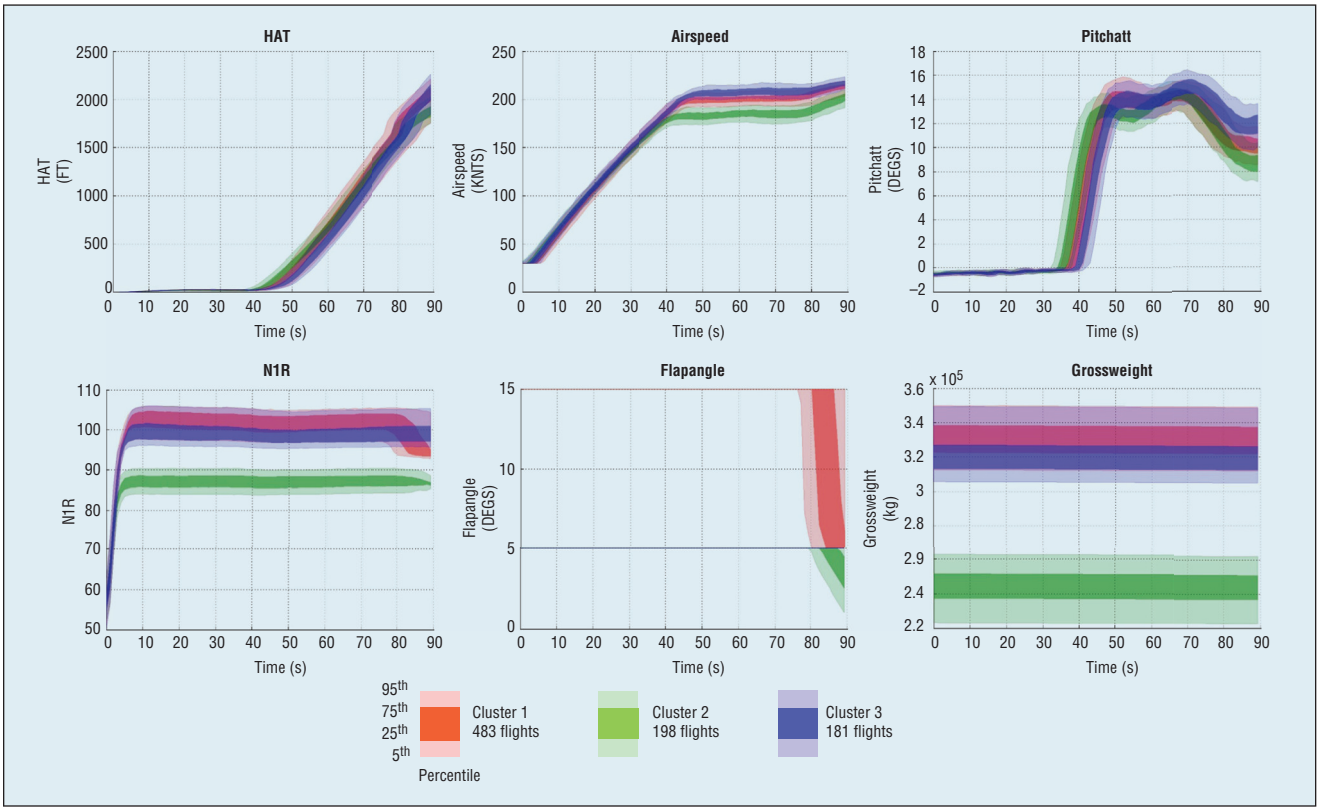

Figure 7. Three clusters identified by ClusterAD-Flight, which correspond to three groups of standard takeoff operations. All the short-haul flights, marked in green and labeled cluster 2, are a distinct group due to their light weight, low takeoff power, low flap setting, and slower-climbing airspeed.

flights, marked in green and labeled cluster 2, are a distinct group due to their light weight, low takeoff power, low flap setting, and slower-climbing airspeed. Long-haul flights can be further grouped into clusters 1 and 3; the former accounted for the majority and the latter consisted of flights with lower power and flap settings due to specific environmental conditions.

Anomalies

Flights with distinct data patterns are clearly identifiable from ClusterAD-Flight results, and their unusual operations could indicate latent risks. We describe three examples from the top 1 percent of abnormal flights detected here.

Figure 8 shows an abnormal flight classified as high-power takeoff, with an altitude and airspeed profile significantly higher than others of the same common pattern. For a short-haul flight with relatively light weight, the use of a full takeoff power setting contradicts the common procedures used by other short-haul flights, as shown in the green cluster. Clearly, this is an uncommon operation inviting further investigation for latent risks. Aside from being an operational mistake, a senior pilot suggested it could be a conscious choice of the pilot in response to potential wind shear.

Figure 9 shows another abnormal flight of low-and-slow takeoff under tailwind conditions. As a short-haul flight, it should be part of the green cluster, yet its altitude and airspeed profile is consistently lower than most others in the common pattern. The 10- knots tailwind could cause this degraded takeoff performance. In addition, there's a sudden 12-degree change of roll attitude 70 seconds after takeoff (shown in the Rollatt chart in Figure 9), while heading and crosswind remained unchanged. Airline safety experts suggested a case of wake turbulence encounter, which is especially hazardous when a takeoff climb is low and slow. Wake turbulence forms behind an aircraft as it passes through the air; it creates risks to other aircraft in the vicinity, and the danger is greatest when an aircraft is operating at low speed and low height because it leaves little room to recover from any upset.

The third example is a takeoff with an unusual flap setting (see Figure 10). The common flap setting is 5 (green and blue clusters) or 15 degrees (red cluster) for takeoff, while this unique flight set the flap angle to 20 degrees. Consequently, its climb rate and acceleration were higher than most other flights. The cause of this high flap angle setting in this anomaly is worth investigating to determine whether it was an intentional maneuver to overcome potential risks of wind shear, turbulence, or other undesired weather conditions during takeoff.

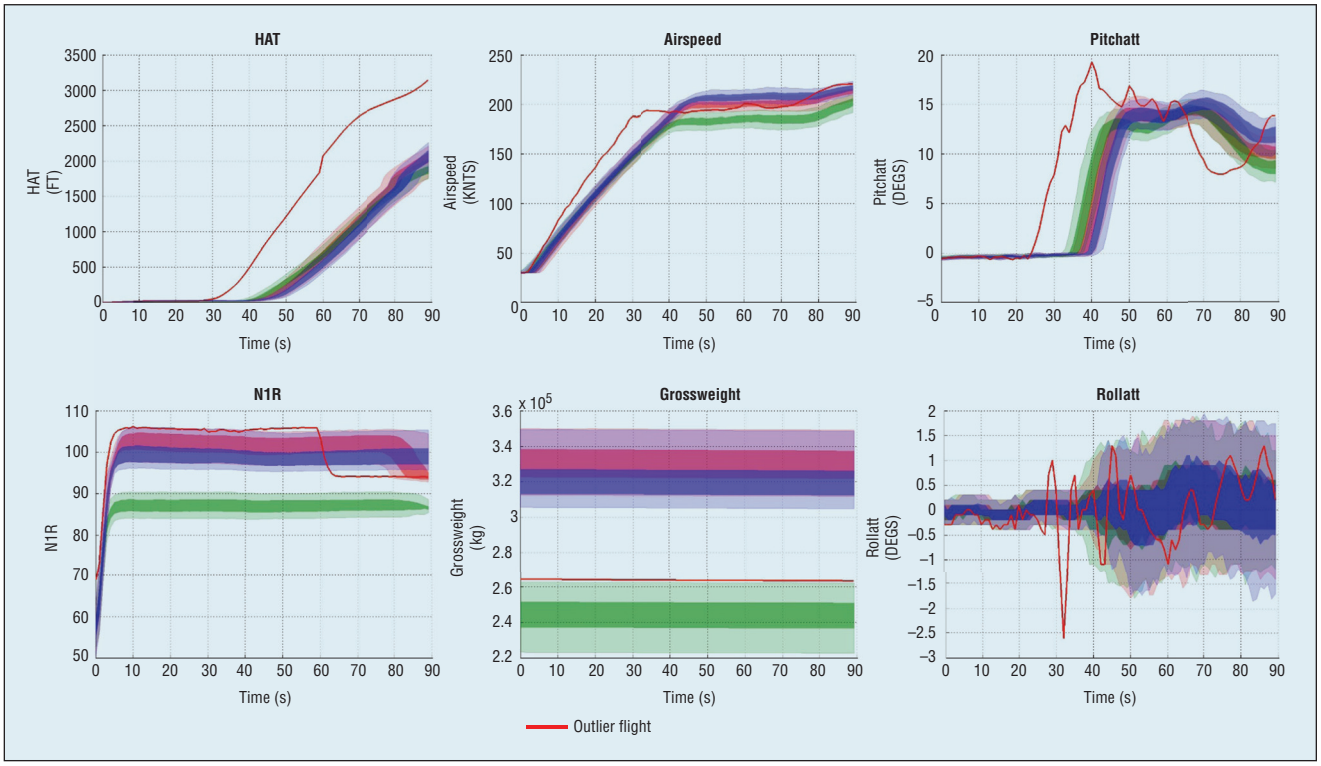

Figure 8. High-power takeoff. For a short-haul flight with relatively light weight, the use of a full takeoff power setting contradicts the common procedures used by other short-haul flights, as shown in the green cluster.

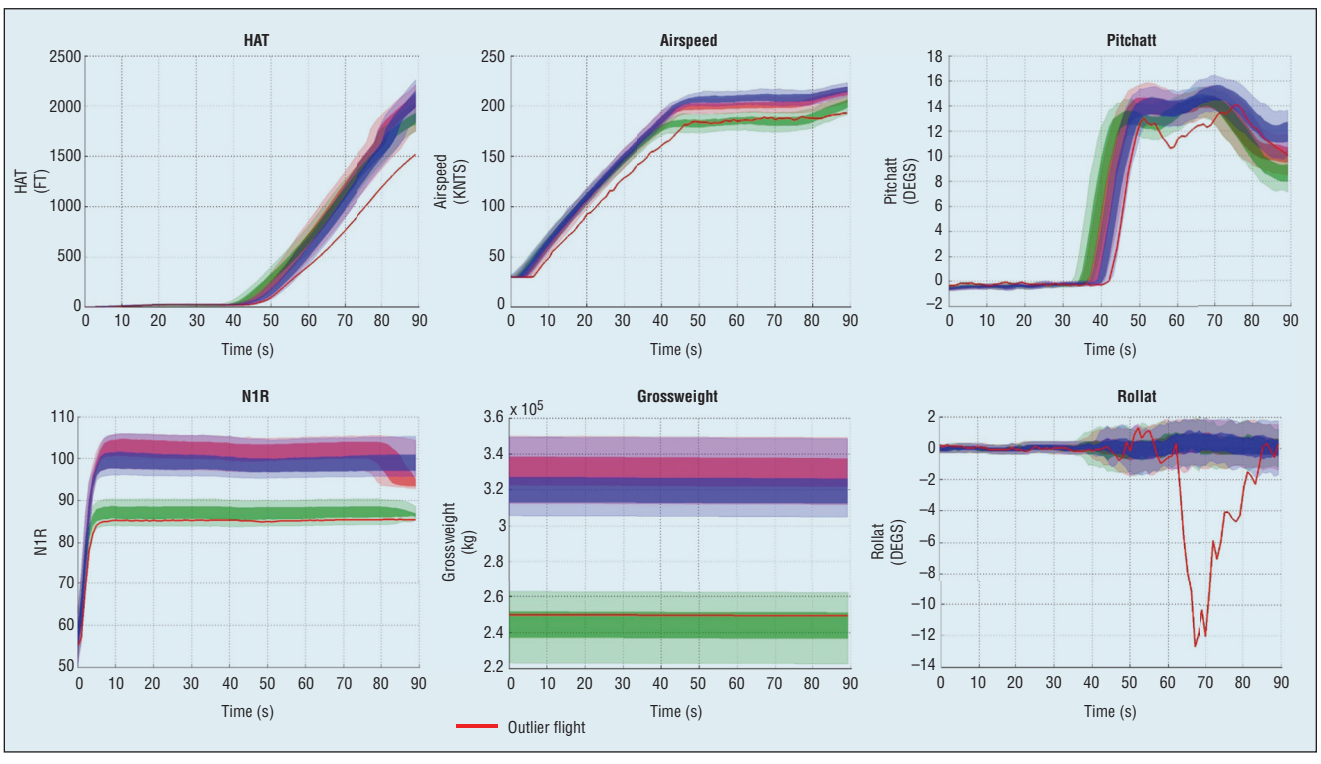

Figure 9. Low-and-slow takeoff with tailwind. As a short-haul flight, it should be part of the green cluster, yet its altitude and airspeed profile is consistently lower than most others in the common pattern.

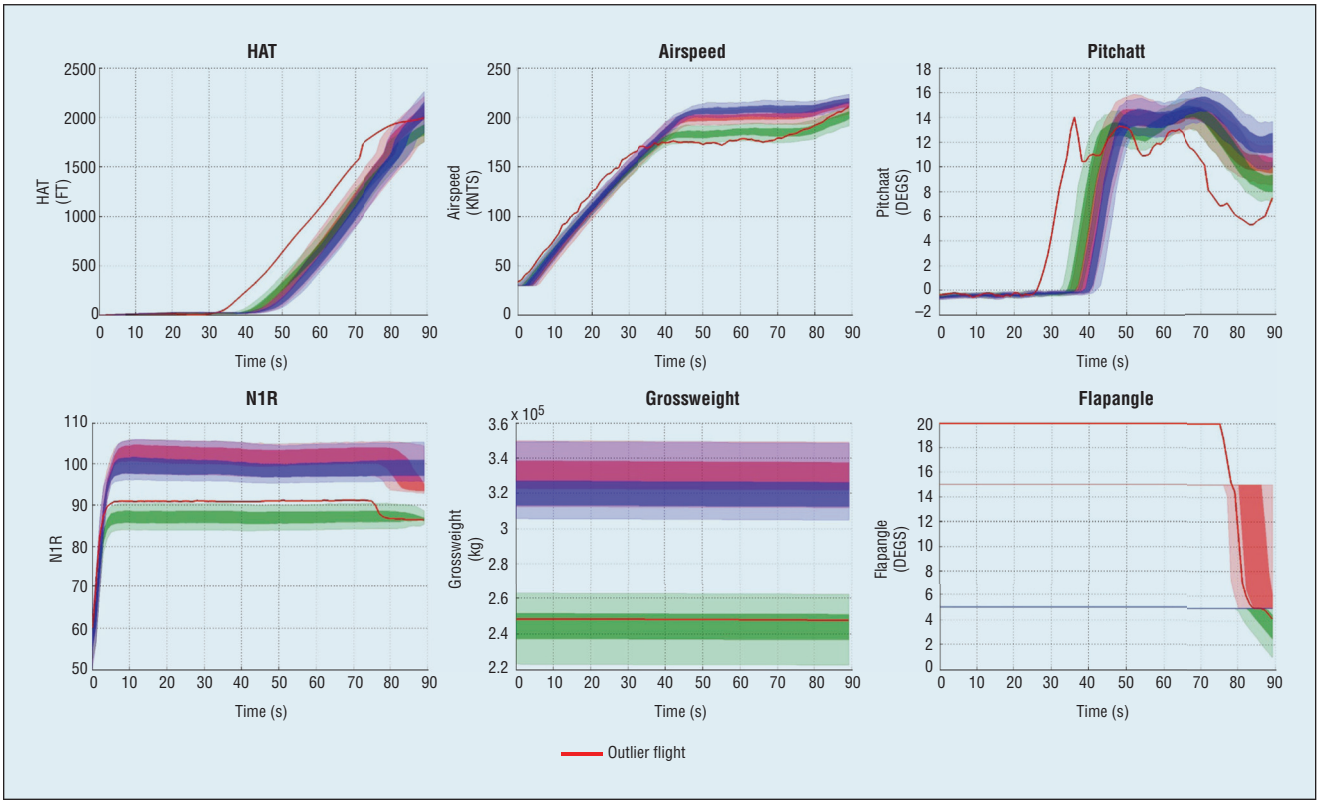

Figure 10. Unusual flap setting takeoff. The common flap setting is 5 (green and blue clusters) or 15 degrees (red cluster) for takeoff, while this unique flight set the flap angle to 20 degrees.

Our results demonstrate several advantages of applying ClusterAD-Flight in flight operations monitoring. Compared with the industry standard method of ED, it's no longer limited to predefined criteria and can detect flights with rare data patterns, allowing airline safety experts to identify patterns worthy of further investigation from tens of thousands of routine flights. Unlike other recently developed anomaly detection methods, ClusterAD-Flight can recognize common patterns in the form of clusters, allowing the management team to check the consistency of current operations. In our case study, the airline found it useful to know that most takeoffs reached at least 1,500 feet (height above takeoff) in 80 seconds, achieving at least 210 knots after 90 seconds, and exhibiting the target pitch between 12 to 16 degrees. Our method is particularly useful in proactive airline safety management, where a broader toolset is needed. ClusterAD-Flight can also be applied in other fields where unknown issues are quickly emerging, while the standard of the norm is absent or not up to date. ❐

Acknowledgments

This work was supported by the City University of Hong Kong (project 7200418). We thank Cathay Pacific Airways for providing flight data, professional knowledge, and insights to enable this research.

References

1. D.L. Iverson, "Inductive System Health Monitoring," *Proc. 2004 Int'l Conf. Artificial Intelligence* (IC-AI04), 2004, pp. 605–611.
2. S. Budalakoti, A.A.N. Srivastava, and M.E.M. Otey, "Anomaly Detection and Diagnosis Algorithms for Discrete Symbol Sequences with Applications to Airline Safety," *IEEE Trans. Systems, Man, Cybernetics, Part C*, vol. 39, no. 1, 2008, pp. 101–113.
3. S. Das et al., "Multiple Kernel Learning for Heterogeneous Anomaly Detection: Algorithm and Aviation Safety Case Study," *Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2010, pp. 47–56.
4. L. Li et al. "Analysis of Flight Data Using Clustering Techniques for Detecting Abnormal Operations," to appear in *J. Aerospace Information Systems*, 2015, DOI: 10.2514/1.I010329.
5. L. Li, "Anomaly Detection in Airline Routine Operations Using Flight Data Recorder Data," PhD thesis, Dept. of Aeronautics and Astronautics, Massachusetts Institute of Technology, 2013.

Florent Charruaud is a master's student in the Department of Systems Engineering and Engineering Management, City University of Hong Kong. Contact him at florent.charruaud@gmail.com.

Lishuai Li is an assistant professor in the Department of Systems Engineering and Engineering Management, City University of Hong Kong. Contact her at lishuai.li@cityu.edu.hk.