A Statistical (Process Monitoring) Perspective on Human Performance Modeling in the Age of Cyber-Physical Systems

Fadel M. Megahed, L. Allison Jones-Farmer, Miao Cai, Steven E. Rigdon and Manar Mohamed

Abstract With the continued technological advancements in mobile computing, sensors, and artificial intelligence methodologies, cyber-physical convergence is becoming more pervasive. Consequently, personal device data can be used as a proxy for the "human operator", creating a digital signature of their typical usage. Examples of such data sources include: wearable sensors, motion capture devices, and sensors embedded in work stations. Our motivation behind presenting this paper is to encourage the quality community to investigate relevant research problems that pertain to human operators. To frame our discussion, we examine three application areas (with distinct data sources and characteristics) for "human performance modeling": (a) identification of physical human fatigue using wearable sensors/accelerometers; (b) capturing changes in a driver's safety performance based on fusing on-board sensor data with online API data; and (c) human authentication for cyber-security applications. Through three case studies, we identify opportunities for applying industrial statistics methodologies and opportunities for future work. To encourage future examination by the quality community, we host our data, code and analysis on an online repository.

Fadel M. Megahed
Miami University, 800 E. High Street, Oxford, OH 45056, USA. e-mail: fmegahed@miamioh.edu

Allison Jones-Farmer
Miami University, 800 E. High Street, Oxford, OH 45056, USA. e-mail: farmerl2@miamioh.edu

Miao Cai
Saint Louis University, 3545 Lafayette Ave., Room 481 St. Louis, MO 63103, USA. e-mail: miao.cai@slu.edu

Steven E. Rigdon
Saint Louis University, 3545 Lafayette Ave., Room 481 St. Louis, MO 63103, USA. e-mail: steve.rigdon@slu.edu

Manar Mohamed
Miami University, 510 E. High Street, Oxford, OH 45056, USA. e-mail: mohamem@miamioh.edu

1. Introduction

With the ever-decreasing costs of wireless networking and continued advancements in mobile computing technologies, we now live in a connected world. The number of internet-connected devices, e.g., sensors, machines/equipment, and medical devices, continues to exponentially increase. The data, networking and infrastructure supporting these devices is commonly referred to as the *Internet of Things* (IoT) (Gubbi et al, 2013). The *International Data Corporation* (IDC, 2019) estimates that there will be 41.6 billion connected devices, generating 79.4 zettabytes ($1\,ZB \approxeq 1$ trillion terabytes) of data in 2025.

There are a large number of opportunities to use these data/devices to transform business operations. It is expected that IoT can lead to new paradigms in: (a) "smart and connected health" in health-care operations/management (Leroy et al, 2014; Chen et al, 2018); (b) "Industry 4.0" (Lasi et al, 2014) or "Industrial Internet of Things" (IIoT) (Jeschke et al, 2017) in manufacturing and supply chain management applications; (c) "smart cities" or "smart grids", where local governments and/or energy companies use IoT sensors to manage its resources more efficiently; (d) "smart farming", where agricultural decisions are informed by embedded sensors and/or drones (Wolfert et al, 2017); and (e) "autonomous" or "connected" vehicles, which capitalize on a large number of internet-connected sensors.

A common theme among these applications is that the use of wireless technology in these domains is now possible due to the development of engineering systems that capitalize on the seamless integration of computation and physical components. For this reason, Helen Gill, PhD, of the United States' National Science Foundation (NSF) coined the term *cyber-physical systems (CPS)* around 2006 as a catch-all phrase to capture those technologies (Lee and Seshia, 2017). The National Science Foundation (2019) states that:

> CPS will enable capability, adaptability, scalability, resiliency, safety, security, and usability that will expand the horizons of these critical systems. CPS technologies are transforming the way people interact with engineered systems, just as the Internet has transformed the way people interact with information ... Moreover, the integration of artificial intelligence with CPS creates new research opportunities with major societal implications ... While tremendous progress has been made in advancing CPS technologies, the demand for innovation across application domains is driving the need to accelerate fundamental research to keep pace.

There are two main observations that need to be highlighted based on NSF's vision for CPS technologies. First, the expectations regarding the transformation potential of CPS technologies are quite high. We agree with the statement that *fundamental research* is needed to help unlock such potential and allow for innovative applications. Second, it is interesting/unfortunate to note that despite the fact that the above synopsis stems from the joint work of several directorates of NSF as well as the research and development (R&D) arms of several U.S. governmental agencies, there is limited discussion of how statistical methodologies can be capitalized on and/or are needed to be developed to advance the current state of CPS technologies. In our view, the utility of statistical approaches (outside of regression, which

non-statisticians often use for baseline comparisons in machine learning applications) is not fully understood by practitioners and researchers engaged with CPS technology. These researchers and practitioners may have had limited exposure to statistical training which explains why statistical methodologies have not been fully considered in such applications.

There are three primary objectives behind this book chapter: (a) review and categorize the literature within the field of industrial statistics examining how CPS technologies can be used in "modeling human performance"; (b) present an overview of the types of statistical modeling approaches that can be considered in the context CPS analysis; and (c) highlight how industrial statistics methodologies can be used/developed to advance the reviewed literature. We focus on the general area of human performance modeling since it: (i) has been largely ignored by the industrial statistics community; and (b) can be considered as an important pillar in many application domains (e.g., advanced manufacturing, motor vehicle safety, and cybersecurity where humans continue to be the most important and vulnerable link).

The remainder of this book chapter is organized as follows. In Section 2, we provide a data-driven review of industrial statistics (quality control/engineering, reliability and experimental design) papers and highlight that our literature has yet to focus on CPS technology applications. In Section 3, we discuss the limitations in the approaches to analyze CPS data and suggest ways in which these problems can be informed by the field of industrial statistics. Then, we examine how specific statistical methodologies can give insights to three CPS applications in Sections 4-6. Our goals in these sections are to: (a) summarize main research streams within these applications; (b) provide an example to explain the data structure in detail; and (c) discuss future opportunities for statistical methodologies. Finally, we present our concluding remarks in Section 7.

2. Background

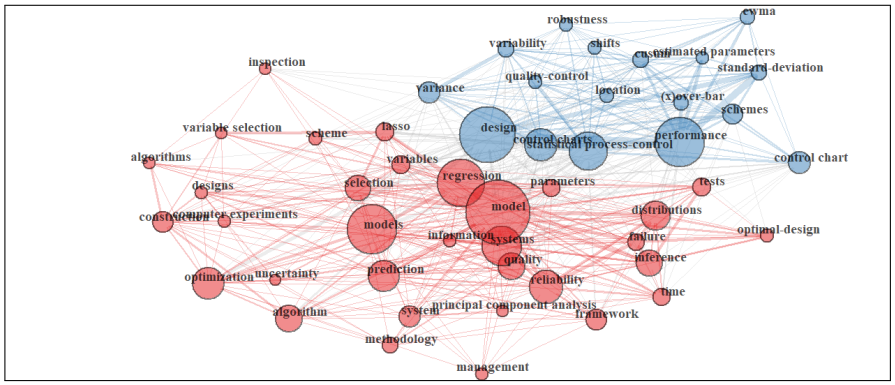
Prior to examining CPS technologies and models, it is important to understand how the field of "industrial statistics" has evolved over the past few years. The goal here is to attempt to see whether statistical approaches that focus on monitoring, experimental design, and reliability have started tackling the data structures and application domains associated with CPS. To achieve this goal, we extracted titles, abstracts, citations, keywords, and authors information from all articles published in *Technometrics*, *Journal of Quality Technology*, *Quality and Reliability Engineering International*, and *Quality Engineering* between 2014 - August 1, 2019. The time span and journals were selected to capture the most recent developments in popular journals associated with the areas of statistical surveillance/monitoring, experimental design, and reliability. Additionally, we have limited the results to the following document types: (a) Article, (b) Early Access, or (c) Review since we did not want to include results from letters to the editor, book reviews, etc. The search was conducted on August 1, 2019 and resulted in 1576 journal papers. Capitalizing on techniques from bibliometrics and natural language processing (NLP), we present an overview of those papers in the paragraphs below.

From a bibliometric stand point, the process of analyzing a large corpus of papers typically follows the following sequential steps: (Börner et al, 2003) (1) data extraction, which we performed using the Web of Science's portal; (2) definition of unit of analysis, in our analysis, we focused on keywords (provided by author and extracted from titles) as well as paper-relationships, which of our 1576 papers were cited by other papers in our corpus of text; (3) selection of metrics for evaluation and computing similarity between units of analysis, the reader can refer to our code in the *Supplementary Materials* Section for more details on how we computed similarity; and (4) data visualization and analytics. Based on this framework, we present our results for three different units of analysis in Fig. 1, which was generated using the *bibliometrix* **R** package (Aria and Cuccurullo, 2017).

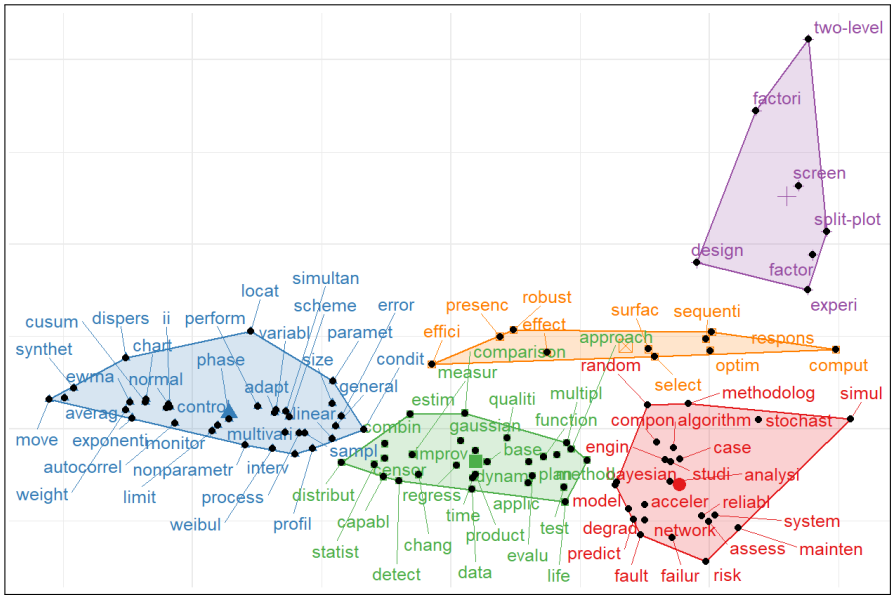
Fig. 1a depicts the 50 most frequently used keywords among the 1576 papers. One can easily see that the majority of these keywords correspond to fundamental concepts/techniques in industrial statistics. Examples include: statistical process control, estimated parameters, computer experiments, regression, and control charts. Note that the size of the node corresponds to the frequency by which its used. Thus, the terms: *model*, *models*, *design*, *performance*, and *statistical process control* are among the most used to describe/summarize our research papers. These should not be surprising, given the methodological nature of our journals. That being said, it was surprising to find that there is not any application domains in the list of the top 50 keywords. Note that: (a) the term construction was primarily used to describe statistical concepts (e.g., construction of control limits) and not the actual field where buildings are built; and (b) our initial guess was to find terms such as *Industry 4.0*, *advanced manufacturing*, and/or *public health* among the top fifty keywords. Moreover, the co-occurrence of keywords indicate the presence of two large clusters: (a) top/blue cluster, captures statistical process control (SPC) methodologies, and (b) bottom/red cluster, which captures other sub-domains within industrial statistics (e.g., experimental design, reliability, dimension reduction and model selection).

To capture an alternative representation of the major concepts in our literature, we have extracted keywords from the titles of the 1576 articles. Here, we have utilized Porter's stemming algorithm (Porter, 2006) to combine keywords/concepts that have a similar root (e.g., measur is used in lieu of measurement and measuring) and *k-means clustering* to identify how of the concepts should be grouped. Additionally, we have limited our analysis to stemmed terms that appeared at least 20 times such that the graph can be readable. Fig. 1b shows that there are five main topics that are being investigated in our literature. When compared to Fig. 1a, we can see that Fig. 1b presents a more detailed perspective on the literature. Similar to Fig. 1a, the stemmed words do not directly capture the types of applications being examined.

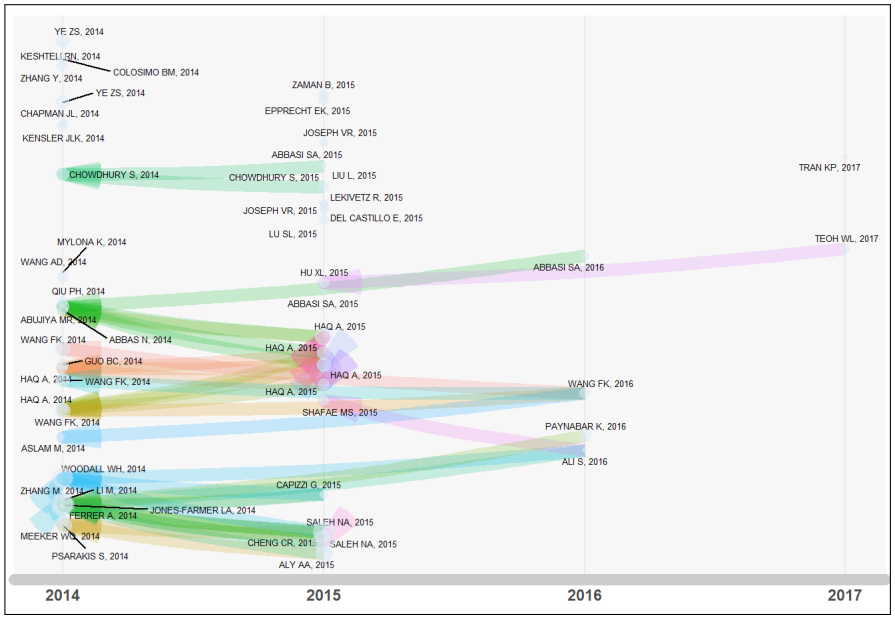
We also examined what of our pool of recently published papers have influenced (i.e. was cited by) other articles among the 1576 documents. To facilitate the graphical representation of this analysis, we reduced the number of papers to only those that were cited at least 10 times (per Web of Science calculations on August 1, 2019). The results of this analysis is depicted in Fig. 1c. From the figure, several observations can be made:



(a) The co-occurrence of the 50 most frequently used keywords



(b) A concept map of the literature, organizing stemmed title words into five clusters



(c) Historical direct citation network of papers with at least 10 citations (among our 1576 papers)

Fig. 1: A bibliometric analysis of 1576 journal papers, published in four popular industrial statistics journals between 2014-2019.

(a) As is expected, the introduction of an absolute value of citations provides an advantage to papers published in the early part of our 5+ year time span. Thus, we can see more papers being captured in 2014-2015 when compared to 2016-2017. No papers published in 2018 and 2019 were captured in this analysis.
(b) Ye and Chen (2014) is the most cited paper in our dataset, with 140 Web of Science citations at the time of our analysis. This paper is captured as Ye ZS, 2014 in our visualization (see top left paper). From the figure, one can see that there are no arrows associated with this paper. We did not expect this result. In our view, there are two possible explanations for the lack of arrows: (i) most of its citations are captured in journals that are not included in our analysis; and/or (ii) none of the citing papers within our dataset have accumulated 10 Web of Science citations.
(c) Self-cites dominate the arrows in the visualization (see e.g., the works of Haq, Chowdhury, and Wang). It is important to note that, in the context of this analysis, this statement is not intended to be a negative comment since papers included in this list had to be published in one of our four selected journals in/after 2014 and have at least 10 citations. Thus, this merely captures potentially important work (limited by the pros and cons of using citations as a proxy to a work's quality/importance), where the author(s) continued pursuing this research stream.
(d) The majority of papers in the figure were related to statistical process monitoring/ control charting methodologies. Examples include: Psarakis et al (2014); Haq et al (2014); Zhang et al (2014); Woodall and Montgomery (2014); Jones-Farmer et al (2014); Haq et al (2015); Capizzi (2015); Paynabar et al (2016); Teoh et al (2017).
(e) Among these papers, CPS-related contributions were limited to: Colosimo et al (2014) and Del Castillo et al (2015), who investigated how profile surfaces can be monitored in advanced manufacturing scenarios.

Based on this data-driven analysis of the literature, we believe that our research community need to be more active in developing methodologies for CPS technologies. As a part of the methodology development, it is important to showcase when our approaches can be used and their benefits/disadvantages when compared to machine learning methodologies.

3. A Conceptual Framework to Categorize Strategies for Approaching CPS Modeling Applications

To illustrate several approaches for modeling CPS applications, let us consider a wearable smart health system that can be used to provide an overall estimation of a user's health condition at any point in time. Table 1, based on Pantelopoulos et al (2010), presents three sensing technologies that can be incorporated in such a system, along with their corresponding measured physiological signals. From an an application point of view, one can envision that (after the necessary preprocessing of data) a wearable system containing one of those sensors can: (a) act as a surveil-

lance tool, i.e. alert user/medical-provider if the heart rate exceeds a pre-determined threshold; (b) provide diagnostic information, e.g., whether a user is diabetic or not based on their blood glucose levels; or (c) categorize activity into sleeping, being awake, exercising, etc based on the accelerometer values. Based on each objective/response type, different models can be used for analysis.

Table 1: Selective bio-signals and their associated sensor devices. The presented content is adapted from Table I in Pantelopoulos et al (2010).

Signal	Sensor Type	Description of Measured Data
Heart rate	Pulse oximeter / skin electrodes	Frequency of the cardiac cycle
Blood glucose	Strip-base glucose meters	Measurement of glucose levels in the blood
Body movements	Accelerometer	Measurement of acceleration forces in the 3D space

From a practitioner's (and/or a machine learning researcher's) perspective, the choice of models used can be informed through several existing frameworks. Among *Python* users, the scikit-learn algorithm flowchart/cheat-sheet is a popular starting-point (Scikit-Learn, 2019), where problems are defined into four groups: (a) classification, where the response variable is categorical, (b) regression, where the response variable is continuous, (c) clustering, where one would like to group observations; however, this approach is unsupervised since no label is provided for classification; and (d) dimension reduction, methodologies. Among each group, the flowchart provides yes/no questions to guide users to an appropriate methodology. Note that, while the flowchart contains several statistical models (e.g., principal component analysis, naive Bayes, LASSO, elastic net and ridge regression), it does not provide any insights on how/when methodologies from statistical surveillance, design of experiments, and reliability modeling may be appropriate. SAS® has provided a similar algorithm cheat-sheet (Li, 2017), where they also divide the algorithms into the same four groups. However, the provide details on which algorithms should be selected based on speed, accuracy, ease of explanation, and/or whether the outcome should be treated as probabilistic. Thus, the SAS flowchart has similar benefits and limitations to those of the scikit-learn framework.

In our view, to augment existing frameworks and overcome their existing limitations, practitioners/researchers have to consider the following:

(a) What is the primary purpose of the modeling application? Specifically, is the goal to explain or predict? Shmueli et al (2010) presents an excellent discussion on the differences between explanatory and predictive modeling. Based on our experience, explanatory modeling is almost always focused on helping one to understand the data through a retrospective analysis (whether that entails supervised or unsupervised learning approaches). On the other hand, the purpose of predictive model is to be used for real-time/prospective analysis.

(b) Building on the previous point, practitioner should also determine if their modeling application can provide prospective predictions? For example, if we use

the blood glucose sensor example, a real-time application would inform a medical practitioner whether based on the current blood glucose level a patient should be considered diabetic? Alternatively, a prospective application would provide insights based on the time-series of glucose levels (or probabilities of being diabetic over time) when a patient can be in risk of being diabetic (i.e., time when blood glucose level is expected to exceed a pre-determined threshold in the future). As can be seen, this question opens the door for reliability and/or time-series modeling applications? This can be either built on the outcomes from machine learning models or be the primary focus of the analysis.

(c) In practice, the process of knowledge discovery and data mining (KDDM) is not a singular step where a machine learning algorithm is deployed. Decisions pertaining to data preparation (cleaning, transformation, and/or feature selection technique), algorithm selection (which machine learning model to use), and algorithm tuning (how to optimize the parameters of an algorithm for the purpose of prediction) will effect the prediction performance. To optimize the KDDM procedure, we recommend the use of designed experiments to study the impact of the factors/decisions-made and their interactions. This structured approach is more efficient than trial-and-error and produces better results than one-factor-at-a-time (OFAT) experimentation strategies. Note that the use of experimental designs to optimize the KDDM procedure in the literature is limited to orthogonal arrays (Taguchi methods), which are used to optimize either feature selection (see e.g., Kwak and Choi, 2002; Yang et al, 2008; Allias et al, 2014; Maji et al, 2016) or algorithm tuning (see e.g., Khaw et al, 1995; Packianather et al, 2000; Sukthomya and Tannock, 2005; Hsu et al, 2010; Kumar et al, 2015; Zare et al, 2019).

(d) In the case of a "binary" classification problem, what assumptions can be made pertaining to both classes? Specifically, let us consider the problem of using accelerometers for fatigue modeling. Is our understanding of fatigue and non-fatigue observations the same? If the answer is yes, then treating this as a binary classification problem is appropriate. If the answer is along the lines of "we have a better understanding of non-fatigue cases and we cannot capture all forms of fatigue", then one should treat this problem as either a one-class classification problem or statistical surveillance problem.

The reader should note that the questions/considerations above present a perspective on how models from industrial statistics and statistical process control can be used to augment the existing frameworks for machine learning. In our view, this allows for a more holistic approach that should allow for the *selection of the most appropriate tool from the Data Science toolbox*. We present a visual summary of the resulting framework in Table/Figure xyz. Allison/Miao, can you help with this?

After the figure/table, we need 1-2 paragraphs highlighting that CPS methodologies largely ignore everything but the algorithm/machine learning step

4. Wearables for Occupational Fatigue Management

Given that manufacturing applications have traditionally been a building block for theory development and evaluation in industrial statistics and statistical process control, we start by examining physical fatigue in manufacturing workplaces. First, it is important to note that the new paradigm(s) of advanced manufacturing/ Industry 4.0 are conceptually different from the computer-integrated manufacturing (CIM) paradigm of the 1990s. The goals of Industry 4.0 are to maximize the impact of a worker's skills by integrating him/her as as an integral component of the cyber-physical infrastructure; however, the end goal of CIM was to achieve a worker-less manufacturing environment (Gorecky et al, 2014). Additionally, recent publications from the *Ergonomics* and *Manufacturing Systems* literatures are showing that the transition to advanced manufacturing is increasing the workload on skilled labor (Brocal and Sebastián, 2015; Romero et al, 2016; Ferjani et al, 2017) and consequently, increasing fatigue levels.

4.1 Importance of the domain

In a recent survey of U.S. advanced manufacturing workers, 57.9% of respondents indicated that they were somewhat fatigued during the past work week (Lu et al, 2017). The high prevalence of occupational fatigue is problematic since they result in detrimental health outcomes (both short- and long-term), increase work-errors, and reduce workers' productivity (see Cavuoto and Megahed, 2017; Lu et al, 2017; Maman et al, 2017; Baghdadi et al, 2018, and references within). Moreover, Ricci et al (2007) estimated that the annual cost of lost production time due to occupational fatigue for U.S. workers exceeds $136 billion.

Wearable devices (hereafter wearables) provide the opportunity to "unobtrusively capture physical exposure information in the workplace, a problem that has challenged the field for several decades" (Schall Jr et al, 2018, p. 351). More specifically, infromation extracted from wearable devices can be used to: (Maman et al, 2017; Baghdadi et al, 2018; Schall Jr et al, 2018) (a) measure body angles; (b) quantify the intensity of workload/physical activity; and (c) capture a time-series of heart rate values. These are important predictors in attempting to quantify physical fatigue (Cavuoto and Megahed, 2017; Maman et al, 2017).

4.2 An illustrative example

The example presented in this chapter is part of the broader study published by Maman et al (2017) and further reported on in Baghdadi et al (2018, 2019). In this example, we utilize the freely available dataset Baghdadi et al (2019), which can be accessed at: https://github.com/fmegahed/fatigue-changepoint/tree/master/Data/Raw/. The data corresponds to a 3-hour manual materials handling (MMH) lab study, which involved a significant period of continuous walking that would allow for analysis of changes in gait over the duration of the session. The study was completed by fifteen subjects, who were instrumented with four small inertial measurement

units (IMU of size 51 mm $\times$ 34 mm $\times$ 14 mm) located at the ankle, hip, torso and wrist and coupled with a heart rate sensor. The IMU is a wearable device that has three sensors: (a) accelerometer, measuring a body's specific force, (b) gyroscope to measure the angular rate, and (c) magnetometer for measuring the magnetic field, which is useful for determining directions based on a global reference field. The IMU captured data at 51.2 Hz, and the heart rate sensor recorded data at 1 Hz throughout the task. Moreover, each of the IMU's sensors, captured data in the x, y, and z directions of the Cartesian coordinate system. We refer the reader to https://fmegahed.github.io/fatigue_case_jqt.html for information pertaining how this data can be pre-processed to extract features that can be used for fatigue modeling.

4.2.1 On the role of experimental design in data collection

In order to understand physical fatigue development, Maman et al (2017) designed a cross-sectional lab study using one-factor within subjects design. The one factor corresponded to the physical workload of the task, divided into three levels: (a) low, which involved an assembly task in a standing position at a workstation, (b) medium, which simulated a supply pickup and delivery task, and (c) high, which simulated a MMH task where participants picked cartoons of varying weights. The tasks were selected based on the range of tasks reported in the survey of Lu et al (2017). Each task level was performed in a separate 3-hours of continuous work session. Per section 4.2, we only focus on the MMH task in our discussion.

4.2.2 Data description

While each of the three IMU sensors can capture data at a high frequency rate in the x, y, and z local channels (i.e. relative to body part positioning), the data needs to be preprocessed prior to use. The goals of preprocessing are to: (a) transform the data to the global X, Y, and X frameworks, i.e. to make them independent of body positioning, (b) overcome the sensor drift problem associated with accelerometer data, and (c) generate/engineer features that can be used for either monitoring or prediction. The reader is referred to the thorough discussion of Baghdadi et al (2019) for more details on this step.

In general, the prediction/monitoring can two different aspects of the data: (a) adjusted and filtered acceleration data, or (b) features extracted from the acceleration data (e.g., statistical summaries of acceleration, jerk, stride length, stride height and stride duration). To assist the viewer to visualize the difference between both "levels" of data analysis, we provide an animated example showing ten consecutive strides in Figure 2. Then, in Figure 3, we show how three features (cumulative sum of stride length, height and duration) vary across participants over the course of the three hour experimental session.

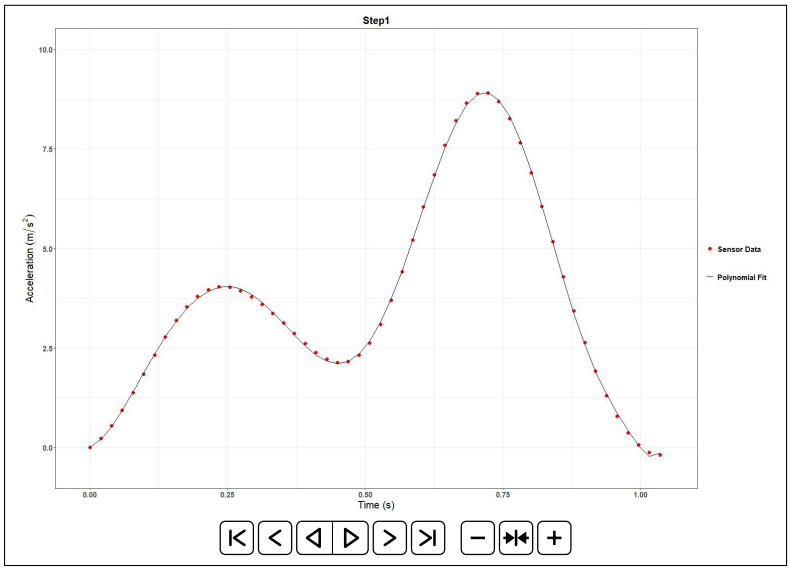


Fig. 2: Ten consecutive gait cycle segments, which can be viewed through interacting with this figure in *Adobe Acrobat Reader* (not through a web browser). The x-axis represents time, and the y-axis captures the magnitude of acceleration.

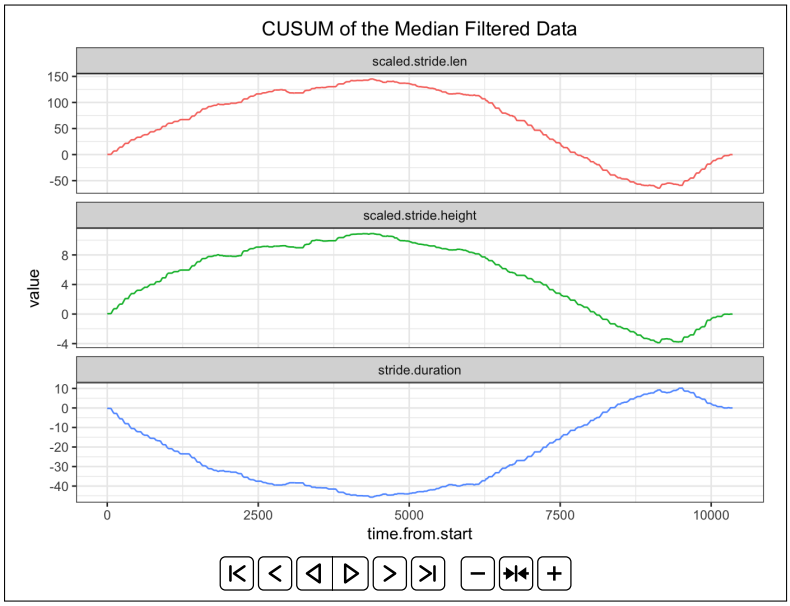


Fig. 3: The CUSUMs of stride length, height and duration for each participant, which can be viewed through interacting with this figure in *Adobe Acrobat Reader*.

4.2.3 Strategies to analyzing the data

Based on our discussion in Section 3, the scoping/framing of this dataset can lead to a number of different data analysis strategies. For example, Maman et al (2017) performed a preliminary analysis (using only eight subjects) using penalized linear and logistic regression to determine whether statistical features extracted from the five sensors' data (e.g., mean, standard deviation, max, min in a non-overlapping 10-minute time window) can be used to explain the variability in users' ratings of perceived exertion (RPE). Their test results showed that the RPE can be predicted with a geometric mean value, $g-mean = \sqrt{sensitivity \times specificity} = 0.88$. It is also interesting to note that their selected features for their model, based on LASSO, included features from all five sensors.

In a follow-up study with a larger sample size, Maman et al (2019) attempted to answer two main research questions: (a) whether five sensors are truly needed for modeling fatigue occurrence; and (b) whether the introduction of kinematic-driven features (e.g., stride length, height, and duration as well as back angle rotation) can improve the prediction. The results showed that: (a) with only one IMU, they can achieve a $g-mean = 0.85$ (compared to 0.87 when all sensors were used); and (b) kinematic features can improve the prediction performance of the model.

Deviating from the statistical/machine learning paradigm, Baghdadi et al (2019) examined whether the combination of multivariate change-point models and clustering can help in understanding how different subjects fatigue over the course of the experimental task. In their analysis, they chose to investigate changes in stride length, height and duration. Based on their analysis, the multivariate change point method revealed systematic changes in walking patterns. From the clustering analysis, the participants were divided into four groups, which reflected changes in both the magnitude and pattern of fatigue development. The participants adjusted their stride to mitigate the effects of fatigue. While maintaining the required pace, some participants elected to have shortened and faster slides, and others had longer but slower strides. This observation has not been observed in the ergonomics literature in part since: (a) wearables have not been thoroughly examined for modeling fatigue development, and (b) the limited number of papers that examined the use of wearables (or vision-based systems) utilized regression or classification approaches, which would not allow for the obtained insights.

4.3 Opportunities for statistical (process control) research

A

5. The Use of On-Board Vehicular Sensors to Capture Changes in Driver's Safety Performance

With the emergence of on-board vehicular sensors and technology, an increasing number of *Naturalistic Driving Studies (NDS)* have been initialized by research

teams around the worlds. NDS continuously records details of the driver, the vehicle, and surrounding environment via unobstructive on-board vehicular sensors and without experimental control (Regan et al, 2012; Eenink et al, 2014). The first large-scale NDS was the 100-Car NDS, pioneered by the Virginia Tech Transportation Institute (Dingus et al, 2006). Following this 100-Car study, NDS has also been explored in other countries, such as the second Highway Research Program (SHRP 2) in the United States and the European Naturalistic Driving Study (UDRIVE), as well as within specific target populations, such as teenagers, older drivers, and commercial truck drivers (Guo, 2019). The NDS supplies researchers with high-volume, high-resolution, and high-dimensional driving data, which create opportunities and pose challenges to data transformation techniques and existing statistical models.

5.1 Importance of the domain

Traditional truck crash prediction studies heavily rely on retrospective data that ultimately trace back to post crash police reports, interviews with witnesses and survivors, and vehicle inspection (Hickman et al, 2018; Stern et al, 2019). Although these post-crash data can be thorough and detailed, they inherently suffer from several limitations.

(a) In real-life data, crashes are *extremely rare* compared with non-crashes. The fatality rate of traffic accidents is 1.13 per 100 million miles driven in the United States, and the property-damage-only crash rate is 313 per 100 million miles driven (National Highway Traffic Safety Administration, 2017). Considering this rareness nature of crashes, hundreds of years of data are needed to achieve the sufficient statistical power to conclude on the difference of fatality rates between autonomous vehicles and human drivers (Kalra and Paddock, 2016; Guo, 2019).
(b) Post hoc reports, interviews, and inspections are subject to *recall bias* and *low-resolution* issue. These retrospective data sources were collected hours, days, or even weeks after the occurrence of the crash, so the accuracy and validity of these data are heterogeneous across different sources. In addition, as the data were collected by police officers, some critical factors in a meaningful time period leading up to the crash, such as distraction, were not regularly collected or reported due to various reasons (Dingus et al, 2011).
(c) Crashes are generally *under-reported*, particularly for minor- and non-injuries. Savolainen et al (2011) estimated that 25% of minor-injury and 50% of non-injury crashes were not reported in the data collected after accident, compared with 100% of fatality-involved crashes were reported. High under-reporting rates of non- and minor injury crashes may cause bias to statistical inference.

In view of these limitations, NDS has been developed by proactively and continuously collecting high-resolution driving data without obtrusive interference. Therefore, NDS has several strengths compared to traditional retrospective data sources. First, NDS collects safety critical events (SCE) such as hard braking events, which have a significantly higher incidence rates than crashes. These unsafe incidents were suggested to be indicative of near-crashes, collision, and crashes (Dingus et al, 2006;

Guo et al, 2010). Second, NDS records high-frequency and detailed traveling data, including but not limited to speed, GPS, and multidimensional accelerometers. In the meanwhile, unsafe incidents will be recorded once a kinematic threshold is triggered. Therefore, researchers can accurately trace back to several seconds prior to an incident and identify risk factors associated with that event. Since all events and data were collected by automated sensors and devices, recall bias, information bias, and under-reporting are minimized in NDS.

5.2 An illustrative example

Here we demonstrate an example of transforming NDS on-board sensor data, external obtaining weather and road geometry data from online APIs, fusing NDS data and online API data, and fitting statistical models based on fused data. For the purpose of illustration, we used a sample of 10-drivers NDS dataset collected by a commercial trucking and transportation company in the United States using on-board vehicular sensors.

5.2.1 Data

The data for this demonstrating example include five sources. Three of them were collected by the NDS study (real-time ping, SCEs, and driver characteristics), while two of them were from online API sources (weather and road geometry) .

1. *Real-time pings*: A small device was installed in each sample truck, and it will ping irregularly (typically every 1 to 10 minutes). Each ping will collect real-time data on the vehicle number, date and time, latitude, longitude, driver identification number (ID), and speed at that second. A sample of the ping data is shown in Table 2.
2. *SCEs*: Real-time time-stamped SCEs and associated GPS locations for were collected by the truck company and accessible to me as outcome variables. Specifically, three types of critical events were recorded: 1) hard brake (HB), 2) headway (HW), 3) collision mitigation (CM), and 4) rolling stability (RS). Once some kinematic thresholds with regard to the driving behavior were met, the sensor will be automatically triggered and the information of these SCEs (latitude, longitude, speed, driver ID) will be recorded. A sample of SCEs data is shown in Table 3.
3. *Driver demographics*: A table that includes the birth date of each driver was provided by the commercial truck company, and the age of the driver can be calculated. The driver's age table is shown in Table 4.
4. *Road geometry data from the OpenStreetMap API*: Two road geometry variables for the sample drivers will be queried from the OpenStreetMap (OSM) project: *speed limits* and *the number of lanes*. The OSM data are collaboratively collected by over two million registered users via manual survey, GPS devices, aerial photography, and other open-access sources (Wikipedia contributors, 2019). The OpenStreetMap Foundation supports a website to make the data freely available to the public under the Open Database License, and could

be queried using the osmar package in statistical computing environment **R**. A sample of road geometry data retrieved from the OpenStreetMap is demonstrated in Table 5.

5. *Weather data from the Dark Sky API*: weather variables, including precipitation intensity, precipitation probability, wind speed, and visibility, were retrieved from the Dark Sky API. The Dark Sky API allows the users to query historic minute-by-minute weather data anywhere on the globe (The Dark Sky Company, LLC, 2019). According to the official document, the Dark Sky API is supported by a wide range of weather data sources, which are aggregated together to provide the most precise weather data possible for a given location (The Dark Sky API, 2019). Among several different weather data APIs we tested, the Dark Sky API provides the most accurate and complete weather variables. A sample of the weather data retrieved from the DarkSky API is shown in Table 6

Table 2: ping data

ping_time	speed	latitude	longitude	driver
2015-10-23 08:00:00	0	33.94360	-118.1681	canj1
2015-10-23 08:08:10	0	33.94358	-118.1681	canj1
2015-10-23 08:09:26	5	33.94288	-118.1681	canj1
2015-10-23 08:22:58	4	33.97146	-118.1677	canj1
2015-10-23 08:23:12	8	33.97178	-118.1677	canj1

Table 3: safety critical events

driver	event_time	event_type
canj1	2015-10-23 14:46:08	HB
canj1	2015-10-26 15:06:03	HB
canj1	2015-10-28 11:58:24	HB
canj1	2015-10-28 17:42:36	HB
canj1	2015-11-02 07:13:56	HB

5.2.2 Data transformation and merging

The research question in this illustration example is *whether the truck driver's cumulative driving time is associated with unsafe driving behaviors*. To answer the question, we will transform the original ping data in the following ways to fit in different statistical models, the SCEs, age of the drivers, road geometry, and weather will be joined back to the transformed data using different combinations of keys.

(a) *Trips*: for each of the truck drivers, if the real-time ping data showed that the truck was not moving for at least 20 minutes, the ping data will be separated

Table 4: drivers

driver	age
canj1	46
farj7	54
gres0	55
hunt	48
kell0	51
lewr10	27
rice30	34
smiv	49
sunc	37
woow59	24

Table 5: Road geometry from the OpenStreetMap API

latitude	longitude	speed_limit	num_lanes
33.426	-84.144	55	2
33.635	-84.306	65	3
33.418	-84.143	55	2
41.801	-91.646	70	2
41.689	-92.073	70	2

Table 6: weather from the DarkSky API

ping_time	longitude	latitude	precip_intensity	precip_probability	wind_speed	visibility
2015-10-23 08:09:26	-118.1681	33.94288	0	0	0.21	9.82
2015-10-23 08:22:58	-118.1677	33.97146	0	0	0.22	9.81
2015-10-23 08:23:12	-118.1677	33.97178	0	0	0.22	9.81
2015-10-23 08:23:30	-118.1678	33.97233	0	0	0.22	9.81
2015-10-23 08:38:00	-118.1798	34.00708	0	0	0.24	9.81

into two different trips. A trip is short continuous driving intervals with a mean length of 1.8 hours.

(b) *Shifts*: the trips will be further divided into different shifts if the driver was not moving for at least eight hours. A shift is therefore a long driving time with intermittent short rests (20 minutes to 8 hours) within shifts.

(c) *Half-hour intervals*: since the length of the trips is heterogeneous (it varies from 5 minutes to more than 8 hours), which makes it difficult to conduct statistical analysis, we further divide trips into half-an-hour fixed intervals, which had one-to-one mapping to the trips. Some of the intervals were less half an hour as some trips were less than 30 minutes.

(d) *A proxy of driver fatigue*: we took the cumulative summation of trip time within a shift for each driver as the cumulative driving time, and used it as a proxy of driver fatigue. The rest time between trips were not counted in the cumulative driving time calculation.

After the transformed data sets were created, different sources of data sets were merged for statistical analyses. Our statistical analyses were based on two merged datasets: *merged half-hour intervals (MHHI)* and *merged shifts*.

- *MHHI*: SCEs were left joined to half-hour intervals if the two datasets had a common driver ID and the event time of SCE were between the start and end time of the half-hour interval. A binary variable of whether SCEs occurred and a count variable of the number of SCEs in each half-hour interval were created using the merged SCEs table. Driver's table was merged to MHHI using a common driver ID. Road geometry data were merged to the ping data by the latitude and longitude coordinates, and weather data were merged to the ping data by the latitude and longitude coordinates, date, and time. The road geometry and weather at ping level were then aggregated to MHHI by taking the mean of each variable.
- *Merged shifts*: SCEs, driver's age, weather, and road geometry were merged to transformed shifts data in a similar way as described in MHHI. The only difference is that the four tables were merged to transformed shifts, instead of transformed half-an-hour intervals.

5.2.3 Statistical models

To answer the question *whether the truck driver's cumulative driving time is associated with unsafe driving behaviors*, we first consider a Bayesian hierarchical logistic regression, where the response is $Y_i = 1$ if a crash occurred in a given segment/time period, and $Y_i = 0$ if no crash occurred. Logistic regression is the most popular statistical model used in traffic safety studies. This hierarchical will be performed based on the transformed MHHI data.

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(p_i) \\ \log\frac{p_i}{1-p_i} &= \beta_{0,d(i)} + \beta_{1,d(i)} \cdot \text{CT}_i + \sum_{j=1}^{J} x_{ij}\beta_j \\ \beta_{0,d} &\sim \text{i.i.d. } N(\mu_0, \sigma_0^2), \quad d = 1,2,\cdots,D \\ \beta_{1,d} &\sim \text{i.i.d. } N(\mu_1, \sigma_1^2), \quad d = 1,2,\cdots,D \end{aligned} \tag{1}$$

Here i is an index of the i-th observation and $d(i)$ is the driver's index of the i-th observation. we assume that each driver has a different baseline probability of having SCEs, which is the random intercept $\beta_{0,d}$. Besides, we also assume that the probability change of SCEs as a consequence of cumulative driving time (CT_i) is different among drivers, which is the random slope $\beta_{1,d(i)}$. we assume exchangeable priors for the random intercepts and slopes respectively. The parameters μ_0, σ_0, μ_1, and σ_1 are hyper-parameters with priors. Since we do not have much prior knowledge on the hyper-parameters, we assigned diffuse priors for these hyper-parameters as recommended by Gelman et al (2017).

$$\begin{aligned}
\mu_0 &\sim N(0, 5^2) \\
\mu_1 &\sim N(0, 5^2) \\
\sigma_0 &\sim \text{Gamma}(1,1) \\
\sigma_1 &\sim \text{Gamma}(1,1)
\end{aligned} \tag{2}$$

$x_{ij}, j = 1,2,,J$ are other covariates. For these covariates, we assigned flat priors $N(0,10^2)$ to their parameters.

Insert logistic regression results here.

Logistic regression considers the probability of an event during a given interval, but it ignores the frequency of events. Therefore, we considered a Poisson regression, another widely used statistical model in traffic safety studies, to account for the frequency of events. In a Poisson regression, the response is the number of events Y_i in a given time intervel T_i. We assume that the number of events has a Poisson distribution with the mean of λ_i.

$$\begin{aligned}
Y_i &\sim \text{Poisson}(T_i \cdot \lambda_i) \\
\log \lambda_i &= \beta_{0,d(i)} + \beta_{1,d(i)} \cdot \text{CT}_i + \sum_{j=1}^{J} x_{ij}\beta_j \\
\beta_{0,d} &\sim \text{i.i.d. } N(\mu_0, \sigma_0^2), \quad d = 1,2,\cdots,D \\
\beta_{1,d} &\sim \text{i.i.d. } N(\mu_1, \sigma_1^2), \quad d = 1,2,\cdots,D
\end{aligned} \tag{3}$$

The model setting is similar to the model in Equation 1, so we used identical priors and hyper-priors as described in Equation 2.

Insert Poisson regression results here.

Despite Poisson regression consider the frequency of SCEs in a given interval, it assumes that the intensity of events is a constant, which may not be true in real-life transportation practice. Here we presented a reliability model, a time-truncated non-homogeneous Poisson process (NHPP) with a power law process (PLP) based on the merged shifts data set. we aim to answer if SCEs occurred more frequently at early stages of shifts, towards the end of shifts, or does not show significant patterns.

Let $T_{d,s,i}$ denote the time to the d-th driver's s-th shift's i-th critical event. The total number critical events of d-th driver's s-th shift is $n_{d,s}$. The ranges of these notations are:

- $i = 1,2,\cdots,n_{d,S_d}$,
- $s = 1,2,\cdots,S_d$,
- $d = 1,2,\cdots,D$.

We assumes that the times of critical events within the d-th driver's s-th shift were generated from a PLP, with a fixed shape parameter β and varying scale parameters $\theta_{d,s}$ across drivers d and shifts s. In a PLP, the intensity function of the NHPP is $\lambda(t) = \frac{\beta}{\theta}(\frac{t}{\theta})^{\beta-1}$. The model is described in Equation 4.

$$
\begin{aligned}
T_{d,s,1}, T_{d,s,2}, \cdots, T_{d,s,n_{d,s}} &\sim \text{PLP}(\beta, \theta_{d,s}) \\
\beta &\sim \text{Gamma}(1,1) \\
\log \theta_{d,s} &= \gamma_{0d} + \gamma_1 x_{d,s,1} + \gamma_2 x_{d,s,2} + \cdots + \gamma_k x_{d,s,k} \\
\gamma_{01}, \gamma_{02}, \cdots, \gamma_{0D} &\sim \text{i.i.d. } N(\mu_0, \sigma_0^2) \\
\gamma_1, \gamma_2, \cdots, \gamma_k &\sim \text{i.i.d. } N(0, 10^2) \\
\mu_0 &\sim N(0, 10^2) \\
\sigma_0 &\sim \text{Gamma}(1,1)
\end{aligned}
\tag{4}
$$

The shape parameter β shows the reliability changes of drivers. When $\beta > 1$, the intensity function $\lambda(t)$ is increasing, the reliability of drivers is decreasing, and SCEs are becoming more frequent; when $\beta < 1$, the intensity function $\lambda(t)$ is decreasing, the reliability of drivers is increasing, and SCEs are becoming less frequent; when $\beta = 1$, the NHPP is simplified as a homogeneous Poisson process with the intensity of $1/\theta$. The $\theta_{d,s}$ is a scale parameter that does not reflect reliability changes.

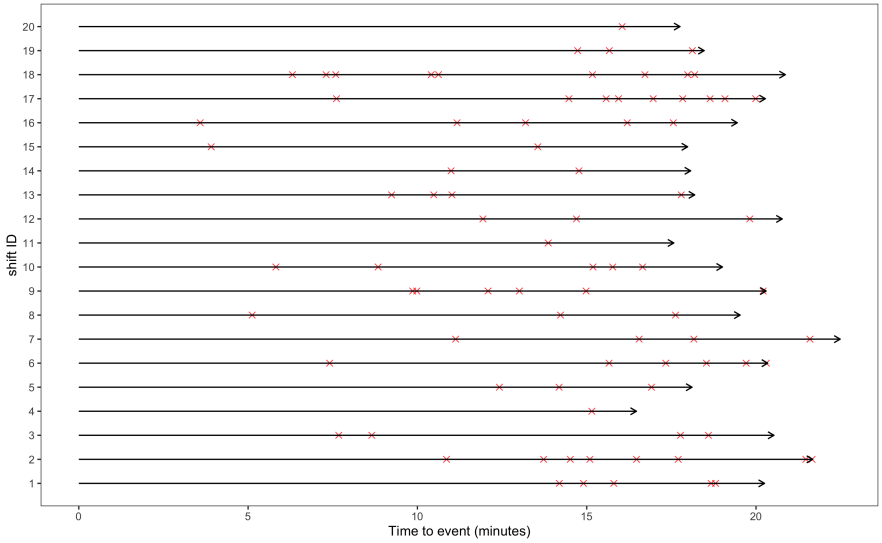


Fig. 4: An arrow plot of SCEs at different shifts

Insert NHPP results here.

5.3 Opportunities for statistical (process control) research

NDS data provide a unique opportunity to understand the risk factors of driving, but also present challenges in statistical analyses. The challenges are in two aspects: 1) high-resolution, high-dimension, and sparse data nature push for faster and less

computationally intensive estimation methods; 2) the fact that multiple SCEs can occur in one trip or shift requires more application of recurrent event models or reliability models.

Although SCEs are more frequent than crashes, they are still rare events compared to the total miles driven. Bayesian estimation, such as Hierarchical Bayes, is especially useful and powerful in the context of sparse data by placing informative or slightly informative priors on parameters and hyper-parameters. However, modern Bayesian estimation is empowered by Markov Chain Monte Carlo (MCMC), which is not scalable in the context of high-volume and high-dimensional NDS data. There are interesting progress on more efficient MCMC estimation strategies for high-volume or high-dimensional data, such as the Firefly Monte Carlo (Maclaurin and Adams, 2015), Pseudo-Marginal MCMC (Quiroz et al, 2019), and energy conserving subsampling Hamiltonian Monte Carlo (Dang et al, 2019), but these algorithms requires specialized statistical theories, self-defined coding and hyper-parameter tuning, and are applied among a limited number of statisticians.

Recurrent event models or reliability models fit naturally with the event generating process of NDS (Guo, 2019). With high-resolution NDS data, recurrent event or reliability models could uncover the patterns of non-homogeneous process, which are important for improving traffic safety. There are several studies that used recurrent event models, such as risk change-point for novice teenage drivers (Li et al, 2017, 2018) and random-effects frailty models (Chen and Guo, 2016). We presented an application of NHPP with PLP among truck drivers in this paper, which could be further improved by adding one more recovery parameter that accounts for the reliability recovery at each rest within a shift.

More opportunities on clustering, classification, and machine learning models. To Fadel: could you write a few sentences in this paragraph. Some high-level descriptions on opportunities to apply machine learning models to NDS data. –Miao

6. Biometric-Driven Computer Security

A

6.1 Importance of the domain

User authentication is a process used to verify that someone who is wanting access to a computing device (remote or otherwise) is who they claim to be. The primary goal of user authentication is to ascertain that only a legitimate user is granted access. The increasing popularity of personal devices and Internet services, and the sensitivity of information they often store (i.e, banking), prompts the need for secure authentication mechanisms.

Although various user authentication mechanisms are widely deployed by almost all devices and web-services, finding a secure mechanism is still remain a challenging problem. As almost all of the deployed mechanisms suffer from well-known security issues. For example, password which is the most widely deployed authentication approach nowadays tend to be insecure as the users normally pick

weak passwords (Florencio and Herley, 2007), or share their password (Singh et al, 2007), also password has intrinsic weakness in password leakage. Moreover, traditional biometrics such as fingerprints often have high error rates, susceptible to impersonation or spoofing attacks.

Behavioral biometrics is the study of using the human unique behavioral patterns to authenticate a person. For long time, handwriting and signature have been used as a mechanism to identify the users. Over the last decades, a lot of studies have been investigating utilizing the behavioral biometrics to authenticate the user to web-services or devices in order to improve the security of the authentication process.

With the increase of the data that can be collected from various sensors on mobile devices, smart-watches or keyboard and mouse on desktops/laptops a lot of work has been done on investigating the ability of using these data to identify the unique behavior of the user and use it to authenticate the user. Either as a stand-alone mechanism or combined with another mechanism. In contrast with passwords and traditional biometrics, behavioral biometrics cannot be forgotten, stolen or shared and are noninvasive.

6.2 An illustrative example

In this section, we provide details of Gametrics (Mohamed and Saxena, 2016), a game-based behavioral biometric system based on simple drag and drop games used to capture the unique user's cognitive ability as well as her unique mouse interactions.

6.2.1 Task Design

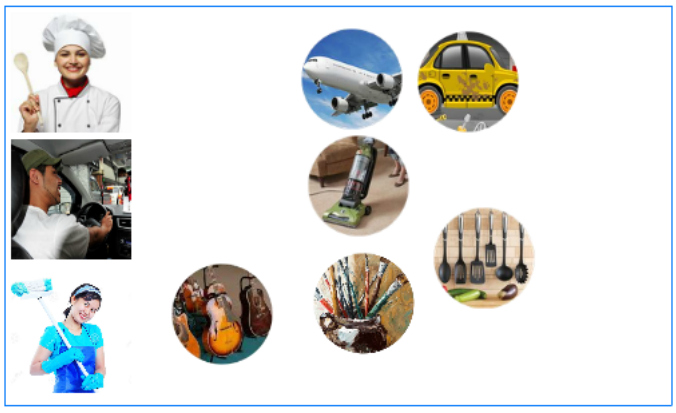


Fig. 5: Gametrics Challenge Instance. Targets, on the Left, are Static; Moving Objects, on the Right, are Mobile. The User Task is to Drag-Drop a Subset of the Moving Objects (Answer Objects) to Their Corresponding Targets.

Gametrics is based on simple drag and drop game challenge. A sample of the game challenges utilized in the study is shown in Figure 5. Each of the game challenges has three target objects and six moving objects. The user's task is to drag a subset of the moving objects (answer objects) to their corresponding target objects. In order to solve the challenge, the user needs to understand the content of the images, find the semantic relationship between the answer objects and the target objects, and drag the answer objects to their corresponding targets to solve the challenge.

Using Adobe Flash ActionScript3 and PHP, we implemented 6 challenges. These challenges are 500 $\times$ 300 pixels in size. The size of each of the moving object is 75 $\times$ 75 pixels and the size of the target objects is 90$\times$ 90 pixels. The challenge starts by placing the moving objects in random locations on the image. Then, each object picks a random direction in which it will move. The object continues moving in its current direction until it collides with another object or with the challenge border. A collision results in an object moving towards a new random direction.

The challenge starts when the user presses a Start" button on the screen center and ends when the user drags all the answer objects to their corresponding targets. While the user plays the game, all the user interaction with the challenge is recorded in a log. Specifically, the gameplay's log stores the objects locations, the mouse location and status (up/down) at each time interval.

6.2.2 Data Collection

In order to evaluate the applicability of identifying the user based on the way she interacts with the game challenges, we pursed data collection from human users. The study was approved by our university's Institutional Review Board.

We recruited the participants using Amazon Mechanical Turk (AMT) service. For the purpose of the study, we created three Human Intelligence Tasks (HITs) distributed over three days. The first HIT was created with 100 assignments to have 100 unique workers. We gathered 98 valid submissions until the HIT expired. On the next two days, we sent the participants emails asking them to participate in the follow-up study. In the first day, the participates were asked to solve 60 challenges and for the second and third day they were asked to solve 36 challenges. 62 participants performed the study on the second day and 29 performed the study on the third day.

In total, the participants successfully completed a total of 9076 challenges. The average time the participants took to complete a game challenge was around 7.5 seconds.

6.2.3 Feature Extraction

From each gameplay log, we extracted a total of 64 features. The features can be categorized into three categories: 1) features that capture the cognitive characteristics of individuals, 2)features that capture the mouse interaction characteristics of the participants, 3) features that are related to both of the cognitive abilities as well as mouse interaction.

The different mechanisms of solving the game challenges are related to the cognitive characteristics of individuals. We capture these characteristics based on the following features:

1. The time between the user pressing on the start button and the first mouse event and the time of the first click/drag. These timing measures capture the time the participates take to comprehend the challenge and start solving it.
2. The average, standard deviation, minimum and maximum of the times between each of the drops and the start of the next drag (these capture the time the user takes to find the next answer object).
3. The total time taken by the user to complete the challenge.

The mouse movement characteristics of the users are captured by following features:

1. The average, standard deviation, minimum and maximum of the speed and acceleration while the user is searching for an answer object and while the user is dragging the object.
2. The average, standard deviation, minimum and maximum of the duration between each two consecutively generated timestamps and the silence during move and during drag.
3. The average, standard deviation, minimum and maximum of time duration between reaching an object and clicking on it, and the time duration between approaching a target object and dropping an answer object on it.
4. The average, standard deviation, minimum and maximum of the angles between the lines that connect each 3 consecutive points in the mouse movement trajectory.

Other mixed features are also extracted that relate to both cognitive and mouse movement characteristics of the participants such as:

1. The total distance the mouse moved within a game challenge.
2. The average, standard deviation, minimum and maximum of the difference between the straight line connecting the start and the end of a move or a drag and the real distance traveled.
3. The average, standard deviation, minimum and maximum of the distance between a click and the object center, and a drop and the target center.

6.3 Data analysis

We utilized the Random Forest classifier in our analysis as it is efficient, can estimate the importance of the features, and is robust against noise Maxion and Killourhy (2010). Random Forest is an ensemble approach based on the generation of many classification trees, where each tree is constructed using a separate bootstrap sample of the data. In order to classify a new input, the new input is run down all the trees and the result is determined based on majority voting.

In our classification task, the positive class corresponds to the legitimate user's gameplay and the negative class corresponds to the impersonator (other user / zero-effort attacker). Therefore, true positive (TP) represents the number of times the legitimate user is granted access, true negative (TN) represents the number of times the impersonator is rejected, false positive (FP) represents the number of times the impersonator is granted access and false negative (FN) represents the number of times the correct user is rejected.

As performance measures for our classifier, we used false positive rate (FPR), and false negative rate (FNR). The FPR measures the security of the proposed system, i.e., the accuracy of the system in rejecting impersonators. The FNR measures the usability of the proposed system as high FNR mean that the system has high rejection rate of the legitimate users.. To make our system both usable and secure, ideally, we would like to have FPR and FNR to be as close as 0.

To improve the accuracy of the classification, We ran a program to find the subset of features that produces the best classification results, as using many features can cause over fitting of the classifier and therefore reduce the accuracy of the future prediction, thus removing some features may improve the accuracy. Therefore, we report the results obtained from selecting the best subset of features per user. Moreover, we study the identification of the user based on a single game challenge and on combining two challenges. As the average time for solving a challenge is around 7.5 seconds, we believe that utilizing two instances of the game challenges to identify the user is not much of an overhead. However, it may improve the accuracy by doubling the amount of captured interactions between the user and the challenges. The merging is done by averaging all the features from the two instances.

Table 7: Study Results: Performance for the classifier. We show the results of using a single challenge and merging of two challenges.

		FPR	**FNR**
Single	**Day 1**	0.06 (0.06)	0.02 (0.04)
	Day 2	0.09 (0.09)	0.07 (0.10)
	Day 3	0.07 (0.06)	0.07 (0.10)
Merge	**Day 1**	0.02 (0.05)	0.02 (0.05)
	Day 2	0.05 (0.09)	0.04 (0.09)
	Day 3	0.04 (0.06)	0.03 (0.05)

Inter-Session Analysis: As mentioned above, we collected data from 98 AMT workers during the first day of our data collection experiment. Each of them completed 60 challenges. We divided the collected data into 98 sets based on the users' identities (ids). In order to build a classifier to authenticate a user based on her gameplay biometrics, we defined two classes. The first class contains the gameplay data from a given user (to be identified), and the other class contains randomly selected gameplay data from other users.

Then, we divided the data into two sets, one for training and the other for testing. The first 40 gameplay instances of each participant and 40 gameplay instances of

the randomly selected set were used to train the classifier, while the other 20 are used for testing.

The results are shown in the first row ("Day 1")of each block in Table 7. We see that utilizing two gameplay instances is better than using a single instance. The best results are acquired by merging two challenge instances in which both the False Positive Rate and False Negative Rate = 2%.

Intra-Session Analysis: Our other main goal was to check the accuracy of the classifier over multiple sessions. As mentioned above, 62 AMT workers participated in the study in the second day and 36 participated in the study in the third day. For each of these users, we used the data of the gameplay of the previous day(s) to train the classifier and then we tested the classifier with the data collected in the next day(s).

The results are shown in the second and third rows ("Day 2" and "Day 3") in each block in Table 7. We find that the performance of the classifier degrades slightly compared to the first day, inter-session analysis. Also, we still found that merging two instances provides better results than using a single instance. The best results are again acquired by merging 2 instances. For the second day, False Positive Rate = 0.05 and False Negative Rate = 0.04 and for the third day False Positive Rate = 0.04 and False Negative Rate = 0.03.

Summary of Results

The results obtained from the classification models show that Gametrics is a viable form of behavioral biometrics. The results show that the classifier can identify the users and reject a zero effort attacker with a high overall accuracy, especially when two game instances are merged together.

6.4 Opportunities for statistical (process control) research

A

7. Conclusions

ABCD

Acknowledgments

This work was supported in part by: the National Science Foundation (CMMI-1635927 and CMMI-1634992); the Ohio Supercomputer Center (PMIU0138 and PMIU0162); the American Society of Safety Professionals (ASSP) Foundation; the University of Cincinnati Education and Research Center Pilot Research Project Training Program; the Transportation Informatics Tier I University Transportation Center (TransInfo); a Google Cloud Platform research grant for data management; and a Dark Sky grant for extended API access (i.e., they increased the number of possible queries per day). Dr. Megahed's research was also partially supported by the Neil R. Anderson Endowed Assistant Professorship at Miami University.

Supplementary Materials

ABC

References

Allias N, Megat MN, Noor M, Ismail MN (2014) A hybrid gini pso-svm feature selection based on taguchi method: an evaluation on email filtering. In: Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication, ACM, p 55

Aria M, Cuccurullo C (2017) bibliometrix: An r-tool for comprehensive science mapping analysis. Journal of Informetrics 11(4):959–975

Baghdadi A, Megahed FM, Esfahani ET, Cavuoto LA (2018) A machine learning approach to detect changes in gait parameters following a fatiguing occupational task. Ergonomics 61(8):1116–1129

Baghdadi A, Cavuoto LA, Jones-Farmer LA, Rigdon SE, Esfahani ET, Megahed FM (2019) Monitoring worker fatigue using wearable devices: A case study to detect changes in gait parameters. Journal of Quality Technology to appear

Börner K, Chen C, Boyack KW (2003) Visualizing knowledge domains. Annual Review of Information Science and Technology 37(1):179–255

Brocal F, Sebastián MA (2015) Identification and analysis of advanced manufacturing processes susceptible of generating new and emerging occupational risks. Procedia Engineering 132:887–894

Capizzi G (2015) Recent advances in process monitoring: Nonparametric and variable-selection methods for phase i and phase ii. Quality Engineering 27(1):44–67

Cavuoto L, Megahed F (2017) Understanding fatigue: Implications for worker safety. Professional Safety 62(12):16–19

Chen C, Guo F (2016) Evaluating the influence of crashes on driving risk using recurrent event models and naturalistic driving study data. Journal of applied statistics 43(12):2225–2238

Chen M, Qu J, Xu Y, Chen J (2018) Smart and connected health: What can we learn from funded projects? Data and Information Management 2(3):141–152

Colosimo BM, Cicorella P, Pacella M, Blaco M (2014) From profile to surface monitoring: Spc for cylindrical surfaces via gaussian processes. Journal of Quality Technology 46(2):95–113

Dang KD, Quiroz M, Kohn R, Tran MN, Villani M (2019) Hamiltonian monte carlo with energy conserving subsampling. Journal of Machine Learning Research 20(100):1–31

Del Castillo E, Colosimo BM, Tajbakhsh SD (2015) Geodesic gaussian processes for the parametric reconstruction of a free-form surface. Technometrics 57(1):87–99

Dingus TA, Klauer SG, Neale VL, Petersen A, Lee SE, Sudweeks J, Perez MA, Hankey J, Ramsey D, Gupta S, et al (2006) The 100-car naturalistic driving study.

phase 2: Results of the 100-car field experiment. Tech. rep., United States. Department of Transportation. National Highway Traffic Safety

Dingus TA, Hanowski RJ, Klauer SG (2011) Estimating crash risk. Ergonomics in Design 19(4):8–12

Eenink R, Barnard Y, Baumann M, Augros X, Utesch F (2014) Udrive: the european naturalistic driving study. In: Proceedings of Transport Research Arena, IFSTTAR

Ferjani A, Ammar A, Pierreval H, Elkosantini S (2017) A simulation-optimization based heuristic for the online assignment of multi-skilled workers subjected to fatigue in manufacturing systems. Computers & Industrial Engineering 112:663–674

Florencio D, Herley C (2007) A large-scale study of web password habits. In: Proceedings of the 16th international conference on World Wide Web, ACM, pp 657–666

Gelman A, Simpson D, Betancourt M (2017) The prior can often only be understood in the context of the likelihood. Entropy 19(10):555

Gorecky D, Schmitt M, Loskyll M, Zhlke D (2014) Human-machine-interaction in the industry 4.0 era. In: 2014 12th IEEE International Conference on Industrial Informatics (INDIN), pp 289–294, DOI 10.1109/INDIN.2014.6945523

Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of things (iot): A vision, architectural elements, and future directions. Future generation computer systems 29(7):1645–1660

Guo F (2019) Statistical methods for naturalistic driving studies. Annual review of statistics and its application 6:309–328

Guo F, Klauer SG, Hankey JM, Dingus TA (2010) Near crashes as crash surrogate for naturalistic driving studies. Transportation Research Record 2147(1):66–74

Haq A, Brown J, Moltchanova E (2014) Improved fast initial response features for exponentially weighted moving average and cumulative sum control charts. Quality and Reliability Engineering International 30(5):697–710

Haq A, Brown J, Moltchanova E, Al-Omari AI (2015) Improved exponentially weighted moving average control charts for monitoring process mean and dispersion. Quality and Reliability Engineering International 31(2):217–237

Hickman JS, Hanowski RJ, Bocanegra J (2018) A synthetic approach to compare the large truck crash causation study and naturalistic driving data. Accident Analysis & Prevention 112:11–14

Hsu WC, Yu TY, et al (2010) E-mail spam filtering based on support vector machines with taguchi method for parameter selection. Journal of Convergence Information Technology 5(8):78–88

IDC (2019) The growth in connected IoT devices is expected to generate 79.4zb of data in 2025, according to a new idc forecast. International Data Corporation. https://www.idc.com/getdoc.jsp?containerId=prUS45213219, [Online. Last accessed August 1, 2019]

Jeschke S, Brecher C, Meisen T, Özdemir D, Eschert T (2017) Industrial internet of things and cyber manufacturing systems. In: Industrial Internet of Things, Springer, pp 3–19

Jones-Farmer LA, Woodall WH, Steiner SH, Champ CW (2014) An overview of phase i analysis for process improvement and monitoring. Journal of Quality Technology 46(3):265–280

Kalra N, Paddock SM (2016) Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transportation Research Part A: Policy and Practice 94:182–193

Khaw JF, Lim B, Lim LE (1995) Optimal design of neural networks using the taguchi method. Neurocomputing 7(3):225–245

Kumar D, Gupta AK, Chandna P, Pal M (2015) Optimization of neural network parameters using grey–taguchi methodology for manufacturing process applications. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 229(14):2651–2664

Kwak N, Choi CH (2002) Input feature selection for classification problems. IEEE Transactions on Neural Networks 13(1):143–159

Lasi H, Fettke P, Kemper HG, Feld T, Hoffmann M (2014) Industry 4.0. Business & information systems engineering 6(4):239–242

Lee EA, Seshia SA (2017) Introduction to embedded systems: A cyber-physical systems approach. Mit Press

Leroy G, Chen H, Rindflesch TC (2014) Smart and connected health [guest editors' introduction]. IEEE Intelligent Systems 29(3):2–5

Li H (2017) Which machine learning algorithm should i use? The SAS Data Science Blog. https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/, [Online. Last accessed August 4, 2019]

Li Q, Guo F, Klauer SG, Simons-Morton BG (2017) Evaluation of risk change-point for novice teenage drivers. Accident Analysis & Prevention 108:139–146

Li Q, Guo F, Kim I, Klauer SG, Simons-Morton BG (2018) A bayesian finite mixture change-point model for assessing the risk of novice teenage drivers. Journal of applied statistics 45(4):604–625

Lu L, Megahed FM, Sesek RF, Cavuoto LA (2017) A survey of the prevalence of fatigue, its precursors and individual coping mechanisms among us manufacturing workers. Applied Ergonomics 65:139–151

Maclaurin D, Adams RP (2015) Firefly monte carlo: Exact mcmc with subsets of data. In: Twenty-Fourth International Joint Conference on Artificial Intelligence

Maji U, Mitra M, Pal S (2016) Imposed target based modification of taguchi method for feature optimisation with application in arrhythmia beat detection. Expert Systems with Applications 56:268–281

Maman ZS, Yazdi MAA, Cavuoto LA, Megahed FM (2017) A data-driven approach to modeling physical fatigue in the workplace using wearable sensors. Applied ergonomics 65:515–529

Maman ZS, Chen YJ, Baghdadi A, Lombardo S, Cavuoto LA, Megahed FM (2019) A data analytic framework for physical fatigue management using wearable sensors. Expert systems with applications (under review)

Maxion RA, Killourhy KS (2010) Keystroke biometrics with number-pad input. In: 2010 IEEE/IFIP International Conference on Dependable Systems & Networks (DSN), IEEE, pp 201–210

Mohamed M, Saxena N (2016) Gametrics: towards attack-resilient behavioral authentication with simple cognitive games. In: Proceedings of the 32nd Annual Conference on Computer Security Applications, ACM, pp 277–288

National Highway Traffic Safety Administration (2017) Traffic safety facts 2015: a compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system

National Science Foundation (2019) Cyber physical systems (CPS) — NSF 19-553. https://www.nsf.gov/pubs/2019/nsf19553/nsf19553.htm, [Online. Last accessed August 4, 2019]

Packianather M, Drake P, Rowlands H (2000) Optimizing the parameters of multilayered feedforward neural networks through taguchi design of experiments. Quality and Reliability Engineering International 16(6):461–473

Pantelopoulos A, Bourbakis NG, et al (2010) A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews 40(1):1–12

Paynabar K, Zou C, Qiu P (2016) A change-point approach for phase-i analysis in multivariate profile monitoring and diagnosis. Technometrics 58(2):191–204

Porter MF (2006) An algorithm for suffix stripping. Program 40

Psarakis S, Vyniou AK, Castagliola P (2014) Some recent developments on the effects of parameter estimation on control charts. Quality and Reliability Engineering International 30(8):1113–1129

Quiroz M, Kohn R, Villani M, Tran MN (2019) Speeding up mcmc by efficient data subsampling. Journal of the American Statistical Association 114(526):831–843

Regan M, Williamson A, Grzebieta R, Tao L (2012) Naturalistic driving studies: literature review and planning for the australian naturalistic driving study. In: Australasian college of road safety conference 2012, Sydney, New South Wales, Australia

Ricci JA, Chee E, Lorandeau AL, Berger J (2007) Fatigue in the us workforce: prevalence and implications for lost productive work time. Journal of Occupational and Environmental Medicine 49(1):1–10

Romero D, Bernus P, Noran O, Stahre J, Fast-Berglund Å (2016) The operator 4.0: Human cyber-physical systems & adaptive automation towards human-automation symbiosis work systems. In: APMS (Advances in Production Management Systems)

Savolainen PT, Mannering FL, Lord D, Quddus MA (2011) The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis & Prevention 43(5):1666–1676

Schall Jr MC, Sesek RF, Cavuoto LA (2018) Barriers to the adoption of wearable sensors in the workplace: A survey of occupational safety and health professionals. Human Factors 60(3):351–362

Scikit-Learn (2019) Choosing the right estimator. https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html, [Online. Last accessed August 4, 2019]

Shmueli G, et al (2010) To explain or to predict? Statistical Science 25(3):289–310

Singh S, Cabraal A, Demosthenous C, Astbrink G, Furlong M (2007) Password sharing: implications for security design based on social practice. In: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, pp 895–904

Stern HS, Blower D, Cohen ML, Czeisler CA, Dinges DF, Greenhouse JB, Guo F, Hanowski RJ, Hartenbaum NP, Krueger GP, et al (2019) Data and methods for studying commercial motor vehicle driver fatigue, highway safety and long-term driver health. Accident Analysis & Prevention 126:37–42

Sukthomya W, Tannock J (2005) The optimisation of neural network parameters using taguchis design of experiments approach: an application in manufacturing process modelling. Neural Computing & Applications 14(4):337–344

Teoh WL, Chong JK, Khoo MB, Castagliola P, Yeong WC (2017) Optimal designs of the variable sample size chart based on median run length and expected median run length. Quality and Reliability Engineering International 33(1):121–134

The Dark Sky API (2019) Data sources. URL https://darksky.net/dev/docs/sources, [Online; accessed 20-June-2019]

The Dark Sky Company, LLC (2019) Dark Sky API Overview. https://darksky.net/dev/docs, [Online; accessed 20-February-2019]

Wikipedia contributors (2019) Openstreetmap — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=OpenStreetMap&oldid=900226891, [Online; accessed 5-June-2019]

Wolfert S, Ge L, Verdouw C, Bogaardt MJ (2017) Big data in smart farming–a review. Agricultural Systems 153:69–80

Woodall WH, Montgomery DC (2014) Some current directions in the theory and application of statistical process monitoring. Journal of Quality Technology 46(1):78–94

Yang CH, Huang CC, Wu KC, Chang HY (2008) A novel ga-taguchi-based feature selection method. In: International Conference on Intelligent Data Engineering and Automated Learning, Springer, pp 112–119

Ye ZS, Chen N (2014) The inverse gaussian process as a degradation model. Technometrics 56(3):302–311

Zare M, Behnia N, Gabriels D (2019) Assessment of land cover changes using taguchi-based optimized svm classification approach. Journal of the Indian Society of Remote Sensing 47(1):45–52

Zhang M, Megahed FM, Woodall WH (2014) Exponential cusum charts with estimated control limits. Quality and Reliability Engineering International 30(2):275–286