Quality Engineering

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lqen20

A mixture model application in monitoring error message rates for a distributed industrial fleet

Bernat Plandolit, Ignasi Puig-de-Dou, Gráinne Costigan, Xavier Puig, Lourdes Rodero & José Miguel Martínez

To cite this article: Bernat Plandolit, Ignasi Puig-de-Dou, Gráinne Costigan, Xavier Puig, Lourdes Rodero & José Miguel Martínez (2023) A mixture model application in monitoring error message rates for a distributed industrial fleet, Quality Engineering, 35:3, 519-533, DOI: 10.1080/08982112.2022.2132866

To link to this article: https://doi.org/10.1080/08982112.2022.2132866

Published online: 01 Nov 2022.

QUALITY ENGINEERING
2023, VOL. 35, NO. 3, 519–533
https://doi.org/10.1080/08982112.2022.2132866

CASE STUDY

OPEN ACCESS

A mixture model application in monitoring error message rates for a distributed industrial fleet

Bernat Plandolit[a,b], Ignasi Puig-de-Dou[a,b], Gráinne Costigan[c], Xavier Puig[a,d], Lourdes Rodero[a,d], and José Miguel Martínez[a,d,e]

[a]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain; [b]Datancia SL, Barcelona, Spain; [c]HP Large Format Big Data Team, Sant Cugat del Vallès, Spain (During development of project); [d]Analysis of Complex Data for Business Decisions Research Group, Universitat Politècnica de Catalunya, Barcelona, Spain; [e]Public Health Research Group, University of Alicante, Alicante Spain

ABSTRACT
Remotely monitoring industrial printers for an unexpected increase of warning and error messages reduces equipment downtime and increases customer satisfaction. Directly tracking raw error messages rates during a given observation period poses some issues. Firstly, when a printer has not been used much during the observation period, its actual printing time is low. In this situation, even a small set of error messages can become an unexpectedly large rate of messages per printing hour. Secondly, classifying printers in error messages groups based on their rate (for instance, low, medium and high) and studying group changes over time, is useful in identifying potential problems. To overcome these issues, a nonparametric estimation method which simultaneously obtains empirical Bayes estimations of error messages rates and the number of error messages groups is used. This approach has been used in epidemiology, mainly in disease mapping research, but not in an industrial reliability context. The objective of our work is to show the application of the mixture model to real-time monitoring of printers' error message rates in a way that addresses the two issues mentioned above.

KEYWORDS
empirical Bayes; preventive maintenance; reliability; repairable systems; statistical process control

1. Introduction

With increased competition and changes in generational buying habits, many companies are diversifying from producing hardware and are increasingly focusing on hardware as a service - from car sharing and public bike schemes that have been introduced in cities worldwide, to intercity trains-as-a-service provided by Hitachi in the UK to personal systems as provided by Microsoft (Surface as a service, Microsoft (n.d.)) and Hewlett-Packard (HP device as a service HP (n.d.)). The success of these new business models, the brand reputation and the preservation of customer loyalty then depend on keeping running costs of machines to a minimum while providing high production times.

The majority of these systems are designed to undergo repairs and preventative maintenance in order to keep contractually required machine availability high. These requirements are also being introduced within cheaper consumer goods (i.e., new EU requirements for repairability on white goods (BBC 2009)). Defining maintenance schedules, capping warranty costs and maintaining a minimum reliability becomes increasingly complicated as different parts are replaced due to preventative or part failure interventions. These become fundamental to keep costs of maintaining service fleets down and production high.

This article focuses on a set of industrial printers from Hewlett-Packard, a major printer manufacturer, installed in printing factories all over the world. These printing factories process customer jobs and sell their printers' output as a product. Any time a printer stops for unplanned reasons, it increases the lead time of a customer job and drops its yield, reducing the profit margin. In order to minimize this situation, these printers generate warning and error messages to let

CONTACT Xavier Puig xavier.puig@upc.edu Universitat Politècnica de Catalunya, Estadística i Investigació Operativa. Edifici H, Avda. Diagonal, 647, 08028 Barcelona, Spain

the operator know in advance that something may not be working as expected.

The messages generated by printers are remotely monitored by a centralized Support Team. An unexpected increase of messages or the appearance of certain types of warnings may provide an early signal of a machine not being operated properly or getting close to failure. Quickly identifying these machines helps support engineers to focus on the equipment in need for follow-up, be it a call or a visit to the customer site. This reduces equipment downtime and increases customer satisfaction.

The challenge this support team faces is the large number of printers being monitored and the diverse type of warning and error messages they generate. Engineers need a prioritizing system that does a fast screening on key printer performance measures to indicate which machines they should focus on. This article describes an implementation to provide this prioritizing system based on the use of Statistical Process Control (SPC) techniques to monitor error and warning messages rates.

Statistical Process Control is the most widely used statistical method to monitor and improve the quality and productivity of manufacturing processes and service operations. SPC primarily involves the implementation of control charts, which are used to detect any changes in a process that may affect the quality of the result (Montgomery 2019). Machine operation monitoring, which can be done through a remote communication system, provides valuable information to detect unusual behavior in order to improve maintenance and supervision policies and anticipate repairs. All this makes SPC a powerful tool for the rapid identification of malfunctioning machines allowing the support team to focus on the equipment that needs the most follow-up, and this could translate into greater customer satisfaction (Stoumbos et al. 2000; Oakland and Oakland 2018).

The SPC methodology consists of two different phases. Phase I focuses on understanding the process variation over time, selecting the model and estimating its parameters. Phase II uses phase I results to monitor the process and to allow quality professionals to detect changes and act whenever necessary. The case at hand is not a standard application of an SPC. Its implementation is not straightforward and requires extra attention into phase I. Phase I parameters' estimates are key (Psarakis, Vyniou, and Castagliola 2014) and model-based clusters can be one of the appropriate techniques to characterize different behaviors, such as in and out of control (Chen, Birch, and Woodall 2016; Jones-Farmer et al. 2014). Finally, it is always recommended to evaluate the performance of the proposed method to detect changes in the monitored process (Jones-Farmer et al. 2014; Psarakis, Vyniou, and Castagliola 2014).

HP monitors its printers' performance using the error and warning message counts gathered in a week. Directly tracking raw error messages rates during an observation period poses some major challenges. First, the information conveyed by printers via error rates is very variable. Second, printers error rates tend to be alike in similar operating conditions. Machines being operated in more extreme settings, with more challenging media or with less experienced operators will generate a higher number of error messages compared to machines being operated in less strict conditions or kept in better shape.

As the monitoring process is based on message counts per time, a natural candidate for such a task is U-charts (Montgomery 2019). They are used to monitor a process based on the number of non-conformities it generates when the observation unit, in our case the printing time in a week, changes from observation to observation. However, this approach comes with some limitations. On the one hand, weekly printing times are very variable between printers and within different weeks of the same printer. Reported weekly error rates (warning and error messages per hundred hours) vary a lot from week to week and U-chart control limits are wide and ragged going from small to very large values depending on the week printing time. The problem is especially visible in low printing time weeks, where message rates and control limits can grow substantially. On the other hand, using weeks as the observation period makes for a lengthy phase I, as a sizeable number of observations are needed for a robust phase I implementation. This problem worsens as printers do not always report in all weeks due to communication issues.

A possible option to improve phase I and reduce its length could be sharing information among similar printers when implementing their U-charts. Printers could be first clustered based on their observed overall performance (e.g., overall error rates and/or average printing times). Control chart parameters could then be estimated by pooling observations from printers included in the same cluster. This approach, although feasible, still has the limitation that a given printer may have different behaviors in different weeks depending on its operating conditions on that week. It also fails as it could potentially include out of control observations when building the chart in phase I.

The methodology we describe in this paper tries to address the above-mentioned issues by exploiting information sharing among printers that behave similarly in a given week. The goal being to improve phase I making it shorter and more robust. This idea fits nicely with HP engineer's perception that printers tend to fail in similar modes in the field depending on their operating conditions such as media usage, job type or operator skillfulness.

The first problem mentioned above of variable printing time, especially severe in weeks of low activity and high warning message rates, can be understood as a "credibility" issue. A warning message rate from a machine that did not print much during the observation period should be weighted less when assessing its relevance than a similar error rate coming from a machine that has been used much more during a week. To minimize this problem a Bayesian method could be used (Carlin and Louis 2000). The empirical Bayes (EB) framework provides a shrunken estimation of the error rate for a specific printer depending on its printing time per week. The empirical Bayes estimation is a compromise between the information provided by the printer observed error rate and the error rate from all printers. When the printing time is large, the Bayesian method weighs more to the information provided by the printer and the estimation is closer to its observed error rate. On the other hand, when the printing time is small it weighs more the information coming from the rest of printers and the estimation for this printer becomes closer to the error rate of all printers (Carlin and Louis 2000).

To address the second issue and exploit printers' operating similarities, one has to consider that printers will tend to fail similarly on alike operating conditions such as customer usage pattern, equipment age or working conditions. This behavior may be used to assign printer observations to groups of high, medium or low error message rates. Relative changes on printer messages rate or printer changes to a higher message rate cluster can be a signal of problems. A traditional method is to use percentiles to classify observations from printers based on their error rate. However, this cutoff is rather arbitrary and there is no guarantee that the percentile classification can validly detect high, medium and low-error rate printers. Böhning (1999) and Schlattmann and Böhning (1993) proposed a mixture model based on a nonparametric estimation method which simultaneously obtains the EB estimations of error rates and the number of risk groups.

As far as we know, this approach has been used in epidemiology, mainly in disease mapping research, but not in an industrial reliability context. For instance, Rattanasiri et al. (2004) used the mixture model to investigate the geographical and temporal distribution of malaria in Thailand. Benach et al. (2004) used it to identify high mortality risk areas clusters in the south-west of Spain.

The objective of our work, presented in this article, is to show the application of the mixture model to real-time monitoring of printers' error message rates in a way that addresses the two issues mentioned above.

Further details on the implementation are given within Section 2 as well as a detailed overview of the data set. The results and the deployment of the solution with the commercial environment are presented in Section 3. The discussion follows in Section 4.

2. Materials and methods

2.1. Case study

This case study is based on data from a large HP commercial printer fleet installed worldwide. This fleet is constantly monitored to provide a full maintenance and repair service and secure high productivity for HP clients.

The firmware of HP printers is designed to be very verbose. A single system may launch several messages or different systems may send concurrent messages all coming from a common root cause. Providing the HP service team on a regular basis with a prioritized list of printers based on their warning message behavior allows for the monitoring of the stability of the entire installed base. To study this problem, we sampled 439 printers from a single product line. Choosing such a subset was done in order to limit variability to a single printer family.

2.2. Dataset

HP printers are by design verbose in reporting when operating parameters diverge from normal constraints. These error messages may or may not indicate that an actual failure will occur, however a large accumulation or error messages may indicate the machine is not operating within design limits. The dataset used for the analysis includes the total number of error messages classified as severe and the total printing time for each printer grouped by calendar week. Severe error messages, contrary to advisory ones, always require operator intervention in order for the printer

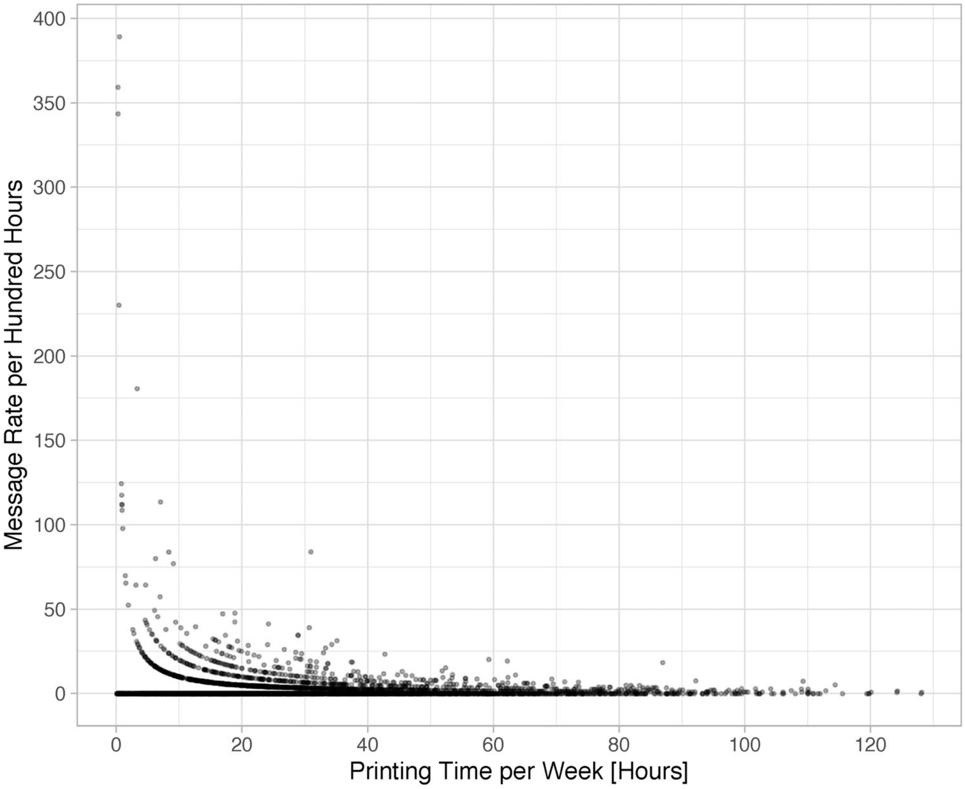


Figure 1. Printer message rate (warning messages per hundred printing hours) versus printing time in hours. Every point represents a machine-week observation. Low printing times show higher variance in reported rates. Only a selection of printers is shown to aid visualization.

to continue operating. These entries are referred to as printer-week observations throughout this article.

Further information on the origin of the error message was not taken into account in the analysis due to the large diversity of messages and root causes that could lead to weakened insights. Only messages generated when the machine is printing were accounted for, as machines can report error messages while idle but these are unlikely to affect productivity.

Finally, a given situation may generate several warning messages all related to the same root cause. To avoid repeatedly counting an already identified situation, messages that belonged to the same cause were grouped together and counted as one.

Data was grouped in weekly buckets to align with the weekly monitoring process that the HP Support Team already had in place.

The final data set included weekly printing time and number of relevant warning messages for 439 printers during 27 consecutive weeks happening from May 29th to November 27th 2017. Weeks are labeled using the ISO week convention from week 22 to 48. The total number of machine-weeks was finally 11,074 as not all machines reported in all weeks. All printers were working in production environments, i.e., clients sites.

2.3. Exploratory data analysis

The average weekly printing time per printer for the chosen sample is 30.5 hours. There are some extreme usage weeks where a printer can print up to more than a hundred hours per week. However, most of the printer-weeks in the sample, 75%, did not print more than 40 hours per week.

The average raw message rate is 3.5 messages per hundred hours. 75% of the printer-week raw message rate lies under 3.4 messages per hundred hours, indicating a large number of weeks with very low values. Only a few weeks have high message rates, up to a maximum of nearly 1,200 per hundred hours in a very extreme case. These high rates may indicate a highly disruptive week or a very low printing time during the given week, as is the mentioned extreme case. This wide range of variability between printing times and messages is characteristic of the data set at hand. This variability may lead to artificially enlarged warning message rates and justifies the usage of the Bayes modeling approach to minimize it.

Figure 1 shows this point for a subset of printers. Message rate variability increases with shorter weekly printing time in the graph as the observed number of messages is divided by a smaller amount compared to weeks with larger printing times.

2.4. Basic likelihood approach

Let O_{it} be the number of observed error messages and T_{it} be the printing time of the i-th distributed machine, $i = 1, \ldots, N,$ in the t-th week, $t = 1, \ldots, T$. We assume

$$O_{it}|\theta_{it} \sim f(O_{it}|\theta_{it}T_{it}), \tag{1}$$

where $f(O_{it}|\theta_{it}T_{it})$ is a Poisson distribution with failure rate θ_{it} and mean $\theta_{it}T_{it}$ of the i-th machine in week t. Under this assumption the maximum likelihood (ML) estimation of θ_{it} is

$$\hat{\theta}_{it}^{ML} = \frac{O_{it}}{T_{it}}. \tag{2}$$

The variance of $\hat{\theta}_{it}^{ML}$ is $\frac{\theta_{it}}{T_{it}}$, large for printer-weeks with small printing time and small in printer-weeks with large printing time, making comparison between printer-weeks performance difficult (Lawson et al. 2000).

2.5. Empirical Bayes approach

We use the empirical Bayes approach for shrinking and clustering machines with respect to their weekly message rate (Schlattmann and Böhning 1993; Böhning, Dietz, and Schlattmann 2000; Böhning 1999). We assume the existence of k possibly unknown clusters with a nonparametric distribution for θ (Böhning, Dietz, and Schlattmann 2000):

$$F = \begin{bmatrix} \phi_1 & \cdots & \phi_k \\ p_1 & \cdots & p_k \end{bmatrix}, \tag{3}$$

where ϕ_j and p_j, $j = 1, \ldots, k,$ indicate the failure rate and the probability of belonging to each cluster respectively.

These parameters are estimated using maximum likelihood through the marginal distribution for O_{it}. This is a nonparametric mixture distribution based on weighted sums of Poisson densities for i-th printer and t-th time period (Böhning, Dietz, and Schlattmann 2000):

$$O_{it}|F \sim \sum_{j=1}^{k} f(O_{it}|\theta_{it} = \phi_j)p_j, \text{ with } \sum_{j=1}^{k} p_j = 1 \text{ and } p_j \geq 0, \forall j. \tag{4}$$

Once we have estimators of k, ϕ_j and p_j, $j = 1, \ldots, k,$ we obtain the empirical Bayes estimation of the θ_{it}, the message rate of the i-th printer in the t-th week using the posterior expectation:

$$\hat{\theta}_{it}^{EB} = \frac{\sum_{j=1}^{\hat{k}} \hat{\phi}_j f(O_{it}|\theta_{it} = \hat{\phi}_j)\hat{p}_j}{\sum_{j=1}^{\hat{k}} f(O_{it}|\theta_{it} = \hat{\phi}_j)\hat{p}_j}. \tag{5}$$

We move forward to classify printer i-th in week t-th in a certain cluster depending on its message rate for that week. Let Z_{itj} be a random variable with value 1 if the i-th machine in the t-th week belongs to cluster j, and 0 otherwise. We can obtain the posterior probability:

$$Pr(Z_{itj} = 1|O_{it}, \hat{F}) = \frac{\hat{p}_j f(O_{it}|\theta_{it} = \hat{\phi}_j)}{\sum_{j=1}^{\hat{k}} \hat{p}_j f(O_{it}|\theta_{it} = \hat{\phi}_j)}. \tag{6}$$

The i-th machine in the t-th week will be classified in the cluster with the highest probability.

The maximum likelihood estimation can be easily obtained using the R package CAMAN (Computer-assisted Analysis of Mixtures (Schlattmann et al. 2016)). Using the nonparametric likelihood estimator (NPMLE) to define the number of components k can lead to an overestimation (Schlattmann and Böhning 1993; Böhning, Dietz, and Schlattmann 2000; Böhning 2003; Schlattmann 2009, 2003). A general strategy to mitigate against overfitting is to compare the NPMLE estimation with mixture models having fewer components using the likelihood ratio test (LR) (Böhning and and others 1999; Böhning 2003; Schlattmann 2009). In this case, the LR does not follow a chi-square distribution (Böhning et al. 1994), and a bootstrap approach is used to provide a reasonable approximation of the true null-distribution (Böhning 1999; Böhning 2003; Schlattmann 2009). In addition, to penalize model complexity in the model comparison, we use the Bayesian Information Criterion (BIC) (Carlin and Louis 2000; Schwarz 1978; Leroux and Puterman 1992).

2.6. Statistical process control proposal

Using the shrunken message rates provided by the presented model, we propose a statistical process control (SPC) methodology to monitor the behavior of the printers. On a weekly basis, data are processed and a list of printers showing above normal message rates is provided to the service team allowing them to focus resources where necessary. As stated in every SPC methodology, we implement a phase I, monitoring system definition, and a phase II, monitoring.

2.6.1. Phase I monitoring system definition

In phase I, we define the entire monitoring system that will be used to detect the printer's message rate's

Table 1. Likelihood ratio test.

Components k	LR	LR boostrap 95% Percentile	BIC
1			29,248.49
2	2,263.54	2.61	27,003.57
3	1,864.16	3.57	25,158.04
4	246.18	4.32	24,930.48
5	61.96	3.78	24,887.15
6	12.17	4.21	24,893.60

Likelihood ratio test (LR), 95% percentile of the LR distribution under the null hypothesis (model with k – 1 components) versus alternative hypothesis (model with k components) obtained with 2,500 bootstrap replications and Bayesian Information Criterion (BIC).

behavior. Using the described data (printing time and messages reported during printing), we obtain the estimation of the number of clusters, their failure rates and the proportion of printer-weeks that belong to each one (i.e., we estimate the F distribution). From this estimation, we obtain an empirical Bayes warning message rate for each printer-week (equation 5). After that, printer-week entries are grouped into each cluster based on their posterior probabilities (equation 6). Equations (5) and (6) are formulae used to smooth future weekly message rates reported and to classify them into each corresponding cluster, a crucial step in phase II.

2.6.2. Phase II monitoring

Once the monitoring system is defined, we make use of equations (5) and (6) to shrink the observed failure rate of every newly observed printer's data point and to classify it accordingly to the most suitable cluster. Once we have this, the way to identify machines with abnormal messaging rate is:

1. Identify printers that have been allocated to the highest message rate cluster or to clusters considered to have unacceptable message rates - i.e. those printers whose estimated message rate is very high. To do this we define a control chart named Posterior Error Rate Cluster (PERC) plot that shows the shrunken message rate of the printer each week and its cluster classification.
2. Identify those printers that have shown an important growth of the estimated message rate with respect to the previous week. That can either happen when a printer changes to a higher message rate cluster or simply has an important increase compared to its regular behavior. The level of importance at which action needs to be taken has to be determined by the experience of support engineers. To do this, in addition to the PERC graph, we compute and plot the percentage of increase in the warning rate from the previous week. We define this indicator as the Weekly Percentage of Change (WPC):

$$WPC(t) = \left(\frac{\hat{\theta}_t - \hat{\theta}_{t-1}}{\hat{\theta}_{t-1}} \right) \times 100. \qquad (7)$$

Table 2. Cluster messaging rates.

$\hat{\phi}_j$	$\hat{p}_j$
0.0	0.202
1.4	0.623
7.2	0.150
23.6	0.025
87.9	0.002

Messages per hundred hours ($\hat{\phi}_j$) and proportion ($\hat{p}_j$) of printer-weeks included for the $k = 5$-cluster model.

These two measures consider respectively the situation in which a printer has a clear poor behavior for a given week and, more subtly, when a significant increase in the message is detected for a printer that could anticipate an irregular performance. Both the PERC and the WPC are used graphically to monitor the fleet as shown in Section 3.2.

3 Results

In the following subsections we present the results of the outlined model to the HP printer fleet (described in Section 2.2). Section 3.1 presents the definition of the clusters for this data set. Section 3.2 details how this can then be used to monitor the state of a printer. Section 3.3 describes the actual model deployment and finally, Section 3.4 assesses the methodology performance.

For the sake of reproducibility, a simulated data set and the R code required to generate the article results is available at https://doi.org/10.5281/zenodo.5675789.

3.1. Phase I monitoring system definition

As mentioned in Section 2.5, the number of clusters may be overestimated. Table 1 shows the likelihood ratio test, BIC and the 95% percentile value of the LR distribution under the null hypothesis (model with $k-1$ components versus alternative hypothesis of model with k components) obtained with 2,500 bootstrap replications. The results show that the model with 5 components is not rejected at the 5% significance level (LR value from the original sample of 61.96 compared with the bootstrapped critical value of 3.78) and in addition, is the model with the lowest BIC value (24,887.15). However, the question remains whether a 6 component model is recommended over the 5 components. To address this we perform a forward strategy obtaining the LR between both models.

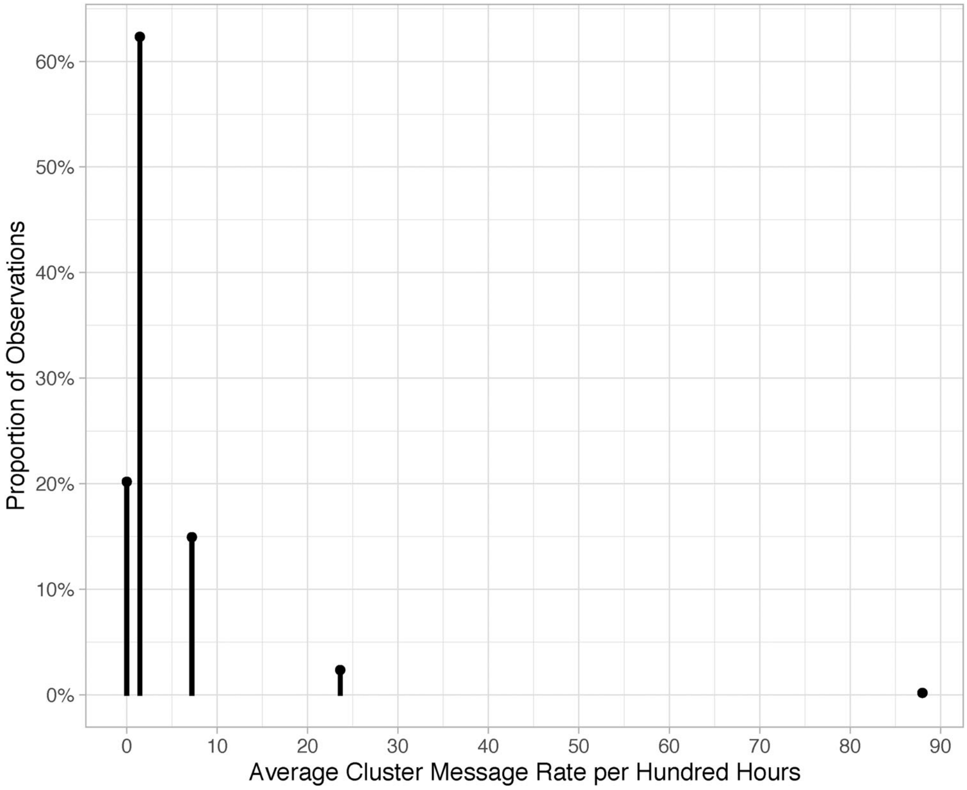


Figure 2. Nonparametric maximum likelihood estimation of F distribution for the $k = 5$ cluster model.

The observed LR obtained was 12.17 compared with a value 4.21 for the 95% bootstrapped critical value. This favors the choice of the 6 components model. However, the BIC for this model is higher than BIC for the 5 component model (24,893.60 versus 24,887.15), so we choose the most parsimonious 5 component model as the final one. The results for the final model are shown in Figure 2 and Table 2.

As we see in Table 2 and Figure 2, there is a no-message cluster encompassing 20.2% of the printer-weeks and a cluster with an average rate of 1.4 messages per hundred hours including 62.3% of the printer-weeks. On the high rate side, 2.5% of the printer-weeks have an average message rate of 23.6 per hundred hours and a marginal 0.2% have a very high average of 88 per hundred hours. Figure 2 also shows that clusters 1 and 2 are close to each other. These two levels can be considered as the normal or acceptable levels of message rate for a printer. Cluster 3 is slightly higher whereas clusters 4 and 5 are far apart. Cluster 3 can be classified as "alert" level and clusters 4 and 5 as "danger".

Figures 3 and 4 show the consequences of taking a Bayesian approach. The left graph on Figure 3 shows the observed message rates per hour and the one on the right the shrunken ones. Figure 4 plots the message rates from two random samples ($n = 50$) of printer-weeks from cluster 2. The graph on the left plots 50 printer-weeks with a low printing time per week and the one on the right 50 printer-weeks with a high printing time per week split by the median printing time in cluster 2 (24.8 hours). The graphs show a larger shrinkage from the maximum likelihood estimation to the empirical Bayes one for the lowest printing times per week rates (from 0.014 to 24.8 hours) and a smaller one for the highest printing times per week rates (from more than 24.8 to 129 hours).

3.2. Phase II monitoring

Once the monitoring system has been defined, phase II starts with the ongoing process of "monitoring" printers using the updated set of observations provided by the printers every week.

The data gathered from the printer's weekly operation is used to estimate its shrunken rate of warnings per unit time and classify it in its corresponding cluster. Then, based on the rules defined in Section 2.6.2, printers that deserve attention are identified in two ways: those that in a particular week have fallen into cluster 4 or 5, and those that have a high WPC that may lead to a change of cluster with higher message rate. These two criteria are used to list all printers

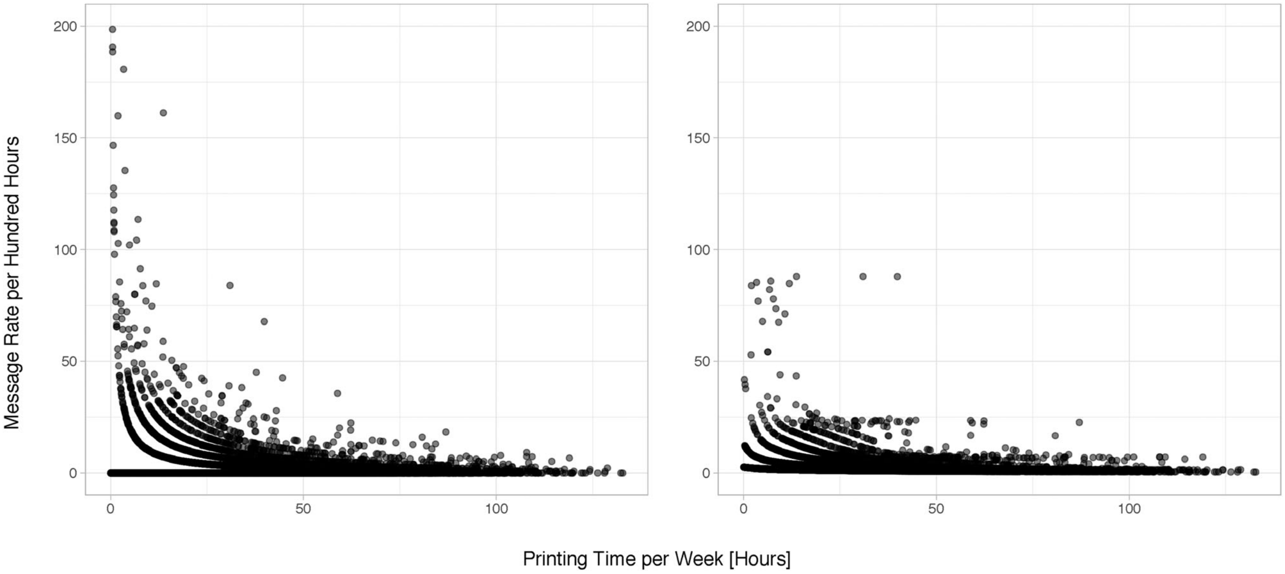


Figure 3. ML estimation (left) and EB estimation warning rates (right) versus printing time per week. Errors per hundred hours cut to 200 to help visualization. Only selected printers shown to ease visibility.

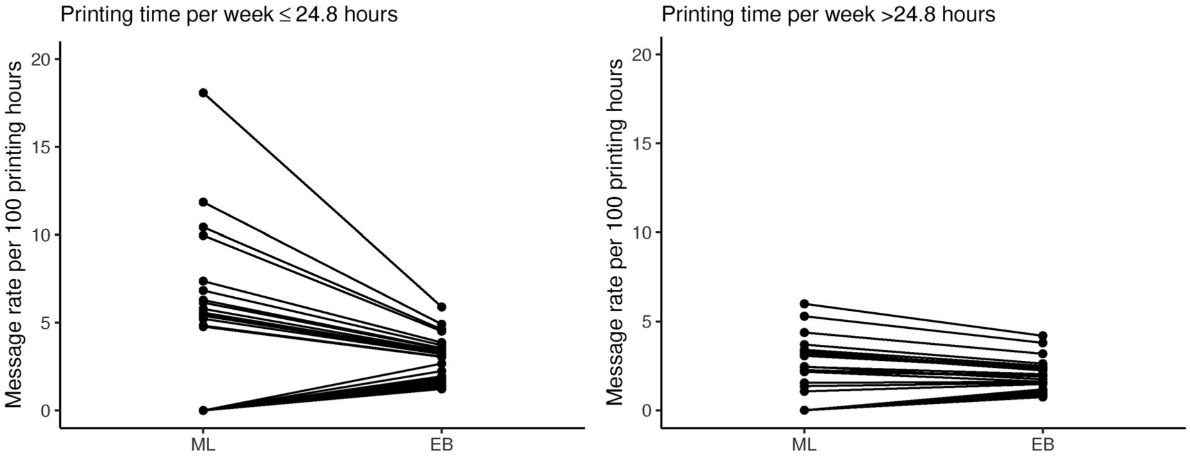


Figure 4. ML error rate estimation paired with the EB error rate for two random samples of 50 printers from cluster 2. The graph on the left shows 50 random printer-weeks with low printing time in cluster 2 (less or equal than 24.8 hours in a week). The graph on the right shows 50 random printer-weeks with high printing time (more than 24.8 hours in a week).

that deserve special attention by the HP support team that week. In order to help assess a printer message rate evolution the PERC plot showing the shrunken warning rate progress, and the WPC indicator plot showing the relative change in warning rates, can be used.

In the PERC plot (Figures 5 and 6, upper graph) the blue, green, yellow, orange and red lines indicate the average message rate of clusters 1 to 5. Each point in the graph represents the printer shrunken rate for the given week. Its color reveals the cluster into which the printer has been classified that same week following the cluster coloring convention. Finally, the grey bars show the weekly printing time for the machine. In the WPC indicator plot (Figures 5 and 6, lower graph) every point represents each week's increase or decrease in the shrunken message rate from the previous week - see formula (7). The color of the point indicates the cluster the printer has been assigned that week, similar to the PERC plot. These plots can be of great help for Support Engineers to decide whether the printer needs any action or not. Figures 5 and 6 provide two real examples of printer behavior regarding the empirical Bayes weekly warning message rates, PERC plot, and WPC plot.

The printer depicted in Figure 5 tends to be in the large printing time range (70 hours a week on average) and in the low warning rate range (1.5 warnings

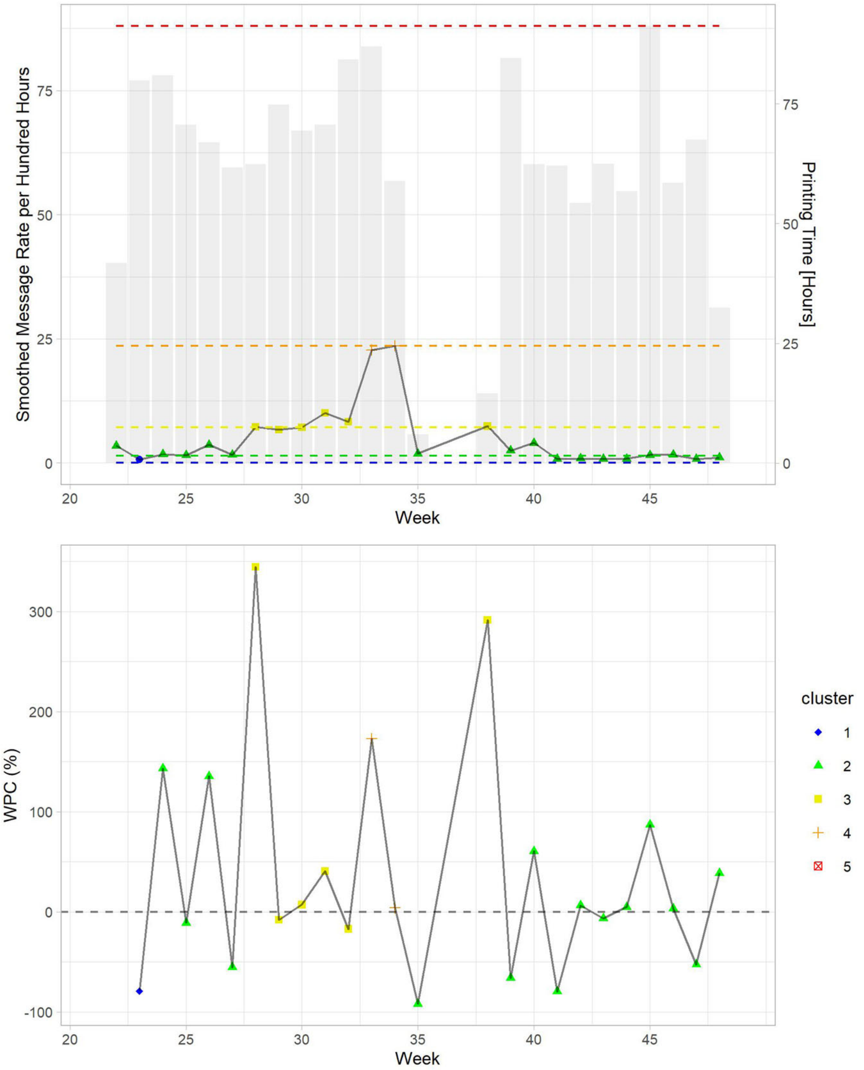


Figure 5. PERC (top) and WPC (bottom) graphs for phase II monitoring. PERC graph plots the empirical Bayes (shrunken) warnings rate versus week for a given printer. The WPC plots the change in the shrunken rate from week to week. A drop in performance following an increase in warnings rate can be seen in weeks 35 to 37.

per hundred hours printing). This is the reason it is classified in clusters 1 and 2 for several weeks and up to week 28. On week 28, the largest-than-usual warning rate classifies it into the third cluster showing a large jump in the WPC metric for that same week. From that week up to week 32 the number of warning messages grows without a relevant change in printing time. This makes the smoothed warning rate to be between 6 and 10 per hundred hours and classifies the machine into cluster 3. The situation degrades even further in weeks 33 and 34 when warning rates jump to 21 messages per hundred hours and the machine is classified in the high-risk cluster 4. The printer shows a major drop in availability on week 35 and a couple of weeks of no-service afterwards related to a machine malfunction. The machine was put back into operation on week 38, and after a rough start that week returned to regular performance levels from then on. This is a clear example of a situation that could have been avoided if the monitoring system were available. Support engineers could have checked printer performance with the customer on week 33 or before and avoid a major stop in production by reviewing operating parameters or performing some preventive maintenance ahead of time.

The machine shown in Figure 6 prints less hours per week and with higher variability than the previous one. Its printing time ranges between 4 and 30 printing hours per week. The printer is very stable regarding its warning rate performance. It is always classified in the low-risk cluster 2 and has small changes in WPC. However, on week 48 it shows a major increase in warning messages rate that classifies it in the high-risk cluster 4, also shown as a major

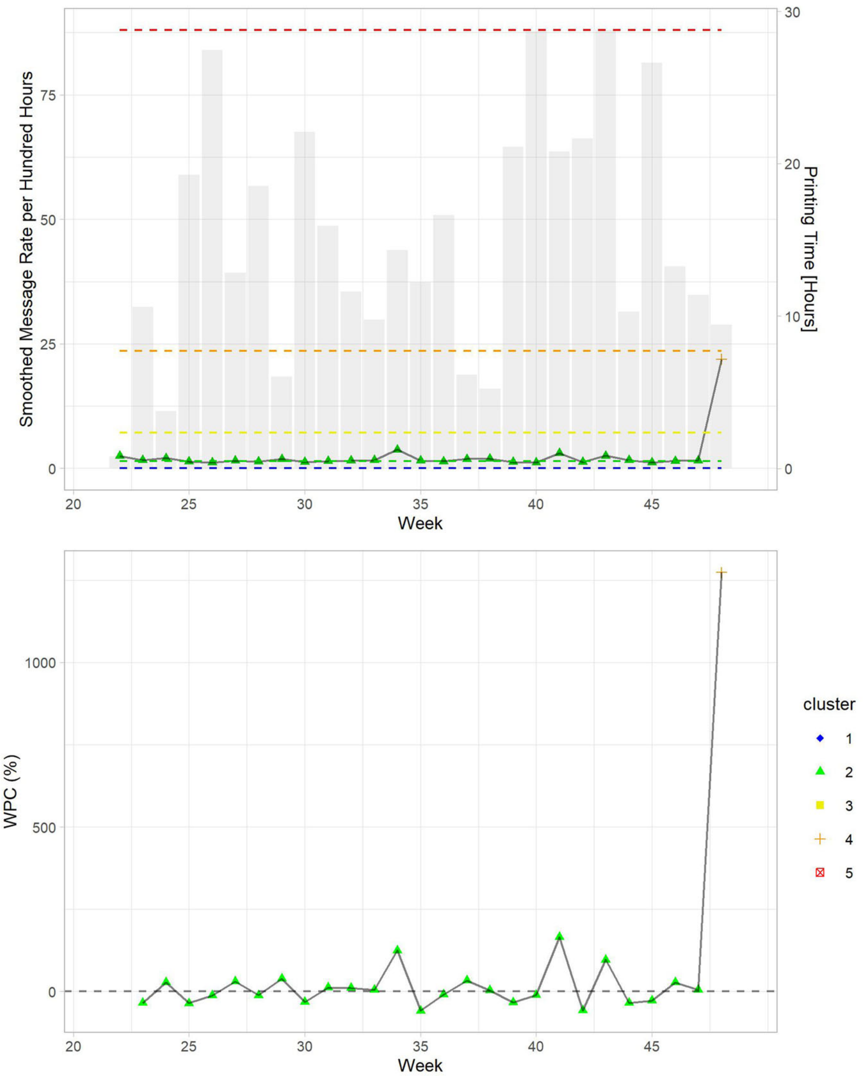


Figure 6. PERC (top) and WPC (bottom) graphs for phase II monitoring. The printer has a good record track up to week 48 where it is classified in cluster 4 from cluster 2 (PERC graph) and there is a large spike in percentage warnings rate growth (WPC graph). This is an example of the procedure catching an early signal of a printer's problem.

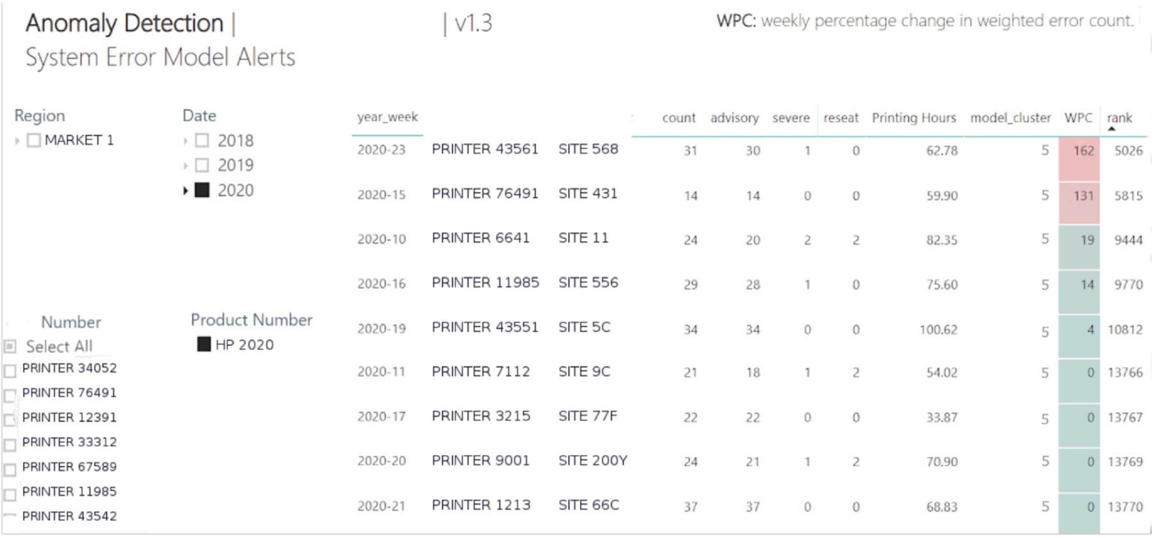
spike in the WPC metric. At this moment, Support Engineers should investigate the machine in more detail and contact the customer to assess any required preventive action.

3.3. Deployment

The resulting data was delivered through an interactive dashboard using Microsoft Power BI (Microsoft 2021). In Figure 7 an example screenshot of the dashboard is shown for a particular printer family.

The first view of the dashboard includes a ranked list of printers by both cluster number and WPC (see Section 2.6.2) allowing for quick identification of which printers need further resources assigned. Note here sensitive data such as printer serial numbers are screened from view. The selectors on the left part of the screen allow for drill down to specific region, time frame and printer. On the right is shown the ranked printers for that selection, where week number, message count, printing hours, cluster numbers, WPC and rank (sorted list of WPC and cluster number) are specified.

A second tab within the same dashboard allows for interactive exploration of individual printers to allow for a better assessment of the state of the printer and identification of the cause of problems (see Figure 8). Again the same selectors are given on the left part of the screen, and the data for a single selected printer is shown on the right. Here a table of each individual reported message is shown with date and message code which allows for the identification of the source of the message.



Anomaly Detection | | v1.3

System Error Model Alerts

WPC: weekly percentage change in weighted error count.

Region
□ MARKET 1

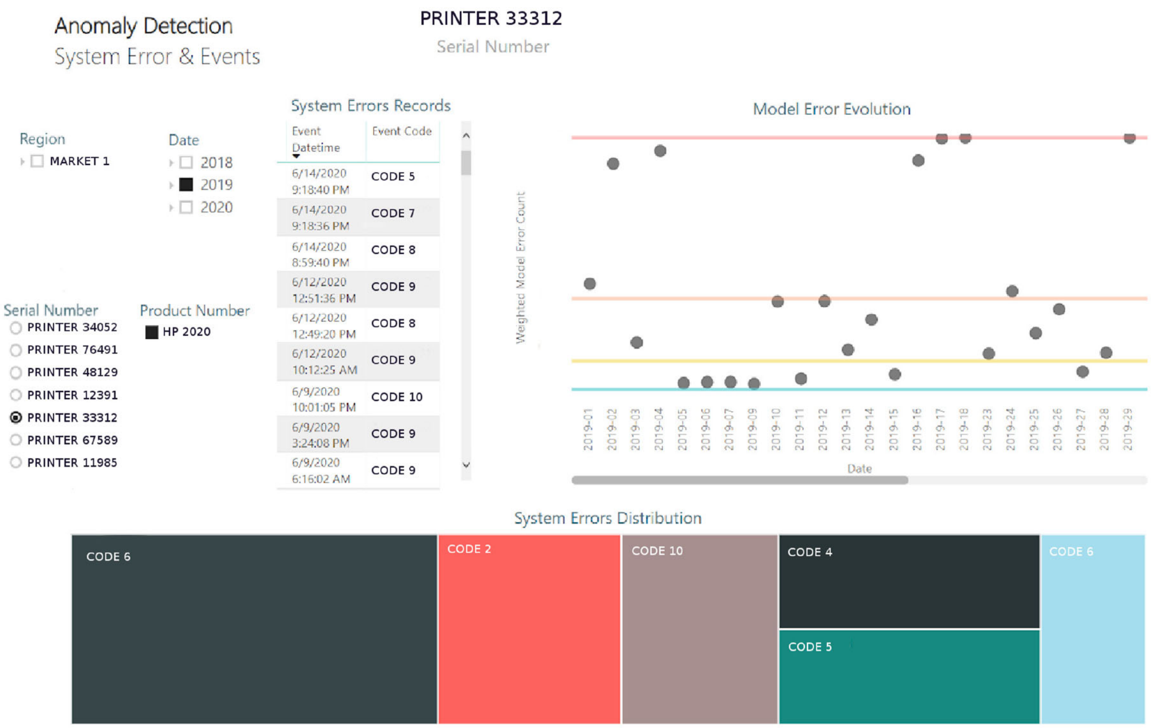
Date
□ 2018
□ 2019
■ 2020

Number
⊟ Select All
□ PRINTER 34052
□ PRINTER 76491
□ PRINTER 12391
□ PRINTER 33312
□ PRINTER 67589
□ PRINTER 11985
□ PRINTER 43542

Product Number
■ HP 2020

year_week			count	advisory	severe	reseat	Printing Hours	model_cluster	WPC	rank
2020-23	PRINTER 43561	SITE 568	31	30	1	0	62.78	5	162	5026
2020-15	PRINTER 76491	SITE 431	14	14	0	0	59.90	5	131	5815
2020-10	PRINTER 6641	SITE 11	24	20	2	2	82.35	5	19	9444
2020-16	PRINTER 11985	SITE 556	29	28	1	0	75.60	5	14	9770
2020-19	PRINTER 43551	SITE 5C	34	34	0	0	100.62	5	4	10812
2020-11	PRINTER 7112	SITE 9C	21	18	1	2	54.02	5	0	13766
2020-17	PRINTER 3215	SITE 77F	22	22	0	0	33.87	5	0	13767
2020-20	PRINTER 9001	SITE 200Y	24	21	1	2	70.90	5	0	13769
2020-21	PRINTER 1213	SITE 66C	37	37	0	0	68.83	5	0	13770

Figure 7. Sample image of Power BI developed by HP for the delivery and analysis of alerts to service teams. Here we show the landing page where a ranked list of printers are shown. The selectors on the left part of the screen allow for drill down to specific region, time frame and printer. See text for further details. Note the image has been edited to screen sensitive data such as printer serial numbers and locations.



Anomaly Detection

System Error & Events

PRINTER 33312

Serial Number

Region
□ MARKET 1

Date
□ 2018
■ 2019
□ 2020

Serial Number
○ PRINTER 34052
○ PRINTER 76491
○ PRINTER 48129
○ PRINTER 12391
● PRINTER 33312
○ PRINTER 67589
○ PRINTER 11985

Product Number
■ HP 2020

System Errors Records

Event Datetime	Event Code
6/14/2020 9:18:40 PM	CODE 5
6/14/2020 9:18:36 PM	CODE 7
6/14/2020 8:59:40 PM	CODE 8
6/12/2020 12:51:36 PM	CODE 9
6/12/2020 12:49:20 PM	CODE 8
6/12/2020 10:12:25 AM	CODE 9
6/9/2020 10:01:05 PM	CODE 10
6/9/2020 3:24:08 PM	CODE 9
6/9/2020 6:16:02 AM	CODE 9

Figure 8. Second page within Power BI where the the weekly evolution of the shrunken message rates (top right) are shown. This corresponds to the PERC plots seen in Figures 4 and 5. In the center of the image the full list of each individual error message is shown as well as time and date of occurrence. Along the foot of the image a treemap chart allows for a quick review of the most commonly occurring error message (here the area of each colored box reflects the number of times an error message is seen, the most commonly occurring placed on the left). The selectors on the left part of the screen allow for drill down to specific region, time frame and printer. Note the image has been edited to screen sensitive data such as printer serial numbers and locations.

The top right graphic shows the evolution of the cluster centers for that printer where the colored lines indicate the cluster numbers 1 to 5 and the black dots show the calculated error rate for that week. Along the bottom of view is a mosaic plot showing the magnitude of the contribution from each type of system error for the selected week. This allows for a quick analysis of what part of the machine is generating the errors and allows for fast troubleshooting.

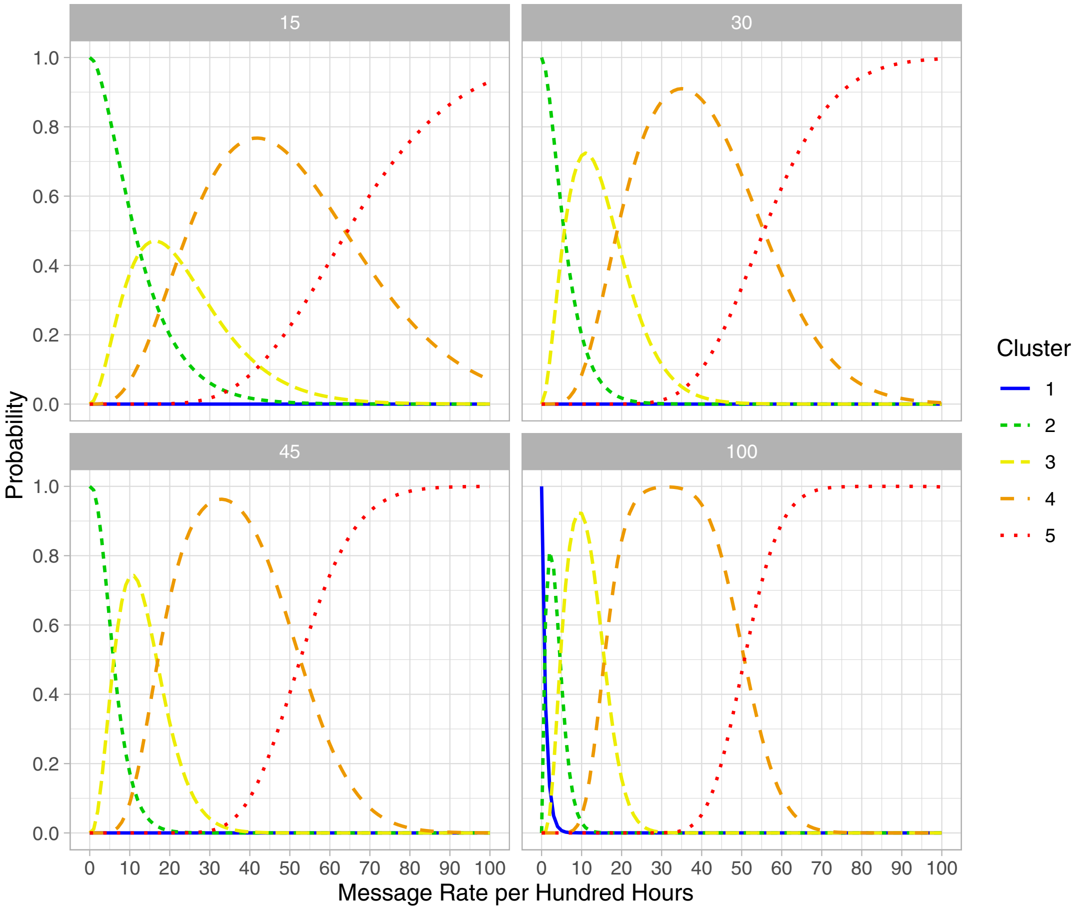


Figure 9. Probability of a printer being assigned to a cluster in a week as a function of the true printer error message rate (x-axis) and printing time (facet). Printing times shown include 15, 30, 45 and 100 hours a week.

3.4. Performance

In order to assess the performance of the proposed charts, we have plotted the probability of a printer-week being assigned to a cluster as a function of the true printer error message rate for different weekly printing times: 15, 30, 45 and 100 hours (Figure 9). These Operating Characteristic curves can be computed analytically without the need for simulation. Given an error rate and a printing time, the probability of seeing any number of error messages follows a Poisson distribution. Once we know the observed number of error messages, the probability of belonging to each cluster is obtained from expression (6) and plotted.

Figure 9 shows, as could be expected, that the higher the error message rate the higher the probability of an observation being assigned to a more extreme cluster. It also shows that the overlap among curves decreases as the weekly printing time increases.

Cluster 1 has a zero-error message rate. Cluster 2 is the second lowest rate cluster and the largest cluster in size, including 62.3% of the printer-weeks (Table 2). When the true printer error rate is zero the model classifies a zero observation in cluster 1 whenever the printing time is larger than 78 hours. Any zero observed in less than 78 printing hours is classified in cluster 2.

Clusters 4 and 5 are the ones with the highest error rate (Table 2). For instance, Figure 9 shows that observations from printers with true error rates larger than 25 and a 45 hours weekly printing time have a probability of being assigned to cluster 4 greater than 0.8, a value that increases as printing time grows. One can also see in the same graph how observations from printers with a true error rate larger than 55 and 45 hours printing time are classified more often to cluster 5 and almost never to clusters 1 to 3. Finally, cluster 3 is somewhere in between clusters 2 and 4.

4. Discussion

The proposed approach helps to define the number of clusters while reducing the variability in estimating error rates coming from low weekly printing times in the repairable system monitoring context. A small number of errors when the printing time is low delivers a large error rate. In these cases, the "credibility" of the error rate should be smaller than a similar error rate estimated over a longer period of time. The empirical Bayesian method minimizes this problem obtaining a shrunken estimate of printing error rates. The mixture model approach has the advantage of letting the model estimate the number of clusters (Schlattmann and Böhning 1993; Böhning 1999; Böhning, Dietz, and Schlattmann 2000; Böhning 2003). Another advantage is an easy implementation of the method through the package CAMAN in the free software R (Schlattmann et al. 2016). In addition, as we showed in our deployment example, it can be easily implemented in big data problems for real data time monitoring (Hong, Zhang, and Meeker 2018).

This methodology has been used in disease mapping research (Rattanasiri et al. 2004; Benach et al. 2004) and other fields as metanalysis, interval-censored data or capture-recapture studies for estimating population size (Schlattmann 2009; Böhning et al. 2007). On the other hand, we should note that given that the marginal distribution of the number of observed errors is a mixture of Poisson distributions, it is a suitable model to deal with zero-inflated behaviors. This is a characteristic present in our data, where many error rates are zero for many printer-weeks.

The model validation, not included in the article, showed that a single Poisson model (one-cluster model) could not recreate the heterogeneity of the observed data, while a 5-cluster one could. The Negative Binomial distribution could be another model of choice when the finite mixture of Poisson does not hold. In our case, grouping printer-weeks in clusters also fitted nicely with the Hewlett-Packard support engineers' perception that printers could be grouped in different clusters based on how "well" they behaved on the field, making the model easy to adopt by users.

Statistical Process Control (SPC) is an important business quality management tool. In this manuscript we have developed a monitoring process with special attention to phase I implementation. The proposed model allows a robust estimation of the parameters estimated in phase I and the classification of different patterns using a model-based cluster. Although there are many theoretical and practical articles related to SPC, it is still an evolving area nowadays. Risk-adjusted charting procedures have been developed for healthcare process control in recent years. They are appropriate when monitoring non-homogeneous observations such as patients with different risk levels to survive a surgical procedure due to preexisting health conditions (Zhang, Gan, and Loke 2012; Steiner 2014). In our case, the risk of printer failure may depend on variables related to the conditions under which the printers are used, and therefore the risk may not be homogeneous. This is a future line of work, where variables that affect the printer failure risk must be first identified and then taken into account. In recent years, modern industrial problems have become increasingly complex, and classical process control techniques may not be sufficient to solve them. The Internet of Things together with new methods such as neural networks or machine learning is likely to open a new paradigm in this field (Iqbal et al. 2019; Park, Fan, and Hsu 2020; Psarakis 2011; Sisinni et al. 2018; Zan et al. 2019).

The approach in this work does not take into account temporal correlation that could improve the estimation of failure rates with high variability. The inclusion of potential specific covariates in the analysis, as mentioned in the previous paragraph, could overcome any existing temporal correlation. Our aim in this project was to monitor and identify printers with high error rates and the proposed approach was useful to achieve this goal.

Finally, the current project, being a pilot, was implemented in a specific product line (latex-based printers) to limit variability within a product family. If one wants to extend the proposed methodology to new product lines care should be given whether to perform the analysis in a broader group of product families or focus independently in each group. The higher the difference between machine technologies and failure modes the less the model will benefit from pooling information. Optimal deployment strategies is an area that requires further thought.

In summary, the mixture model is a useful alternative to real-time monitoring of printers' error message rates in an industrial reliability context.

About the Authors

Bernat Plandolit Department of Statistics and Operations Research. Universitat Politècnica de Catalunya. 08028 Barcelona. Spain. Datancia SL. 08006 Barcelona. Spain linkedin: https://www.linkedin.com/in/bernat-plandolit-lopez/

Ignasi Puig-de-Dou Department of Statistics and Operations Research. Universitat Politècnica de Catalunya. 08028 Barcelona. Spain. Datancia SL. 08006 Barcelona. Spain linkedin: www.linkedin.com/in/ignasipuig

Gráinne Costigan. HP Large Format Big Data Team. 08174 Sant Cugat del Vallès. Spain (During development of project).

Xavier Puig. Department of Statistics and Operations Research. Universitat Politècnica de Catalunya. 08028

Barcelona. Spain. Analysis of Complex Data for Business Decisions Research Group. Universitat Politècnica de Catalunya. 08028 Barcelona Spain.

Lourdes Rodero. Department of Statistics and Operations Research. Universitat Politècnica de Catalunya. 08028 Barcelona. Spain. Analysis of Complex Data for Business Decisions Research Group. Universitat Politècnica de Catalunya. 08028 Barcelona Spain.

José Miguel Martínez. Department of Statistics and Operations Research. Universitat Politècnica de Catalunya. 08028 Barcelona. Spain. Public Health Research Group. University of Alicante. 03690 Alicante. Spain.

ORCID

Bernat Plandolit http://orcid.org/0000-0001-5043-6783
Ignasi Puig-de-Dou http://orcid.org/0000-0003-1747-8564
Gráinne Costigan http://orcid.org/0000-0001-5879-2416
Xavier Puig http://orcid.org/0000-0001-6525-0498
Lourdes Rodero http://orcid.org/0000-0002-8794-7541
José Miguel Martínez http://orcid.org/0000-0002-9633-1204

Acknowledgements

The authors would like to thank the support from the Hewlett-Packard Large Format Big Data Team in developing the project.

Funding

This work was partly supported by the Spanish Ministry of Science and Innovation and the European Regional Development Fund (grant number PID2021-125380OB-I00 (MCINN/AEI/FEDER)).

References

BBC. 2009. EU brings in ‘right to repair’ rules for appliances. Last Modified: Sept. 30 2019. https://www.bbc.com/news/business-49884827.

Benach, J., Y. Yasui, J. M. Martínez, C. Borrell, M. Pasarín, and A. Daponte. 2004. The geography of the highest mortality areas in Spain: A striking cluster in the southwestern region of the country. *Occupational and Environmental Medicine* 61 (3):280–1. doi: 10.1136/oem.2002.001933.

Böhning, D. 1999. *Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping and others*. New York: Chapman and Hall/CRC.

Böhning, D. 2003. Empirical Bayes estimators and non-parametric mixture models for space and time–space disease mapping and surveillance. *Environmetrics* 14 (5):431–51. doi: 10.1002/env.598.

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay. 1994. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics* 46 (2):373–88. doi: 10.1007/BF01720593.

Böhning, D., E. Dietz, and P. Schlattmann. 2000. Space-time mixture modelling of public health data. *Statistics in Medicine* 19 (17–18):2333–44. doi: 10.1002/1097-0258(20000915/30)19:17/18<2333::AID-SIM573>3.0.CO;2-Q.

Böhning, D., W. Seidel, M. Alfó, B. Garel, V. Patilea, and G. Walther. 2007. Advances in mixture models. *Computational Statistics & Data Analysis* 51 (11):5205–10. doi: 10.1016/j.csda.2006.10.025.

Carlin, B. P, and T. A. Louis. 2000. *Bayes and empirical Bayes methods for data analysis*. New York: Chapman & Hall/CRC.

Chen, Y., J. B. Birch, and W. H. Woodall. 2016. Effect of phase I estimation on phase II control chart performance with profile data. *Quality and Reliability Engineering International* 32 (1):79–87. doi: 10.1002/qre.1727.

Hong, Y., M. Zhang, and W. Q. Meeker. 2018. Big data and reliability applications: The complexity dimension. *Journal of Quality Technology* 50 (2):135–49. doi: 10.1080/00224065.2018.1438007.

H. P. 2021. “HP device as a service. n.d. ” Last visited October 11th 2022. https://www.hp.com/us-en/services/daas.html.

Iqbal, R., T. Maniak, F. Doctor, and C. Karyotis. 2019. Fault detection and isolation in industrial processes using deep learning approaches. *IEEE Transactions on Industrial Informatics* 15 (5):3077–84. doi: 10.1109/TII.2019.2902274.

Jones-Farmer, L. A., W. H. Woodall, S. H. Steiner, and C. W. Champ. 2014. An overview of phase I analysis for process improvement and monitoring. *Journal of Quality Technology* 46 (3):265–80. doi: 10.1080/00224065.2014.11917969.

Lawson, A. B., A. B. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel, A. Clark, P. Schlattmann, and F. Divino. 2000. Disease mapping models: An empirical evaluation. Disease Mapping Collaborative Group. *Statistics in Medicine* 19 (17):2217–41.

Leroux, B. G, and M. L. Puterman. 1992. Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* 48 (2):545–58. doi: 10.2307/2532308.

Microsoft, B. n.d. “Microsoft surface as a service.” Last Modified Jan. 9, 2019. https://news.microsoft.com/uploads/2016/09/SurfaceasaServiceFS.pdf.

Montgomery, D. C. 2019. *Introduction to statistical quality control*. 8th ed. Hoboken: John Wiley & Sons, Inc.

Oakland, R. J, and J. S. Oakland. 2018. *Statistical process control*. New York: Routledge.

Park, Y.-J., S.-K S. Fan, and C.-Y. Hsu. 2020. A review on fault detection and process diagnostics in industrial processes. *Processes* 8 (9):1123. doi: 10.3390/pr8091123.

Psarakis, S. 2011. The use of neural networks in statistical process control charts. *Quality and Reliability Engineering International* 27 (5):641–50. doi: 10.1002/qre.1227.

Psarakis, S., A. K. Vyniou, and P. Castagliola. 2014. Some recent developments on the effects of parameter estimation on control charts. *Quality and Reliability Engineering International* 30 (8):1113–29. doi: 10.1002/qre.1556.

Rattanasiri, S., D. Böhning, P. Rojanavipart, and S. Athipanyakom. 2004. A mixture model application in disease mapping of malaria. *Southeast Asian Journal of Tropical Medicine and Public Health* 35:38–47.

Schlattmann, P. 2003. Estimating the number of components in a finite mixture model: The special case of homogeneity. *Computational Statistics & Data Analysis* 41 (3–4):441–51. doi: 10.1016/S0167-9473(02)00173-1.

Schlattmann, P. 2009. *Medical applications of finite mixture models*. Berlin: Springer.

Schlattmann, P, and D. Böhning. 1993. Mixture models and disease mapping. *Statistics in Medicine* 12 (19–20):1943–50. doi: 10.1002/sim.4780121918.

Schlattmann, P., J. Hoehne, M. Verba, and M. P. Doebler. 2016. "Package 'CAMAN'." Based on C.A.MAN. R package version 0.74. Accessed December 12, 2020. https://CRAN.R-project.org/package=CAMAN.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6 (2):461–4.

Sisinni, E., A. Saifullah, S. Han, U. Jennehag, and M. Gidlund. 2018. Industrial internet of things: Challenges, opportunities, and directions. *IEEE Transactions on Industrial Informatics* 14 (11):4724–34. doi: 10.1109/TII.2018.2852491.

Steiner, S. H. 2014. Risk-adjusted monitoring of outcomes in health care. *Statistics in Action: A Canadian Outlook* 14:225–41.

Stoumbos, Z. G., M. R. Reynolds, Jr, T. P. Ryan, and W. H. Woodall. 2000. The state of statistical process control as we proceed into the 21st century. *Journal of the American Statistical Association* 95 (451):992–8. doi: 10.1080/01621459.2000.10474292.

Zan, T., Z. Liu, Z. Su, M. Wang, X. Gao, and D. Chen. 2019. Statistical process control with intelligence based on the deep learning model. *Applied Sciences* 10 (1):308. doi: 10.3390/app10010308.

Zhang, L., F. F. Gan, and C. K. Loke. 2012. Phase I study of surgical performances with risk-adjusted Shewhart control charts. *Quality Technology & Quantitative Management* 9 (4):375–82. doi: 10.1080/16843703.2012.11673299.