A SCALABLE PLATFORM APPROACH FOR ANALYZING CHANGE IN MANUFACTURING DATA

Alan H. Brothers, Ph.D.[1]
Benjamin II Carlos[1]
Kenneth Raymundo[2]

[1] Manufacturing Engineering Business Intelligence and Data Analytics
[2] Operations Excellence
Western Digital Corporation, 109 Technology Ave., SEPZ, Laguna Technopark,Binan, Laguna, Philippines 4024
alan.harold.brothers@wdc.com
benjamin.carlos@wdc.com
kenneth.raymundo@wdc.com

ABSTRACT

More than 1 million magnetic recording heads, known as sliders, are produced by a back end process consisting of more than 150 individual steps, which are controlled using several hundred key process indicators (KPIs). Even when changes to the manufacturing process are planned and carefully implemented, engineers are challenged to validate the impacts to all these KPI; when changes are unplanned, such as in cases of power outages or natural calamities such as typhoons, volcanic activity, pandemics, etc., it becomes effectively impossible.

We have developed an analytics platform, known internally as Protect, to empower engineers to collect, prioritize, and visualize end-to-end KPI data coming from our suppliers, our factory, and our customer, in both planned and unplanned change situations, within a matter of hours and entirely in a browser environment. It does this by collecting data from our cloud data warehouse, applying machine learning and statistical algorithms to quantify change and prioritize the KPIs with the most significant change, and then making the KPIs accessible for interactive visualization online using business intelligence software.

1. 0 INTRODUCTION

The Protect platform arose because our customers were concerned about how changes were managed in the manufacturing process, both planned changes – such as new process recipes, new raw materials, etc. – and unplanned changes – such as power outages or natural calamities. During planned changes, engineers would typically monitor only a handful of KPIs which their domain knowledge suggested would be impacted. This approach was highly dependent on engineer experience and failed in the face of unanticipated side effects. In the face of unplanned changes, the same approach was used, but with an even worse outcome, due to the variety and unfamiliarity of the changes.

Overcoming these challenges by analyzing the full set of KPI was impractical due to the high amount of expertise required to assemble such datasets, and to prioritize / screen only the most interesting results from them; and by the need to do the analysis quickly, especially in cases of unplanned change. Instead, a platform was needed that had at least the following features:

- A centralized, cloud-based data warehouse where hundreds of KPIs could be quickly accessed.
- A flexible KPI creation system where new KPIs could be quickly onboarded.
- A flexible definition system for defining two populations whose KPI could be compared, e.g. control vs. experimental, pre- vs. post-event, or known-good vs. known-bad.
- An algorithm or algorithms for measuring change in KPIs between the two populations, allowing significant changes to be ranked/prioritized.
- A flexible, interactive interface for visualizing and exploring the ranked KPIs, and performing simple statistical tests.
- A zero-code, web-based, high-availability interface for configuring analysis and exploring results, accessible to all users regardless of experience level.

Such a platform would enable engineers to quickly answer important questions such as:

- Was there significant change in the manufacturing data?
- What processes/zones showed change, and which specific KPI showed the most change?
- When did the change happen?
- Was the change sudden or gradual?
- Has the change already resolved, or was it permanent?

The operational impact of such a platform would be a significant improvement in time to action for all types of change, leading to fewer surprises, more consistent quality to

our customer, and less yield losses from at-risk inventory whenever harmful changes did occur.

2. 0 REVIEW OF RELATED WORK

Not Applicable.

3.0 METHODOLOGY

3.1 Source Data

In response to this need, Protect was designed to batch analysis of KPI changes between two user-defined populations, referred to as the Reference and Comparison. Whenever analysis is needed, the user simply defines each group using a combination of process dates, batch numbers, slider serial numbers, product names, and other useful filters; and then specifies which of the 500+ pre-configured KPI to analyze.

The user's criteria are used to dynamically query 'lookup tables' in a cloud-based data warehouse, provisioned by a major global cloud infrastructure leader and managed internally. The results of these queries are a set of matching slider serial numbers for the Reference and Comparison groups.

The lookup tables also contain a set of join keys, used to join the matching serial numbers to KPI data stored in a set of 'KPI tables', and resulting in the final analysis-ready datasets. While significant effort was needed to set up 500+ KPI columns, onboarding new KPI is as simple as specifying the table and column name where the KPI is stored, and the corresponding join key to join it to the lookup table.

Changes from the Reference to Comparison dataset are analyzed in 3 steps: (1) anomaly detection for generalized change detection / initial screening; (2) follow-up change analysis for numerical KPI; and (3) follow-up change analysis for categorical KPI.

3.1 Generalized Change Detection / Initial Screening

Anomaly detection starts by fitting an anomaly detection algorithm to the entire Reference dataset, assuming an arbitrary proportion of anomalies such as 5%. Several algorithms may be used for this, although the best algorithms for creating a scalable platform will be: (1) scale-invariant, to minimize costly pre-processing of the KPI data; and (2) capable of generating continuous anomaly scores, rather than normal/anomalous class labels.

An example of a suitable anomaly detection algorithm is the isolation forest, illustrated schematically below. This algorithm relies on univariate splitting and is thus scale-invariant; and it yields anomaly scores based on how 'difficult' it is to isolate each slider from the others in the Reference population, with anomalous sliders more easily isolated and normal sliders less easily isolated.

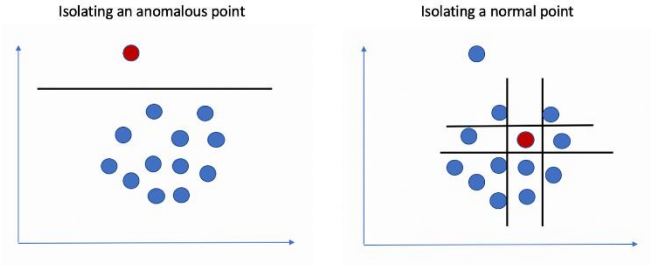

Figure 1: Schematic illustration of the isolation forest algorithm, which is one algorithm suitable for use in a generalized platform such as Protect.[1]

The fitted isolation forest or other model may be used to score all sliders in both populations, and the scores can then be compared for the purpose of giving a 'bird's-eye view' of change. Specifically, similar anomaly scores in the two populations indicate that the Comparison data fell within the envelope fitted around the most normal sliders of the Reference data, i.e. there was no major shift between populations. Conversely, higher or lower anomaly scores indicate that the Comparison data fell either less frequently, or more frequently, inside the envelope, i.e. there was a shift between the two populations.

This heuristic is useful for deciding whether further exploration of individual KPI is required or not in a particular analysis, and fitting can be performed on subsets of KPIs as needed, for example to indicate change within a particular process zone only, or change in KPIs relevant to a certain business outcome only. It is important to note that most anomaly detection algorithms were designed for use with numerical features, so generalizing the approach to allow for categorical data requires more care in the selection of algorithms[2].

3.2 Follow-Up Change Analysis

When detailed analysis is indicated by the initial change detection, it is necessary to rank individual KPIs, so that users may focus on only those with significant change. This prioritization must be done separately for numerical and categorical KPI, since comparisons across those data types are generally not meaningful.

Numerical KPIs may be analyzed using a variety of methods, for example the Earth Mover distance algorithm, which envisions each KPI distribution as a pile of 'dirt', and computes the work required to transform the Reference pile into the Comparison pile, as shown below. Earth Mover distance relies on minimal assumptions about the underlying

distributions of the KPIs and offers sensitivity to changes in both the magnitude and shape of the distribution, although standardization of the distances is required to allow the KPI to be compared, and thus ranked.

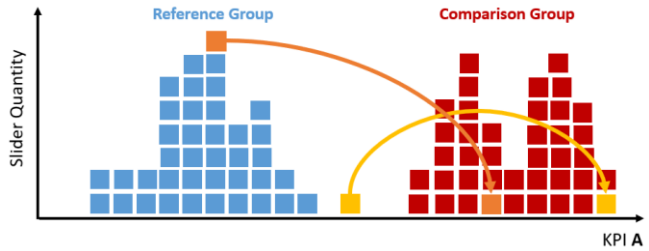

Figure 2: Schematic illustration of the Earth Mover distance for ranking numerical KPI based on change.

For categorical KPIs, change can be quantified using flexible, non-parametric hypothesis tests like the Chi-Squared Test of Independence, or related methods like Cramer's V. For example, a useful KPI change score may be derived from the test statistic computed when evaluating the null hypothesis that the categorical KPI is not associated with the dataset (Reference or Comparison), i.e. from the level of statistical confidence that the distribution of values in the Reference is or is not different from the distribution in the Comparison population. Once again, some post-processing is needed to handle edge cases, for example KPIs with very low or high cardinality.

3.3 Visualization and Exploration

Protect stores raw data, anomaly scores, and KPI change scores in a cloud-based object store and makes these available for interactive visualization and exploration using commercial business intelligence software such as Tableau, Power BI, or Spotfire. As described above, the goal of these visualizations is to quickly answer key questions such as whether meaningful change occurred or not, and if so, when it occurred and in what form; and to provide quick access to the full KPI data, particularly for those KPI ranked highest in terms of change.

4.0 RESULTS AND DISCUSSION

As already discussed, the value of the Protect platform comes mostly from the avoidance of quality and yield issues associated with undetected and unwanted KPI changes, which it makes possible by substantially reducing the turnaround time on data analysis, i.e. data collection, cleaning, visualization, and interpretation. While a comprehensive manual analysis without Protect could easily take several days, a Protect analysis can be configured and executed in a few hours or less, allowing changes to be detected, and initial hypotheses generated, almost as quickly as the data become available. Put another way, Protect allowed engineers to bypass the tradeoff between a comprehensive analysis and a responsive analysis, providing analysis that is both comprehensive of all KPI data, and highly responsive and timely.

When algorithms for anomaly detection and KPI change scoring are carefully chosen, the resulting platform easily accommodates a variety of KPIs, allowing it to easily scale both within a factory, and across factories / production processes, allowing greater quality and cost benefits with relatively small investment.

Examples of Protect use include validation of key data before and after factory power outages, typhoons, the eruption of the Taal volcano, and the transition to a work-from-home system during the Covid-19 pandemic. Protect is well adapted for such situations, since engineers have relatively little experience to guide them on which KPIs will / will not be affected in such situations, and thus are unable to effectively guarantee quality through conventional methods.

An example visualization is provided below. Key features in the visualization include the list of KPIs organized by change score (top right), and the full raw KPI data corresponding to whichever KPI has been selected (left).

Figure 3: Example visualization for KPI analysis, showing the KPI with the most change, and allowing users to view the corresponding raw data.

5.0 CONCLUSION

The vast quantity of data becoming available in modern manufacturing offers huge potential for improvements in both quality and yield. However, leveraging this potential without proportionately increasing manpower requires platforms capable of sifting through data and presenting engineers with only those data which are most relevant.

In this report, a platform was described to do this, by providing engineers with a fully web-based system for analyzing vast quantities of KPI data in both planned and unplanned change situations, and recommending to them only the data showing significant differences. Such a platform can be scaled across processes, across factories, and

even across industries, as a way to improve control over both quality and yield.

6.0 RECOMMENDATIONS

As outlined above, the key components of the Protect platform are the algorithms selected for the initial screening, and the individual KPI ranking. Although some well-known example algorithms were already described here, a variety of other methods would be suitable as well, and customization of these algorithms and metrics for particular processes and data environments is highly recommended in order to maximize value.

In particular, algorithms that operate on both numerical and categorical KPIs simultaneously are highly valuable, since these allow for even faster analysis by ranking all relevant data in a single list.

7.0 ACKNOWLEDGMENT

The author gratefully acknowledges the support of the Western Digital IT, analytics, and engineering teams, without whom such a complex platform would not be possible. In particular, the author would like to thank Yimin Zeng, Abhishek Kumar Gaur, Alan Huang, Kenneth Raymundo, Sorani Dejsook, Chee Kheng Lim, Maria Venus Gambito, and Luis Geoffrey Macapayag.

8.0 REFERENCES

1. A. Mavuduru. **Towards Data Science**, 2021, https://towardsdatascience.com/how-to-perform-anomaly-detection-with-the-isolation-forest-algorithm-e8c8372520bc.
2. Ayman Taha and Ali S. Hadi, **ACM Computing Surveys**, 52(2), 2019, 1-35.

9.0 ABOUT THE AUTHORS

Alan H. Brothers is a Data Scientist at Western Digital Corporation, focusing on analytics within the slider fabrication process. He has more than 15 years of experience as a researcher, developer, process engineer, and data scientist in the data storage, electronics manufacturing, and defense industries.

Benjamin G. Carlos II is Program Manager for Business Intelligence and Analytics for Manufacturing Engineering at Western Digital Corporation, with more than 15 years of experience in Test and HDD Recording Sub Systems Integration.

Kenneth F. Raymundo is a Data Engineer at Western Digital Corporation. He contributes data engineering proficiency honed over 13 years in the industry. His engineering background, coupled with experience across various departments, equips him to tackle complex data challenges.