Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 81 (2019) 24–27

www.elsevier.com/locate/procedia

52nd CIRP Conference on Manufacturing Systems

# RFID Data Driven Performance Evaluation in Production Systems

Ray Y. Zhong*

*Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong*

* Corresponding author. Tel.: +852-2859-2953; fax: +852-2859-6535. *E-mail address:* zhongzry@hku.hk

**Abstract**

Radio frequency identification technology commonly known as RFID technology is now widely used because it can increase productivity, efficiency and convenience. RFID has also been used in the manufacturing industry like shop floors where production data can be collected from production lines and sent to the information center for further analysis. Big data analysis provides a good opportunity, which can help the management of the shop floors. This paper reports on a case study using RFID datasets from a manufacturing shop floor to achieve performance evaluation. The datasets are processed by machine learning algorithms in R language. Some findings and observations have been obtained, which can be used as a reference for the future production.

*Keywords:* RFID, Big Data, Production Shop floor, Machine Learning.

## 1. Introduction

With the development of Radio frequency identification (RFID) technology, it has been widely used in manufacturing shop floors to collect data and identify objects [1-5]. RFID tags have been built-in chips so that they can be attached to certain objects whose data could be collected [6]. These tags convert data by responding to signals from RFID readers. In the manufacturing processes, a large amount of data is transmitted and stored [7]. The data is related to the batch of the product, the ID of the worker, the quantity of the product, the sequence of processing, and so on. In general, as time goes on, the data volume will become larger and larger. Therefore, using traditional analytical methods to analyze the data is not only difficult, but also needs to meet management requirements. Big data analytic tools such as statistics approaches, data mining models and machine learning algorithms, provide a good opportunity for processing the data [8-10]. Some available software like Python, R and MATLAB, all of which can be used for big data analysis. This paper aims to process and analyze a given dataset from a RFID-enabled manufacturing shop floor by using R. R is a programming language which is suitable for statistical computing by figuring out the hidden trends and valuable information and knowledge.

This paper uses a dataset from a RFID-enabled manufacturing shop floor including 9 columns of different data attributes and 413,473 rows of data which is obtained from daily production. The nine columns are: 'ID', 'BatchMainID', 'UserID', 'ProcCode', 'ProcSeqnum', 'Quantity', 'Good Number', 'Time' and 'Location'. The specific meanings of these nine columns are as follows:

- ID: auto-generated ID in SQL database;
- BatchMainID: representing a batch of product. Each BatchMainID corresponds to multiple 'UserID', 'ProcCode', 'ProcSeqnum', 'Quantity', 'Good Number', 'Time' and 'Location' data;
- UserID: indicating a specific worker. Each UserID corresponds to multiple 'BatchMainID', 'ProcCode', 'ProcSeqnum', 'Quantity', 'Good Number', 'Time' and 'Location' data;

10.1016/j.procir.2019.03.005

- ProcCode: representing a typical processing such as milling, drilling etc.

- ProcSeqnum: indicating the sequence of processing;

- Quantity: the total pieces for a batch. The value ranges from 0 to 180;

- Good Number: indicating the amount after inspection. The value of this column of data ranges from 0 to 180;

- Time: a time stamp reflecting the finished time of a processing. There are 36,726 out of 413,473 rows where the value is NULL.

- Location: representing a specific machine. Each Location corresponds to multiple 'BatchMainID', 'UserID ', 'ProcCode', 'ProcSeqnum', 'Quantity', 'Good Number' and 'Time' data;

Since this database is large and there are many complex relationships among the data, R language will be used to cleanse this data and a machine learning algorithm is used for analyzing the datasets. This paper is focusing on the research question: How can working time, sequences of processing and machines affect product quality?

The rest of this paper is organized as follows. Section 2 reports on the methodology. Section 3 talks about the results and discussions. Some findings and observations are discussed in this section. Conclusions are given in Section 4.

## 2. Methodology

### *2.1. Data cleansing*

There are 36,726 rows of data without time stamp which should be removed. There are 1,353 distinct UserID. User 48940 has the highest appearance rate in the data. Some UserID like 19831, 23973 and 24019 only appear once. It is possible that their records are in the deleted data trunk where time is missing.

There are 282 different types of processes numbering from 2 to 744. The most frequent type of processing is 456. Creating good item rate (GIR) = Good number/Quantity, GIR represents the quality ratio of products. One finding is that among the 376,746 rows of data which were cleansed, 11,284 rows of data have good item rate < 1.

Some machines have been used more frequently than others, for instance, machine 11002 been used 18,925 times. Whereas machine 20609 only has 1 use rate. One possible explanation is that this machine is mostly used in the deleted data trunk where time is missing.

Of the time data retained, there are 20,007 records that are found in the overnight work shift. The creation of time length is as follows:

*Same User-ID – same ProcCode – same ProcSeqnum – same Location*

As 20,007 pieces of data are overnight, that means date +1. Regardless the overnight data, another logic can be used:

*Same User-ID – same ProcCode – same ProcSeqnum – same Location – same Date*

Then using [Date (Last row) - Date (first row)] to get the time length to produce X amount of product (here X stands for unknowns).

### *2.2. Machine learning algorithms*

The goal of using machine learning algorithms is to predict the quality of the product (i.e. good item rate = Good number/Quantity) based on existing data.

Modelling to start with is least absolute shrinkage and selection operator (lasso – a kind of regression method), because it considers the interaction between variables (2-way interactions). Considering a sample of $N$ cases, each case consists of $m$ covariates and a single outcome. Let $y_i$ presents the outcome and $x_i := (x_1, x_2, ..., x_m)^T$ is the covariate vector for the $i^{th}$ case. Then, we can get:

$$\min_{\beta_0,\beta}\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\}$$

$$\text{s.t. } \sum_{j=1}^{m}|\beta_j|\le t$$

Where $t$ is a prespecified free parameter that determines the amount of regularization. $\beta\in\Box^{m}$ is the standard $l^m$ norm.

There are 1,353 workers and 547 different machines recorded in the dataset. So interaction between worker and machine will require 740,091 degree of freedom, which largely exceed the amount the data: 61,659.

Using Lasso, as it is a shrinkage method, it will perform variable selection, which eliminate (coefficient=0) any combination of interaction that is not useful in the model. That will free up the degree of freedom, inform the useful interactions, and cooperate with other machine learning approaches like neural network, support vector machine, etc [11, 12]. The lasso model matrix cannot deal with 2-way interaction because it requires cluster computing [13]. Thus, general lasso regression with no interaction is used in this research.

The dataset is randomly split into training set (70%) and test set (30%). Some UserIDs only have 1 or 2 rows. If setting them as factors, some UserID only occur in the training set while not the test set vice versa. Then the trained set cannot be tested. Therefore, the UserID is set as continuous and this method is applied to ProcCode, ProcSeqnum and Location to keep them as continuous. Using 10 folds cross validation to find the smallest λ for lasso regression, this λ is used to predict the good item rate in the test set.

The root mean square error of lasso regression is 0.05941, which is smaller than standard deviation of the original good item rate in the test set 0.05943. That means generally, the predicted good item rate is within the standard deviation.

## 3. Results and Discussions

### *3.1. ProcSeqnum and quality*

Fig.1 shows the impact of different ProcSeqnum (the sequence of processing) on product quality (good item rate). There are three main findings in this figure. Firstly, as it is shown in Fig.1, the data amount of the sequence of processing of No. 3, No. 49, No. 55 to No. 71 is much smaller than that from the other numbers.

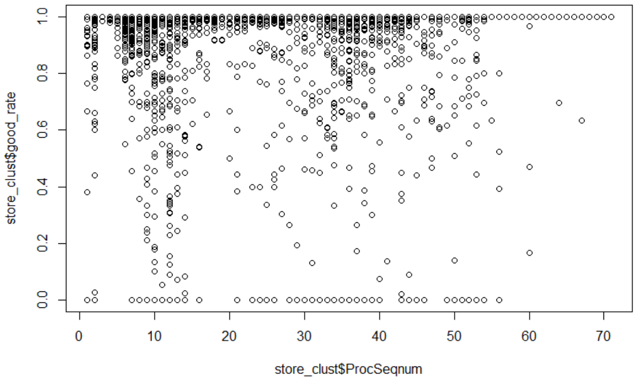

Fig. 1. The effect of different ProcSeqnum on quality

The data samples are less than five. There are two possible explanations for that. One is that these sequences of processing are not used much in this shop floor and the other is that their data has been removed in the data cleansing step due to large number of noises. Secondly, the quality of products produced in the sequence of processing of No. 18, 19, and 22 is relatively high and stable. While, the good item rate of No. 9 to No. 14 is very unstable as the floating range of this rate is quite large.

### 3.2. Working time and quality

In Fig. 2, the horizontal axis indicates the working time, from 0:00 to 24:00 and the vertical axis indicates the good item rate between 0.0 to 1.0. There are two important findings in this figure. Compared with the good item rate of products produced from 8:00 to 23:00, this rate is slightly decreased. Another observation is that the number of products produced between 5:00 and 7:00 is significantly reduced compared to other working times. This may be attributed to the inefficient worker's productivity during this time period as it is the early morning when people get fatigue usually.

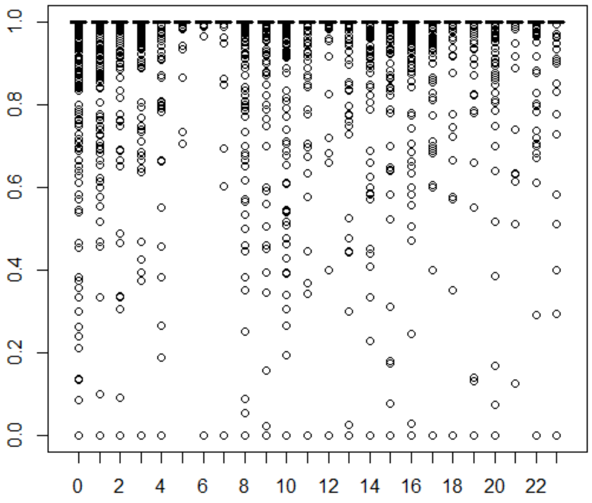

Fig. 2. The effect of different working time on quality

### 3.3. Location and quality

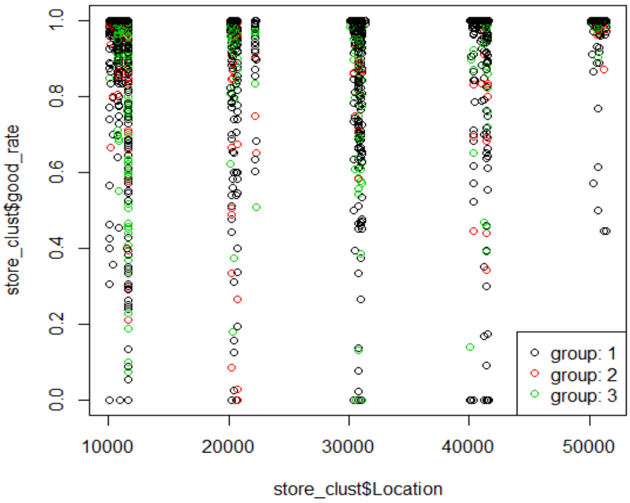

Fig. 3. The effect of location time on quality

Fig. 3 shows the impact of different machines on product quality. Three results can be derived from this figure. Firstly, machines numbered starting with 10,000 are used most in this shop floor. Secondly, in terms of the good item rate, machines with numbers starting at 10000 and 30000 are not as stable as the other three groups. The good item rates are between 0.4 and 1. Lastly, the good item rate of products manufactured by machines with a number starting at 50000 can be stabilized at around 0.9. That means the quality of the products from these machines is more stable.

## 4. Conclusion

This paper describes a study on RFID datasets from a manufacturing shop floor. The impacts of three perspectives including working time, the sequence of processing and machines on product quality are examined. In the data cleansing section, the duplicated time of different 'ID', 'BatchMainID', 'UserID', 'ProcCode', 'ProcSeqnum', and 'Location' has been removed. Three main findings can be summarized. The first is that the products produced from No. 18, 19 and 22 sequences have good item rate. Second, workers working in the early hours of the morning (0:00 to 7:00) may affect the quality of the products. Third, the quality of products produced by machines with numbers starting at 50000 is more stable.

Based on the above observations, there are several suggestions for this shop floor. Firstly, regarding the sequence of processing, if the same type of parts can be produced with No. 9, 14, 18, 19, and 22, it is better to use No. 18, 19, and 22 for production instead of No. 9 and 14, or technical improvements to No. 9 and No. 14 is needed. Secondly, for product batches with strict quality requirements, production arrangement between 0:00 to 7:00 should be avoided. Thirdly, this shop floor needs to carry out equipment maintenance or upgrades for machines with numbers starting with 10000 and 30000. In the meantime, it is advantageous to use machines with numbers starting with 50000.

Future research could be carried out in the following aspects. Firstly, the shop floor performance could be extended in terms of the production cycle and total output. Data-driven approaches will be further examined to investigate these aspects. Secondly, RFID data is one of the captured datasets in the IoT-enabled manufacturing sites. Other data from different sensors such as temperature, vibration and force could be integrated to the further decision-making.

## References

[1] Zhong, R. Y., Lan, S., Xu, C., Dai, Q., & Huang, G. Q. (January 01, 2016). Visualization of RFID-enabled shopfloor logistics Big Data in Cloud Manufacturing. The International Journal, Advanced Manufacturing Technology, 84, 5-16.

[2] Zhong, R. Y., Huang, G. Q., Lan, S., Dai, Q. Y., Chen, X., & Zhang, T. (July 01, 2015). A big data approach for logistics trajectory discovery from RFID-enabled production data. International Journal of Production Economics, 165, 260-272.

[3] Zhong, R. Y., Huang, G. Q., Dai, Q., & Zhang, T. (January 01, 2013). Estimation of Lead Time in the RFID-Enabled Real-Time Shopfloor Production with a Data Mining Model. 321-331.

[4] Zhong, R. Y., Dai, Q. Y., Qu, T., Hu, G. J., & Huang, G. Q. (January 01, 2013). RFID-enabled real-time manufacturing execution system for mass-customization production. Robotics and Computer Integrated Manufacturing, 29, 2, 283-292.

[5] Zhong, R. Y., & 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). (March 01, 2018). Analysis of RFID datasets for smart manufacturing shop floors. 1-4.

[6] IEEE International Conference on RFID-Technologies and Applications, Institute of Electrical and Electronics Engineers., & Institute of Electrical and Electronics Engineers,. (2014). 2014 IEEE RFID Technology and Applications Conference (RFID- TA).

[7] Finkenzeller, K. (2014). Rfid handbook: Fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication. Hoboken, N.J: Wiley.

[8] Katal, A., Wazid, M., Goudar, R. H., & 2013 Sixth International Conference on Contemporary Computing (IC3). (August 01, 2013). Big data: Issues, challenges, tools and Good practices. 404-409.

[9] Lee, J., Kao, H.-A., & Yang, S. (January 01, 2014). Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment. Procedia Cirp, 16, 3-8.

[10] Gandomi, A., & Haider, M. (April 01, 2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35, 2, 137- 144.

[11] Stanford University., Tibshirani, R., & National Science Foundation (U.S.). (1994). Regression shrinkage and selection via the lasso.

[12] Jerome Friedman, Trevor Hastie, & Rob Tibshirani. (February 01, 2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33, 1.)

[13] Yu, Q., & Li, B. (January 01, 2014). Regularization and Estimation in Regression with Cluster Variables. Open Journal of Statistics, 4, 10, 814-825.