Industry 4.0 Case Study
Big Data Analysis on an Industrial Database

Dr. Hajnal Éva, Dániel Marton, Mátyás Rick
Óbuda University Alba Regia Technical Faculty, Székesfehérvár, Hungary
hajnal.eva@amk.uni-obuda.hu
marton_dani@hotmail.com
owermaty@gmail.com

***Abstract*— A pilot project on an existing production line was accomplished to help to join to industry 4.0 objectives. The essential of the project is a big data analysis, which tries to find new information in the collected complex dataset and find correlation between the product quality inspection data and the data of the production line. Our final mission is to discover the cause of size and figure deviancies of a cylindrical product. One unit of a whole complex production line was investigated with statistical and intelligent data analysis methods. All the products are labelled with a unique identifier, but these were not included in the manufacturing data. The connection of the production line and the quality inspection was established by using the timestamp values. During this research classic statistical methods and smart data analysis methods, SOM data analysis were used. In the examination we could extract some interesting information. The results in the following step are utilizable by an automatic evaluation system, after that the information could be tracked back to the production line.**

I.INTRODUCTION

Industry 4.0 is today's important endeavor, which tries to change the industrial production in the direction of more intelligent, cost-efficient, flexible, adaptive and robust enough. In this system the collection and processing of information is the key element [1]. There is a continuous information gathering at the work place, the generated dataflow is integrating with the information of the corporation and the surroundings of the corporation, and the analysis of the created dataset is executed and results a loopback to the production [1],[2], [3].

With this structure of such system almost all experts agrees [3],[4]. The implementation of this process starts with the creation of the IT background. This is followed generally by form of data analysis pilot projects and primary data strategy forming, which are followed by the automation of the system with continuous data gathering, evaluation, and formation of intervention pipeline [5], [6]. The first and third steps of the process can be easily generalized, but the second is harder[5]. In every situation the involvement of an expert is mandatory to ask the appropriate questions from the system, to define the information to be processed and the way of this process and evaluation.

In this paper we will present a case study about a practical application of this concept from the gathered data to the practical decision making support via data cleaning and analyzing and concluding. The main questions are: can we extract any interesting information from the collected quality control and manufacturing data, and can we combine the quality management data and manufacturing data to get closer to the targeted predictive maintenance?

II.DATA

A. Manufacturing data from PLC

The data collected from the PLC was stored in an MS-SQL database. Data records between 2018.01.01 and 2018.06.30 were processed in this research. It contains the name of a particular workstation and its parameter (e.g. feed, rpm), the timestamp and value of the given parameter. The time values are recorded when a significant changes happen. This MS-SQL compatible backup contains approx. 3 million data records, with 27 variable (each recorded with timestamp). But only one workstation's five parameters - namely a lathe- were selected for the data analysis.

B. Quality inspection data of the production line

On the production line cylindrical product is produced. Every product is inspected and checked during the manufacturing process. The measured dataset contains 1-7 data record per product (depending on the measurement protocol). Each product is identified with a virtual unique product ID.

C. Unique product identifiers and the corresponding timestamp

This dataset contains the time value of the beginning and endings of a manufacturing process of a corresponding product on every machine queried from the robotic arms. This was recorded in an SQL database, so it is possible that there are some delay compared to the rough PLC data. By the investigations, the beginning of a product is differ from the PLC data, but the recorded endings have generally 1-2 second delay in a 50 s long production period.

III. METHODS

In the project MS-SQL database, R 3.5.0 statistical software, R studio Version 1.1.453 environment and MS Power BI were used [7].

A cleaned and homogenous dataset has to be involved in the investigation. The quality inspection features and production line data were processed according to this concept. The second goal was to link the manufacturing data to the quality inspection data, so the correlation between them can be evaluated.

In the first phase one workstation, which is a lathe, was investigated, its data were separated and in the future we can make similar examinations for the other workstations.

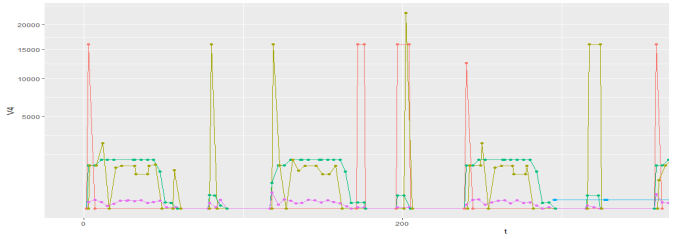

Figure 1 Raw data from PLC

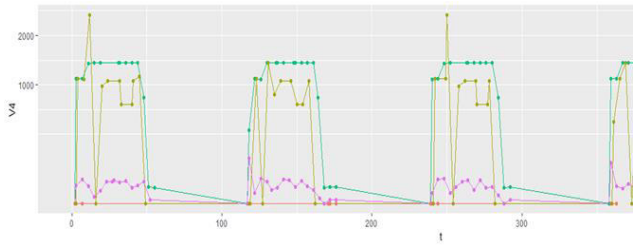

Figure 2. Data after the cleaning algorithm

Preparation of the production line data

The production line data was processed with the upcoming methods during the preparation

- Separation of the parameters connected to different workstations
- Data cleaning. On figure 1 we can see a section of data from the PLC. On the image we can see the rapid movements of the nonworking tool. While it is not relevant for the manufacturing process, it can be taken out from the dataset. The other disturbing factor is the spin up and down of the spindle between manufacturing of different products for the automatic cleaning of the workspace, so these should be taken out as well. The data cleaning was built upon the rotation of the spindle, because the manufacturing starts when the product starts spinning and it stops, when the spin stops. The identification of the manufacturing of one product is trivial on Figure 2.
- For each products (from the graphs on Figure 2) each parameter's maximum, average, integral value, biggest steepness of up going phase and down going phase were calculated. To sum up: For every workstation in case of 5 characteristics of the totally 25 variable was calculated. These data was take into the upcoming statistical analysis.

Preparation of quality control data

- Quality control data were separated by the workstation which is responsible for the given parameter.
- Values were standardized and normalized by the next method. Actual values of all parameters were subtracted from the optimal value and normalized with the tolerance of the given parameter, so the majority of data is in the [-1,+1] interval.

$$E_{norm} = \frac{K_{value} - Kta_{rget}}{K_{bound} - Kta_{rget}} \quad (1)$$

- If a radially symmetric attribute had more measured values on one product, the average and the most different from the targeted value was calculated. So totally 14 characteristics were gained.

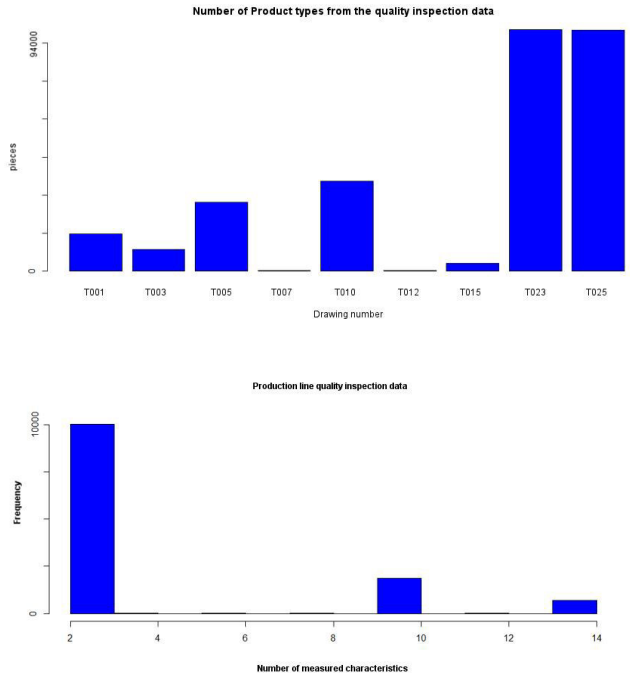

Figure 3 A. Number of product types from the quality management database. The most frequent product dataset was selected for further analysis. B. Distribution of the measured quality features of the selected product. All of the products have one measured feature, and according to the quality management policy some selected products took part in a more detailed investigation.

Fitting the product IDs to the manufacturing data

To this fitting data from section 'A' and 'C' were used, so the PLC data was possible extended with the product ID, which makes possible to track every product during the manufacturing, and makes the finding of correlations available. Fitting is possible, if we try to fit the calculated end of the manufacturing from PLC data to the manufacturing timetable. Because the database processes could have a 1-2 second latency in the records which have to take into account in the data processing.

It is very important to supply the statistical investigations with homogeneous data. To reach it, the most relevant product type and the most relevant characteristics were selected after a distribution calculation (fig. 3). Most of the products have only 1-2 measured features, which is later appropriate for time series investigations, the smaller part have 10-14 measured features which was involved in the correlation hunting statistics.

The prepared dataset were used for traditional statistics and for smart statistical methods. These were histogram calculation and visualization with R script and Self Organizing Map investigations. Finally ~5000 pieces of the products were investigated in the statistical analysis. The distribution of the characteristics is normal but some characteristics have no variance. These are excluded from the statistical investigations (fig 4).

SOM analysis were made with R Kohonen package [8][9][10]. Data were centered and scaled for the calculations. The map size was 20*25, type is hexagonal, the length of the training was 1000 epoch, the supersom method was used to avoid the problems of missing data [8][10].

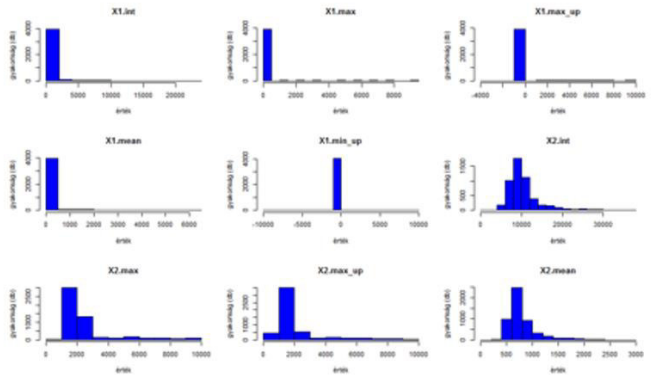

Figure 3. Histograms of PLC data of one workstations in the production line

IV. Results and Discussion

On figure 3 and 4 the histograms of the production line data and the quality inspection data can be seen. The production line data has minimal variance, however few extremely different values are in the histograms. The quality management data's histograms show larger variance. It can be presumable that the variance of the product does not show a direct correlation with one or few features of the production line, but the consequence of the accumulation of bit differences. However the extreme values that were measured on the workstation could be interesting. It is important to investigate if these are some artificial products of the data recording or cleaning algorithms or real extreme values. Furthermore the extreme process data and the connected product attributes have to be queried and investigated.

The distribution of the product quality data (fig 4.) was accepted to be normal by the Shapiro-Wilk normality test and the q-q test. However few histogram's maximum values have on offset from 0, which means that the largest part of the product is not optimal. It could be a normal consequence of the technological decisions if the variance is low enough. However some histograms are too wide - large variance -, which means that few products are out of the tolerance. In this case the narrowing of the histogram very likely need a technological change, but change of the offset needs only the setting of the workstation, which can cause the decrease of the number of rejected products. Some histograms have a local minimum in the middle of the graph instead of a global maximum. It suggest that the setting is often made ad hoc instead of deeper consideration, or might not accessible a correct model for the settings.

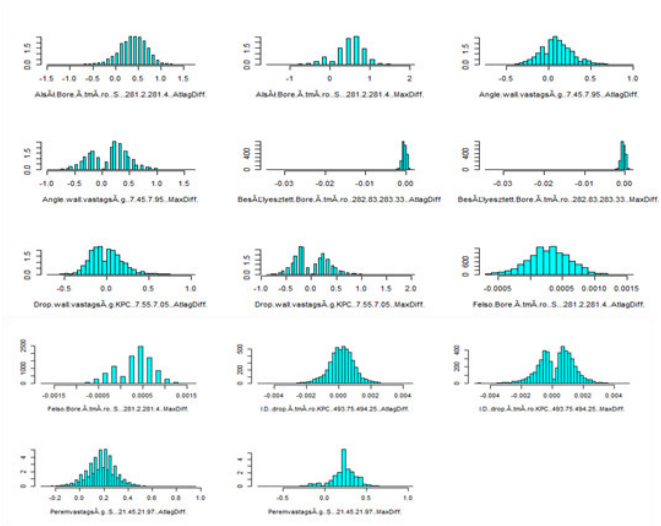

Figure 4. Histograms of quality inspection data measured during the manufacturing process

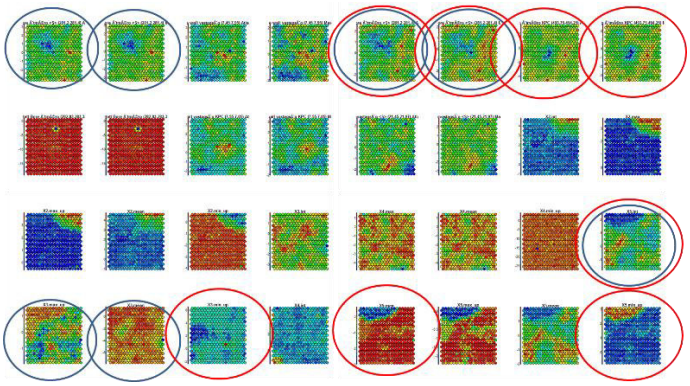

Figure 5. SOM feature maps of the quality inspection data and the relevant production line PLC data.

Self Organizing Map is a smart data analyzing method for medium or large size of dataset of which dimension number is smaller than 150. In this feed forward type neural network the neurons learn the dataset and create a map which is fitting to the dataset. The distance of the neurons shows the grouping property and clustering of the data, moreover each feature can be seen on a feature map, which makes possible the visual correlation hunting. Every single neuron has got weight vector, what is possible to reflect the processed data value, thus the SOM map characteristic planar for the purpose of the values of the different features, such as a topographic map can be evaluated visually. The akin map elements show the correlation between different characteristics. On fig 5 two correlation groups are signed with blue and red circles. However not the whole feature maps are similar, but only parts of them. Two interesting groups are in the map line. There is a group In the top left third section of the blue circled map is being depicted recurrently in more characteristic plane and the same groups can be found in characteristic of production line process, which would justify a conclusion that occurrence of certain deviation accompanied, at the same time it can be related to the characteristic of production line. It anticipates a correlation between the adjusted rotational speed of the spindle (main axle may reach the expected rotational speed slower) and the speed of stoppage next to the different values of the engine characteristics.

The other correlation can be observed in the red surrounding map by a line running from top right corner towards to two-thirds of the left down corner. The pattern is a characteristic occurred recurring in more maps which coinciding with more production characteristics. These are the cut- off speed of spindle speed rotation (rev), in addition to this there is a correlation between this pattern and different engine characteristics. It makes difficult to analysis of data that the same deviation of cutting off at the spindle, implies in one case positive in another case negative differences in size, combined with other production data. This occurrence due to the fact, that we try to analyze a complex system, and in many cases several reason can explain the same phenomena. It should analyze the statistical methods due diligence separating them properly.

V. Conclusion

The goal was to statistically analyze the 2018-year data of the products of a company. The task was to find a correlation between the quality characteristics of the products, most importantly the surface and size characteristics and the production line processes. On the production line the production process is quite complicated moreover, multiple products were made on the line which makes data analysis even further complicated. The statistic examinations must be executed on homogeneous data in order to avoid any statistical error. For this reason, the data of certain workstations and product types must be analyzed separately. From the histograms it is possible to determine that in case of some characteristics the maximum of the normal distribution is not at the optimal value. The correction of this is worth considering, because probably it would lead to the improvement of the quality without additional costs. Create histogram from quality control data could be a daily routine and could be easily achieved, because the data for this are online available at this moment.

Using the data of production line, quality inspection laboratory and the attached production line process, SOM analysis have been made. During the analysis a conclusion was formed, which includes that some of the quality characteristic and the production line data could be correlated by the lathe spindle load and lathe spindle rotation speed. This question was used to create a null hypothesis. To analyze the null hypothesis PCA, LDA, Boxplot diagrams can be performed on the data structures. For the first step only a few well-understood characteristic and production line characteristics as description values have been analyzed. Every statistics showed the possible correlation between the subset of the describer and the description values. The problem is that most of the variance of the description characteristics cannot be explained, these are under complex environmental control. Using other kind of interpretation, the describer values are not forming bordered clusters, which would make possible the very reliable, robust prediction that can be used in an industrial environment.

There are two options to get closer by executing future examinations. At the of time series analysis, the variance of the base noise and the autocorrelation can be distributed. If the analysis of the other product types will give a similar result, then it will be possible to combine them for the analysis of the

Time Series using proper precautions. The other possible method is the analysis of the combination of the described factors, which would probably return a good result as well. If the number of correlations reaches the required amount, then a simple decision tree can make the result applicable in industrial environment. All in all, a statistic null hypothesis has been created and in the next step the professional background must be clarified by involving company's employees.

Further research is required using the mentioned methods in order to prove the statement in the results. To ensure full investigation of the production line the other workstations are to be examined as well by using similar methods as before.

ACKNOWLEDGMENT

I would like to thank the help and the effort of colleagues and students. We acknowledge the financial support of Arconic 2019 grant.

REFERENCES

[1] J. Yan, Y. U. E. Meng, L. E. I. Lu, and L. I. N. Li, "Industrial Big Data in an Industry 4 . 0 Environment : Challenges , Schemes , and Applications for Predictive Maintenance," vol. 5, 2017.

[2] M. Habib, E. Ahmed, I. Yaqoob, I. Abaker, T. Hashem, M. Imran, and S. Ahmad, "Big Data Analytics in Industrial IoT using Concentric Computing Model," pp. 1–11.

[3] F. Tao, Q. Qi, A. Liu, and A. Kusiak, "Data-driven smart manufacturing," *J. Manuf. Syst.*, no. January, 2018.

[4] J. Lee, H. Kao, and S. Yang, "Service innovation and smart analytics for Industry 4 . 0 and big data environment," *Procedia CIRP*, vol. 16, pp. 3–8, 2014.

[5] M. Fernandes, A. Canito, V. Bolón-Canedo, L. Conceição, I. Praça, and G. Marreiros, "Data analysis and feature selection for predictive maintenance : A case-study in the metallurgic industry," *Int. J. Inf. Manage.*, vol. 46, no. 431, pp. 252–262, 2019.

[6] A. Kumar, R. Shankar, and L. S. Thakur, "A big data driven sustainable manufacturing framework for condition-based maintenance prediction," *J. Comput. Sci.*, vol. 27, pp. 428–439, 2018.

[7] F. Juretig, *R Statistics Cookbook*. BIRMINGHAM - MUMBAI: Packt Publishing Ltd., 2019.

[8] F. Lobo and V. Bação, *Introduction to Kohonen's Self-Organising Maps*. NOVA Information Management School - Universidade Nova de Lisboa, 2010.

[9] "How to label the som notes by the majority vote." [Online]. Available: https://stat.ethz.ch/pipermail/r-help/2010-June/241009.html. [Accessed: 26-Apr-2019].

[10] M. R. Wehrens, "Package ' kohonen ,'" 2018.