Ecosystem for COllaborative Manufacturing PrOceSses – Intra- and Interfactory Integration and AutomaTION
(Grant Agreement No 723145)

D3.4 Computational Modelling, Simulation and Prediction of Production I

Date: 2018-02-27

Version 1.0

Published by the COMPOSITION Consortium

Dissemination Level: Public

Co-funded by the European Union's Horizon 2020 Framework Programme for Research and Innovation under Grant Agreement No 723145

Document control page

Document file: Computational Modelling, Simulation and Prediction of Production I
Document version: 1.0
Document owner: CERTH

Work package: WP3 – Manufacturing Modelling and Simulation
Task: T3.3 – Simulation and Forecasting in Production and Logistics
Deliverable type: R

Document status: ☒ Approved by the document owner for internal review
☒ Approved for submission to the EC

Document history:

Version	Author(s)	Date	Summary of changes made
0.1	Thanasis Vafeiadis, Alexandros Nizamis (CERTH)	2018-01-20	ToC and Initial text
0.2	Tracy Brennan (BSL)	2018-02-01	Contribution on section 5.4
0.3	Thanasis Vafeiadis (CERTH)	2018-02-05	Contribution on Section 3,6 and 7
0.4	Nikolaos Alexopoulos, Christos Ntinas (CERTH)	2018-02-09	Contribution on Section 6
0.5	Dimosthenis Ioannidis, Alexandros Nizamis (CERTH)	2018-02-12	Contribution on Sections 4 and 5
0.6	Theofilos Mastos (KLE)	2018-02-14	Contribution to section 5.2, 5.3 and 5.4
0.7	Thanasis Vafeiadis (CERTH)	2018-02-16	Contribution to section 6
0.8	Ifigeneia Metaxa, Vasiliki Charisi (ATL)	2018-08-21	Contribution to section 4.3
0.9	Dimosthenis Ioannidis, Alexandros Nizamis (CERTH)	2018-02-21	Finalization for the internal review
1.0	Alexandros Nizamis (CERTH)	2018-02-27	Final version submitted to the European Commission

Internal review history:

Reviewed by	Date	Summary of comments
Luis Martins (BSL)	2018-02-27	Minor changes made to Section 1, 4 and 6
Aggelos Papadopoulos (KLE)	2018-02-26	The deliverable is well structured and written. Minor modifications required

Index:

1 Executive Summary

The present document is a deliverable of the "***Ecosystem for COllaborative Manufacturing PrOceSses – Intra- and Interfactory Integration and AutomaTION***" - (COMPOSITION) project, funded by the European Commission's Directorate - General for Research and Innovation (D-G RTD) under Horizon 2020 Research and Innovation programme (H2020). The deliverable presents an initial version of Computational Modelling, Simulation and Prediction of Production I developed until M18 of the project.

The COMPOSITION project needs simulation and prediction in both intra- and inter-factory scenarios. This report is focused on intra-factory scenarios which are related to production. In this first stage of the project and Task 3.3 - Simulation and Forecasting in Production and Logistics the research has been conducted to use cases which were selected as the cases with the highest priority from both pilot and technical partners of the project. Besides that, the first steps of the work that has been done to the rest of the use cases are also presented in this document.

The final version of the COMPOSITION's Simulation and Prediction engine and of the corresponding analysis results will be delivered at M28 with the second part of this deliverable, *D3.5 Computational Modelling, Simulation and Prediction of Production II*.

2 Abbreviations and Acronyms

Table 1: Abbreviations and acronyms are used in this deliverable

Acronym	Meaning
API	Application Programming Interface
CMMS	Computerised Maintenance Management System
DFM	Digital Factory Model
DSS	Decision Support System
ERP	Enterprise Resource Planning
LOF	Local Outlier Factor
MQTT	Message Queuing Telemetry Transport
NFA	Nondeterministic Finite – state Automata
SSP	Slope Statistic Profile

3 Introduction

3.1 Purpose, context and scope of this deliverable

This document presents the computational modelling, simulation and prediction functions on production developed until M18 of the COMPOSITION project. This document is part of the "*Task 3.3 – Simulation and Forecasting in Production and Logistics*" mean to design and implement trend analysis techniques for and on significant and key process variables. This deliverable defines the initial approaches for the core set of algorithms, techniques and methodologies dedicated on predictive maintenance. With the implementation of such techniques, we aim to provide detection of possible deviations from normal conditions.

3.2 Content and structure of this deliverable

The content of this deliverable is organized as follows:
In Section 4, an overview of the Simulation and Forecasting tool in overall project architecture is provided while in Section 5, a brief description of the data used for each use case that belongs to the area of computational modelling, simulation and prediction of production, according to *D2.1 – Industrial Use cases for an Integrated Information Management System* is provided. Section 6 provides a description of the functions and methodologies developed (new) and utilized or modified (existing ones from scientific literature) until M18 of the project, along with their application on the use cases. In Section 7, the next steps of the analysis are briefly described and in Section 8, the conclusions of this initial research are drawn.

4 Simulation and Forecasting tool in Overall COMPOSITION Architecture

This section describes the position of the Simulation and Forecasting Tool in the COMPOSITION project. The main interactions of this component with the rest of the project's components are described too.

4.1 Overview

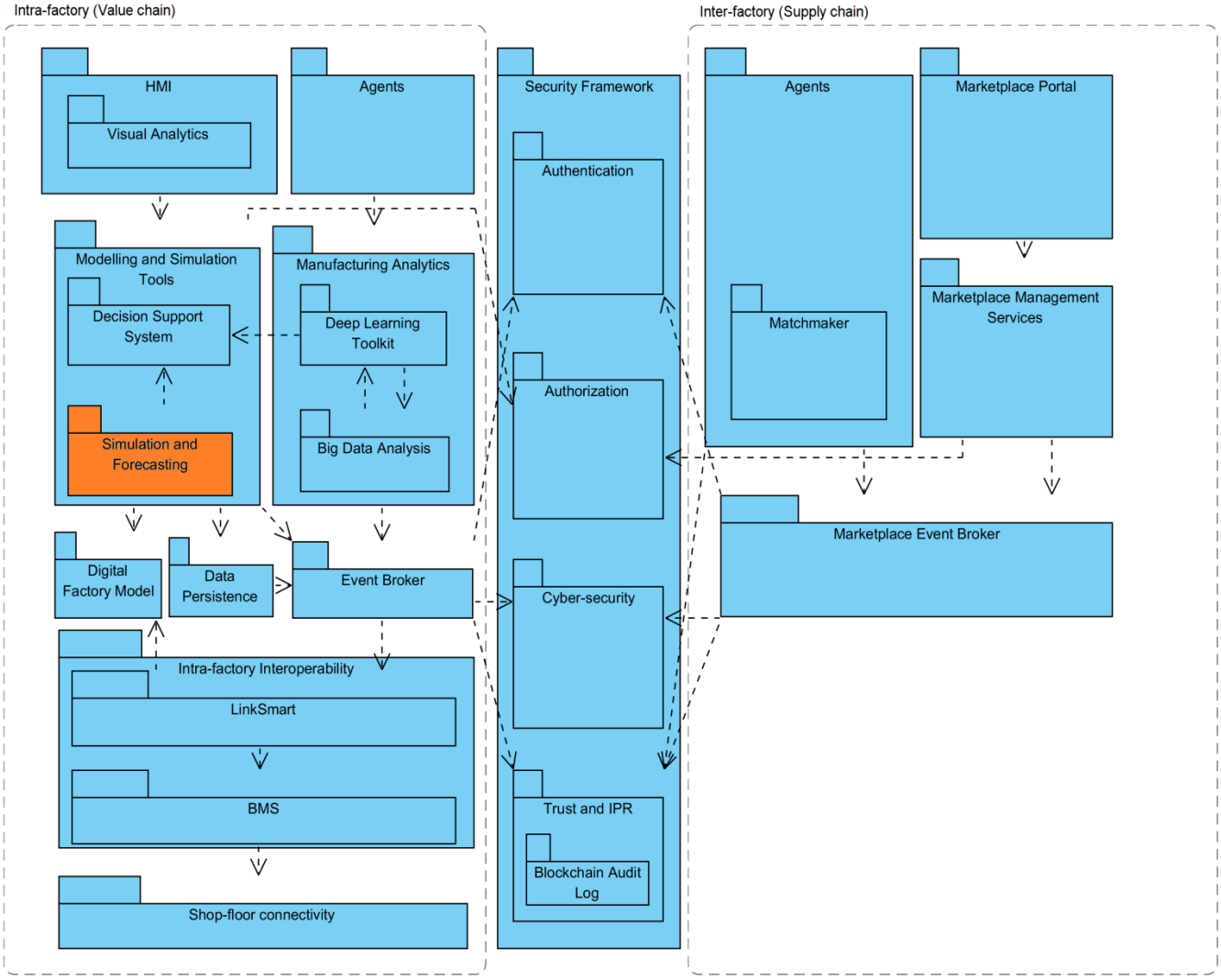


Figure 1: COMPOSITION architecture functional view

As depicted in Figure 1, the Simulation and Forecasting tool belongs to the Intra-factory package. More precisely, it is part of the Modelling and Simulation Tools module as described in *D2.3 - The COMPOSITION architecture specification I*. The Simulation and Forecasting tool is strongly correlated with DSS and Data Store and DFM components. Furthermore, it is also correlated with the Event Broker. The Broker will be responsible to transfer a Simulation tool's prediction to the Agent Marketplace for the project's Interfactory scenarios. As the correlation with the Event Broker is only matter of connectivity with the Marketplace, we decided to present in more details only the interaction with DFM and DSS components.

4.2 Digital Factory Model

The Digital Factory Model is a core component of the COMPOSITION system. The DFM enables the digitalization of industrial aspects. Data which are provided from different system's parts in a heterogeneous format finally are described in a common format using DFM schema. This means that all the data are

modelled and provided with the same format to all related components. The Digital Factory Model is able to describe all the information related to a factory such as buildings, assets, actors, processes and measurements.

The Simulation and Forecasting tool is able to load data from DFM using the DFM API's services. Besides that the Forecasting tool's predictions become available to other components using DFM API and the corresponding factory instances. Every prediction is described as an Event using the defined format from the DFM schema. More precisely, the DFM Event format is covered by OGC Observation and Measurements JSON format. Every prediction will be displayed as an Observation. In the figure below, an example of Forecasting tool output for UC-KLE 1 Maintenance decision support in terms of DFM schema is presented. The output presents the probability of future faults (electrical fault and mechanical fault) as a collection of OGC observations.


```
{
    "member": [
        {
            "href": "http://www.composition-project.eu/KLE_1_Fault_Probability"
        },
        {
            "resultTime": "09:02:36.656514",
            "result": {
                "uom": "http://www.composition-project.eu/uom#percentage",
                "value": 0.06599286563614744
            },
            "observedProperty": {
                "href": "http://www.composition-project.eu/bossi_fault_probability"
            },
            "id": "electrical_fault_id",
            "procedure": {
                "href": "http://www.composition-project.eu/predictive_maintenanceKLE"
            },
            "type": "CategoryObservation"
        },
        {
            "resultTime": "09:02:36.656514",
            "result": {
                "uom": "http://www.composition-project.eu/uom#percentage",
                "value": 0.06956004756242569
            },
            "observedProperty": {
                "href": "http://www.composition-project.eu/bossi_fault_probability"
            },
            "id": "mechanical_fault_id",
            "procedure": {
                "href": "http://www.composition-project.eu/predictive_maintenanceKLE"
            },
            "type": "CategoryObservation"
        }
    ],
    "phenomenonTime": {
        "instant": "09:02:36.656514"
    },
    "featureOfInterest": {
        "href": "http://www.composition-project.eu/faultProbability"
    },
    "_id": {
        "$oid": "5a7b0431c4e4cd27a4082c41"
    },
    "id": "KLE_1_SFT_Event"
}
```


Figure 2: SFT output as DFM event

The Simulation and Forecasting tool algorithm for UC – KLE 1 Maintenance decision support related to future faults of KLEEMANN's Bossi machine, has been dockerized and deployed to COMPOSITION Docker container. This Docker image sends continuously the probabilities of future fault to the DFM instance for the KLEEMANN's factory using the DFM API's service *setObservationCollection.* After that, the DSS component is able to get the posted probabilities from the aforementioned DFM instance by using the *getObservation* service from DFM API. So, the usage of DFM instance and DFM APIs service enables the communication between DSS and Simulation and Forecasting tool.

More details about DFM are available at *D3.2 Digital Factory Model I* which is published at M15.

4.3 Decision Support System

Decision Support System should be designed in order to accommodate the needs in a manufacturing environment. DSS integrates Digital Factory Models with events data, and other information and knowledge about the products, manufacturing, planning, simulation, communication and controls at all levels of planning and manufacturing. Raw data from sensors on factories are acquired and transformed according to the Digital Factory Model Schema. Processed data is accessed by the DSS through the DFM API. Accessing and processing the transformed data is easier and DSS implementation can be applied without the complicated need of transforming data in a suitable format. As a deliverable related to DSS is not yet available a brief description of the component is presented below in order to clarify the component's functionality.

Designing a DSS data and algorithm specification should be considered. Data specifications derive form DFM API. The algorithms suitable to be applied on a DSS in a manufacturing environment are both data mining algorithms in order to retrieve suitable data from a repository and decision-making algorithms for the decision-making process. The most used data mining algorithms are classifications trees, generic algorithms, support vector machine, Naïve Bayes. Various combinations and modifications of the above algorithms are considered to designing the data mining part of the DSS. Additionally, Nondeterministic Finite – state Automata (NFA) could be used in the decision-making process. The automata provide the possibility to be expanded during the process. Originally, DSS can implement a rule engine based on Finite State Machines, where the rules applied are defined based on the use cases. The initial rules can be used during a time period to train data and then NFAs and non – deterministic algorithms can be implemented for the decision – making process with the collaboration of the Deep Learning Toolkit.

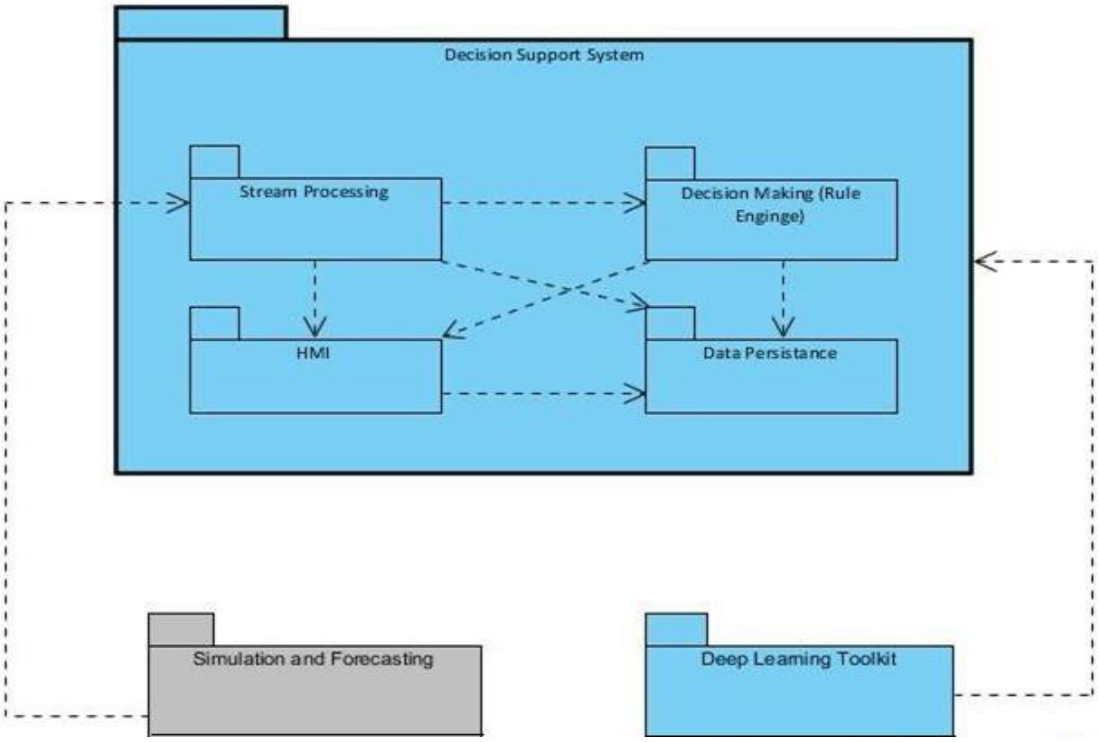


Figure 3: Decision Support System architecture

DSS architecture relates to the use of common data for all systems. Most manufacturing shop floors use the same sensors and they provide the same data. The differences are spotted to the levels of integration at the shop floors with the virtual factory models. Digital Factory Models create a virtual model of the factory, the DSS can work with. The process of making the virtual models of the factories follows at the section below.

Briefly, the sub-components of decision support system are:

- HMI - This component is responsible for the interaction with the user.
- Decision Making - The rule engine of decision support system
- Stream processing - This component processes the data of all external systems and extracts the information necessary for the decision support system
- Data persistence - This sub-component handles the storage of information for decision support internal use.

More details about the internal structure of these components are presented in deliverable D3.9 Manufacturing Decision Support System

Regardless the internal structure the decision support system deploys as one using a composed Docker image.

4.3.1 Planning and Preparation

Planning virtual factory models should consider the existing architecture of the system described in Figure 3. Data streams transformed from DSS into meaningful indicators and actionable data to provide knowledge. Based on B2MML models and definitions, we created the underlying module and data structures.

4.3.2 Analytic steps and modelling

In the COMPOSITION project there are several types to be modelled to aid the decision support process for maintenance prediction and process optimization. These are:

- Assets
- Asset Schematics
- Workers
- Workload
- Tasks
- Maintenance procedures
- Stream data from sensors
- Incident models

Based on the data above the models are categorized as follows:

- Context data
- Event data
- Performance indices
- Rules
- Actionable data
- Ingest the data into the data platform

The next step is to bring the relevant data from various sources, either from within or from outside the enterprise, into an analytic environment where it can be processed. The format of the data at source may differ from the format required at the destination. Data transformation may also be needed by the ingestion tool.

In addition to the initial ingestion of data, many intelligent applications are required to refresh the data regularly as part of an ongoing learning process. The learning process readapts data to provide KPIs for the decision – process. The readapted data can be inserted in the rule engine and define the states, transitions and parameters needed to implement a rule. This can be done by setting up a data pipeline or workflow. It is a part of an iterative process that includes rebuilding and re-evaluating the analytical models used by the intelligent application deploying the solution.

4.3.3 System Functionalities and API for the DSS

The next step is to obtain a deeper understanding of the data by investigating its summary statistics, relationships, and by using techniques such as visualization. This is also where issues of data quality and integrity, such as missing values, data type mismatches, and inconsistent data relationships, are handled. Pre-processing transformations are used to clean up the raw data before further analytics and modelling can take place.
Statistics provide the initial relationships between different kinds of types defined above. Also, statistics can be used to know the nominal operational procedures as well as the failure modes and exceptions. Data visualisation provides an easily readable format for the acquired data. Extraction of knowledge is visualised in the decision – making process and graphs are created with valuable information.
Identification of incosistancies in data relationships and corrupted data are able to be handled with statistical analysis and visualisation. Outliners are spotted and whether or not they are included in the data set is decided by decision makers, according to pre – defined process in the factories. Outliners provide significant knowledge in the maintenace processes of a factory, because the indicative faults, fault frequencies and causes and they should be processed in order to extract maintenance KPIS.

Table 2: DSS services

Service	***Description***	***Methods***
SyncData	*Synchronizes data*	- *SyncUser* - *SyncTaskWorker* - *SyncAvailableWorker* - *SyncAssets* - *SyncModels* - *AddKPI* - *RemoveKPI* - *GetKPIS* - *AddRule* - *RemoveRule* - *GetRules* - *AddNotification* - *RemoveNotification* - *GetNotifications* - *AddNotificationChannel* - *RemoveNotificationChannel* - *GetNotificationChannels*

GetData	*Acquire data*	- *getAssetsOffline* - *getDescriptionFromAssetId* - *getTaskKindOffline* - *getWorkersOffline* - *getAvailableWorkersOffline* - *getTaskDataOffline* - *getCorrectTask* - *getTaskOffline*
NotifyOnAction	*Updates rules related data*	- *NotifyUser* - *NotifyUserOnAction* - *HandleSuccess* - *HandleFailure*

5 Industrial Data Description

In this section a brief description of the data used for each use case that belongs to the area of computational modelling, simulation and prediction of production, according to D2.1 – Industrial Use cases for an Integrated Information Management System is provided.

5.1 Use Cases Prioritisation

Before starting with use cases' data description in this section and the data processing in the following section, it is worth to mention that the work has been done at Task 3.3 by M18 was mainly focused in scenarios which were selected as of highest priority from both technical and pilot partners.

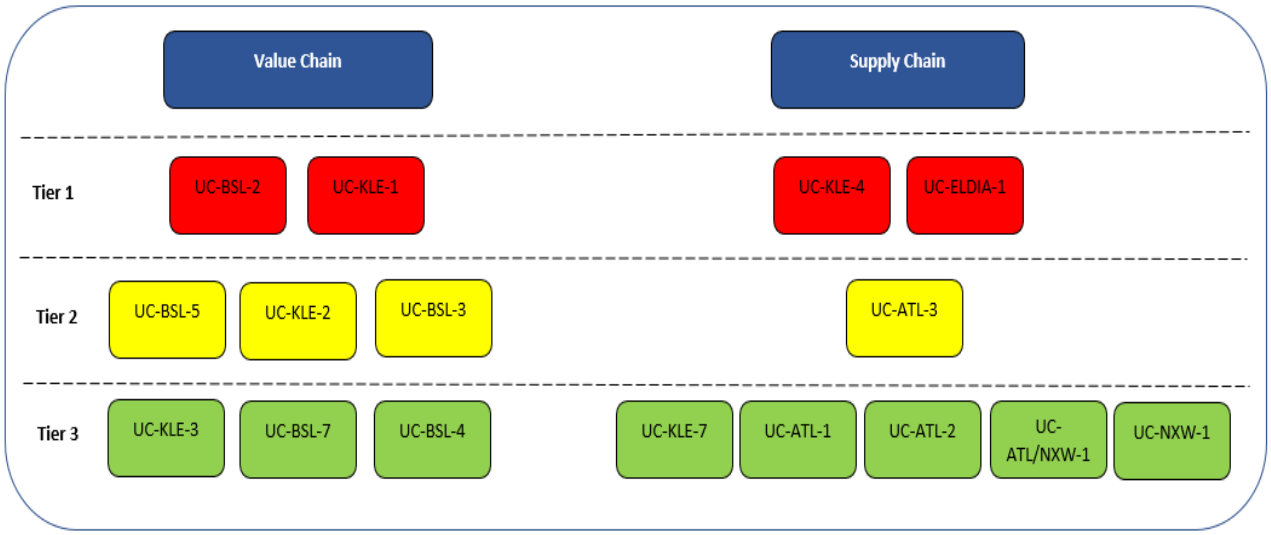


Figure 4: COMPOSITION Use cases prioritisation

As depicted in figure 4 the UC – KLE 1 Maintenance decision support and UC – BSL 2 Predictive Maintenance are the highest priority intra-factory scenarios cases. The main focus of this report is the documentation of research and development related to these two use cases. However, the first results of other intra-factory use case scenarios such as UC - KLE 2 and UC – KLE 3 are presented as well.

5.2 UC – KLE 1 Maintenance decision support

This use case focuses on the early detection of machine failure in the BOSSI polishing machine at KLEEMANN's shop-floor. A dataset generated by the Computerised Maintenance Management System (CMMS) is analysed. CMMS maintains a database including information about the company's maintenance operations, such as failure/problem description (mechanical, electrical, hydraulic), duration of breakdown repair, cost of machine breakdown repair, cost of person hours, cost of parts required for repairing etc. This set of data is extracted from CMMS in excel as a report file. 590 breakdowns have been recorded in a period of 10 years (2007-2017). CERTH has already provided a descriptive analysis on the types of faults, actions and changed parts. The probability of the type of the next breakdown to happen has also been calculated by utilizing time series.

A second set of data, will be generated from the sensors that will be installed inside and outside of the BOSSI machine by CERTH. The sensors that will be used in order to capture vibration data are accelerometers. More information about the sensor types can be found in *"D7.6 On site Readiness Assessment of Use Cases based on Existing Sensor Infrastructure I"*.

The data of both datasets generated from CMMS and the installed sensors, will be analysed together to give early indication that a motor inside or outside of the BOSSI machine will face a near future breakdown. This will then be communicated to the maintenance planner and maintenance manager via email or via the COMPOSITION platform.

5.3 UC – KLE 2 Delayed Process Step

This use case focuses on the prediction of delayed process step that leads to bottlenecks in the piston roughing and polishing steps. A set of productivity data is generated by the company's ERP system. The ERP maintains a database including information about the productivity in specific work stations, such as start-end time, human duration per piston, standard time etc. This set of data is extracted from ERP in excel as a report file. Around 13.800 records have been produced in a period of more than 1,5 years (01/2016-09/2017). The COMPOSITION system will perform forecasts for bottlenecks based on the production's historical data and will provide suggestions to the Production Manager to adjust the process and ultimately erase the bottlenecks. CERTH is currently working on the analysis of productivity data.

5.4 UC – KLE 3 Scrap Metal and Recyclable Waste Transportation

This use case focuses on the detection of bin and container fill levels and the calculation of the optimal route for collecting bins inside KLEEMANN's shop-floors. A set of quantity data is generated by the company's ERP system. The ERP maintains a database including information about the produced scrap metal, the paper, wood and plastic wastes. This set of data is extracted from ERP in excel as a report file. In 2016 around 1.000 tons of scrap metal were produced. Also, 7 tons of plastic, 4 tons of wood and 79.5 tons of paper have been recycled. In this use case, a time of flight distance sensor will be used. More information about the sensor types can be found in *"D7.6 On site Readiness Assessment of Use Cases based on Existing Sensor Infrastructure I"*.

5.5 UC – BSL 2 Predictive Maintenance

This use case focuses on the early detection of motor failure in the reflow ovens at BSL. In order to detect this there are two data sets which will be analysed. The first set of data is the machine log files which are generated from the machines themselves. For each blower the machine logs SP (Set point) which is the temperature set by the user, PV (Present value) which is the real temperature measured by the thermometer and OP (Output power) which is the power measured as a percentage where 100% means full voltage is applied and 0% means no voltage is applied. These log files are daily files which are updated every five minutes. The second set of data is generated from the sensors which have been placed in the reflow oven by Tyndall. These sensors can 'listen' and monitor performance (temperature, vibrations, power consumption) on and near fans (blowers) in reflow ovens. It is intended that the 'signature data' generated from these sensors and the log files from the oven will be analysed together to give early indication that a fan will fail in the near future. This will then be communicated to relevant personal via email and displayed on large visualization screens in the factory floor.

6 Data Preparation and Processing

This section provides a description of the functions and methodologies developed (new) and utilized or modified (existing ones from scientific literature) until M18 of the project, along with their application on the use cases.

6.1 UC – KLE 1 Maintenance decision support

6.1.1 Methodologies

6.1.1.1 Heatmaps

A **heat map** (or **heatmap**) is a graphical representation of data where the individual values contained in a matrix are represented as colours. The term 'heat map' was originally coined and trademarked by software designer Cormac Kinney in 1991, to describe a 2D display depicting financial market information (US Patent, 1993).Heat maps originated in 2D displays of the values in a data matrix. Larger values were represented by small dark gray or black squares (pixels) and smaller values by lighter squares.

For this use case, we have applied heatmaps to describe the metric of Pearson's correlations among the variables of the dataset.

6.1.1.2 Regression

The earliest form of regression was the method of least squares, which was published by Legendre in 1805 (Legendre, 1805) and by Gauss in 1809 (Gauss, 1809) Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821 (Gauss, 1821)] including a version of the Gauss–Markov theorem.
Linear and Logistic regressions are usually the first algorithms people learn in predictive modelling. Due to their popularity, a lot of analysts even end up thinking that they are the only form of regressions. The ones who are slightly more involved think that they are the most important amongst all forms of regression analysis. Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. As mentioned above, regression analysis estimates the relationship between two or more variables. There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the significant relationships between dependent variable and independent variable.
2. It indicates the strength of impact of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.
There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line). The types of regression most mentioned in scientific and statistics literature are: linear regression, logistic (or logit) regression, polynomial regression, stepwise regression, ridge regression, lasso regression and elastic net regression. Due to the nature of the data of this use case, we had applied only the linear regression.

Linear Regression
It is one of the most widely known modelling techniques. Linear regression is usually among the first few topics which people pick while learning predictive modelling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear. Linear Regression establishes a relationship between dependent variable, denoted here after as Y, and one or more independent variables, denoted hereafter as X, using a best fit straight line, also known as regression line. It is represented by an equation:

$$Y = a + b * X + e$$

where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s). The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.

Logistic Regression
Logistic regression is used to find the probability of event = Success and event = Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Logistic regression is widely used for classification problem and doesn't require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio.

Polynomial Regression
A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

$$Y = a + b * X\text{^}2$$

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

Stepwise Regression
This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves *no* human intervention. This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping co-variates one at a time based on a specified criterion. Some of the most commonly used Stepwise regression methods are listed below:

- Standard stepwise regression does two things. It adds and removes predictors as needed for each step.
- Forward selection starts with most significant predictor in the model and adds variable for each step.
- Backward elimination starts with all predictors in the model and removes the least significant variable for each step.

The aim of this modelling technique is to maximize the prediction power with a minimum number of predictor variables. It is one of the methods to handle higher dimensionality of data set.

Ridge Regression
Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimates (OLS) are unbiased their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Lasso regression
Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero.

ElasticNet Regression
ElasticNet is a hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features which are correlated.

Due to the nature of the data of this use case, the most suitable type of regression is the linear regression.

6.1.1.3 Probability Theory

Probability theory is the branch of mathematics concerned with probability. Although there are several different probability interpretations, probability theory treats the concept in a rigorous mathematical manner by expressing it through a set of axioms. Typically these axioms formalise probability in terms of a probability space, which assigns a measure taking values between 0 and 1, termed the probability measure, to a set of outcomes called the sample space. Any specified subset of these outcomes is called an event. Central subjects in probability theory include discrete and continuous random variables, probability distributions, and stochastic processes, which provide mathematical abstractions of non-deterministic or uncertain processes or measured quantities that may either be single occurrences or evolve over time in a random fashion. Although it is not possible to perfectly predict random events, much can be said about their behaviour. Two major results in probability theory describing such behaviour are the law of large numbers and the central limit theorem.

For this use case, we have developed an approach where the probabilities of an upcoming event are calculated based on scenarios prior to that event.

6.1.2 Application

In this section the application of correlation heatmap, linear regression and probability theory methodologies are provided and briefly described below.

6.1.2.1 Heatmap - Linear Regression

The variables of the dataset's use case are: machine fault type (PROBLEM DESCRIPTION), duration of problem solution (in hrs) (DURATION), the actual person hours (PERSON HOURS), the actual cost of person hours (COST OF PERSON HOURS), the actual cost of parts to be replaced (COST OF PARTS), the type of parts to be replaced (PARTS) and the action taken by the personnel to solve the problem (ACTION). A correlation heatmap of all variables mentioned above is given in Figure 5.

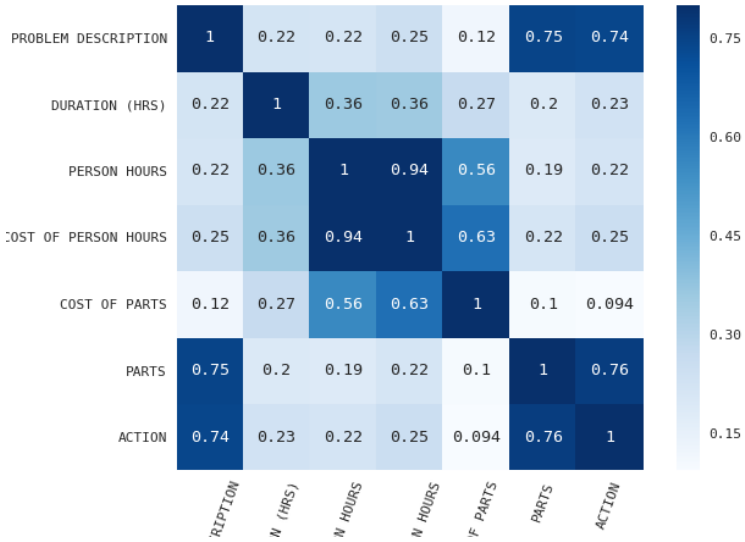


Figure 5: Correlation heatmap of all use case dataset's variables.

The heatmap on figure 5, point to the fact that there is strong positive correlation between the machine fault type, the parts and the action taken by personnel so as to solve the problem. The actual problems of the industrial machine are three: electronic, hydraulic and mechanical. Figure 6, provides the distribution of machine fault types in a time range of 10 years. One can see that the most common fault types are electrical and mechanical ones.

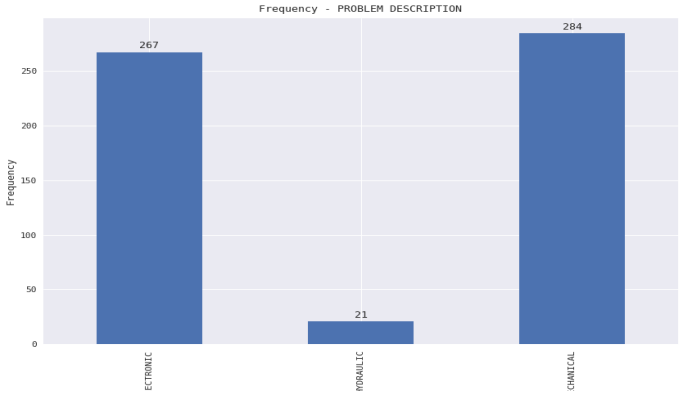
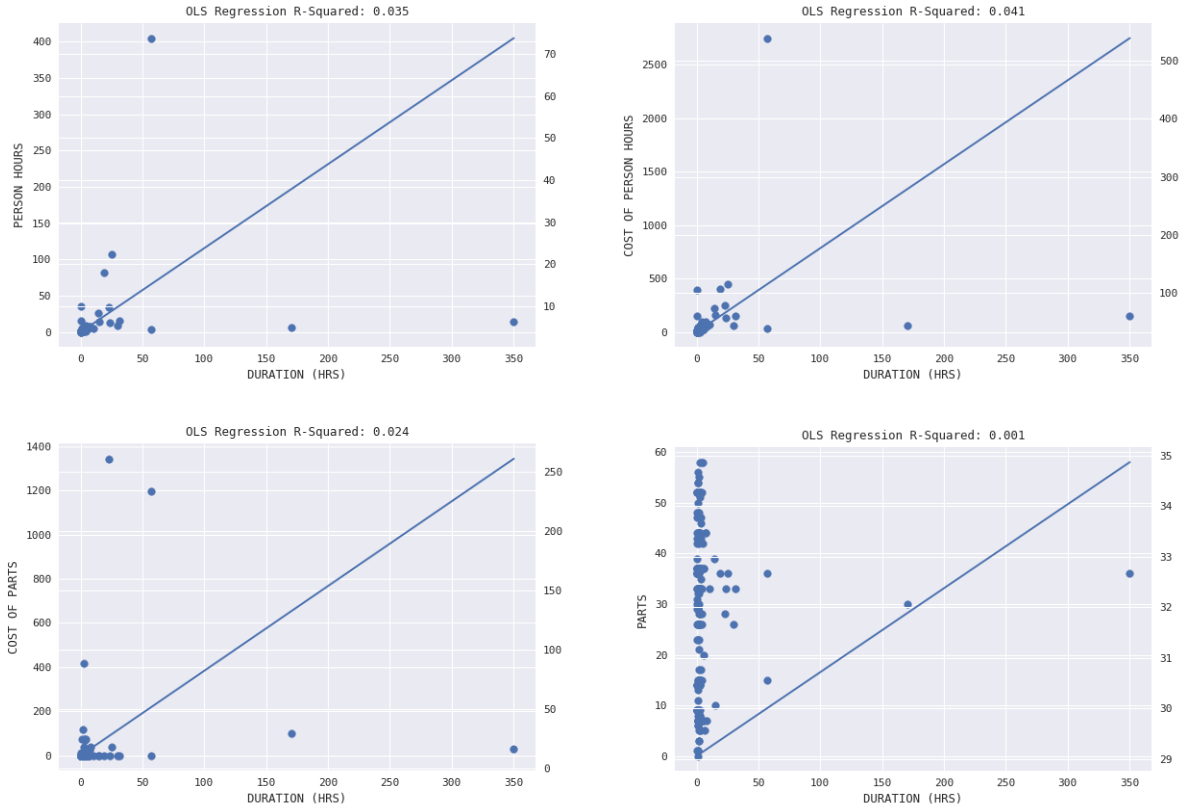
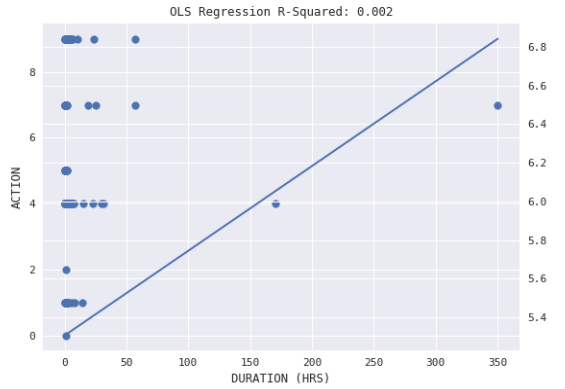
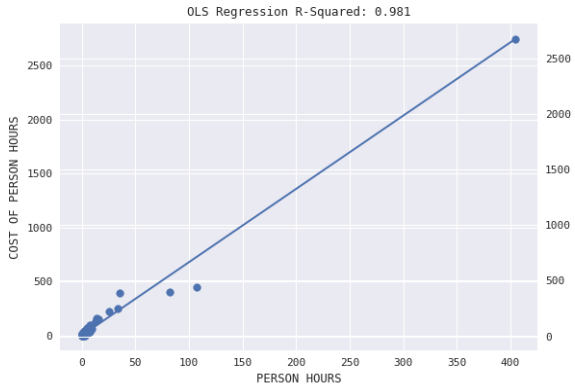


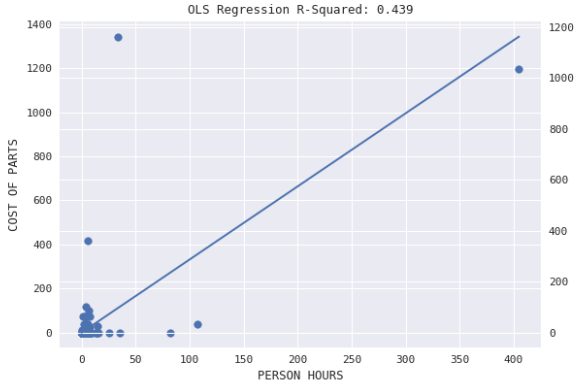
Figure 6: Frequency of machine fault types in a time range of ten years

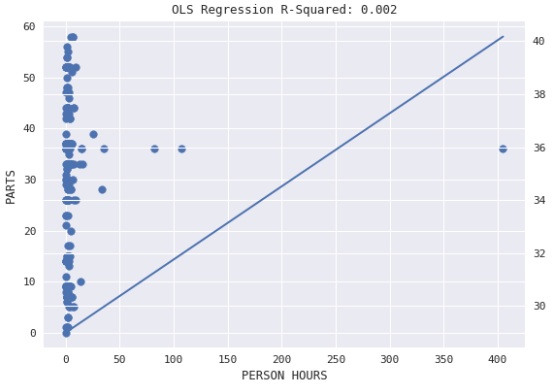
In figure 7, the linear regression model between different variables of the dataset is provided. The variables used for this task are DURATION, PERSON HOURS, COST OF PERSON HOURS, COST OF PARTS, PARTS and ACTION. For the validation of the models, the metric of coefficient of determination, denoted R^2 or r^2 and pronounced "R squared", is used. The coefficient of determination is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). The coefficient of determination ranges from 0 to 1, where the closer to 1 the better the fit of the linear model on the data.

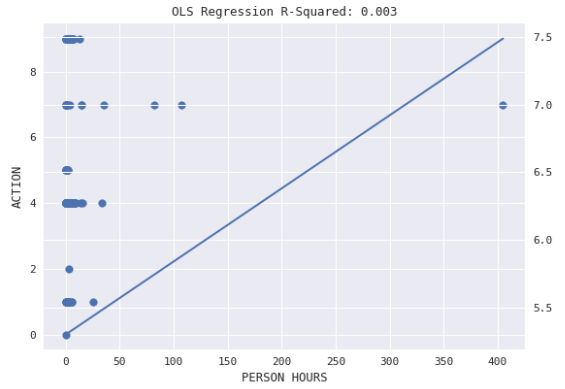


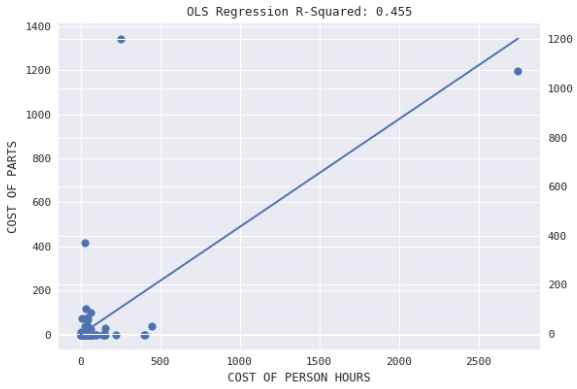

OLS Regression R-Squared: 0.002
ACTION
DURATION (HRS)

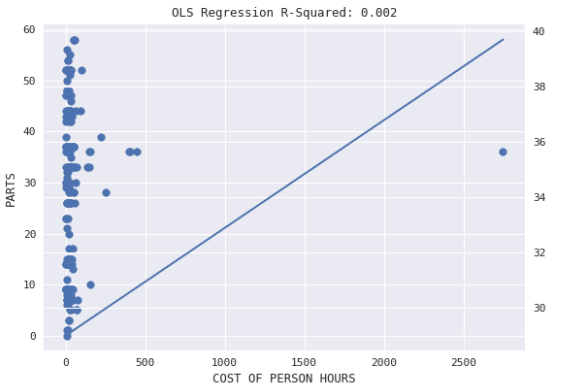

OLS Regression R-Squared: 0.981
COST OF PERSON HOURS
PERSON HOURS

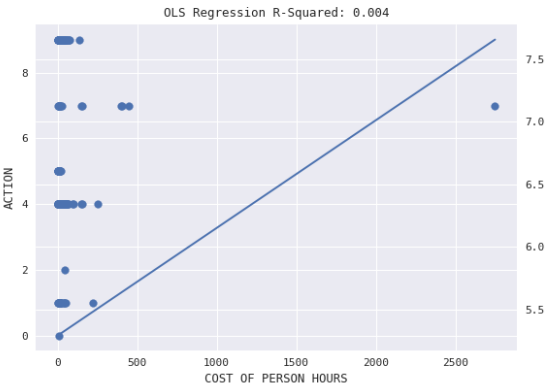

OLS Regression R-Squared: 0.439
COST OF PARTS
PERSON HOURS

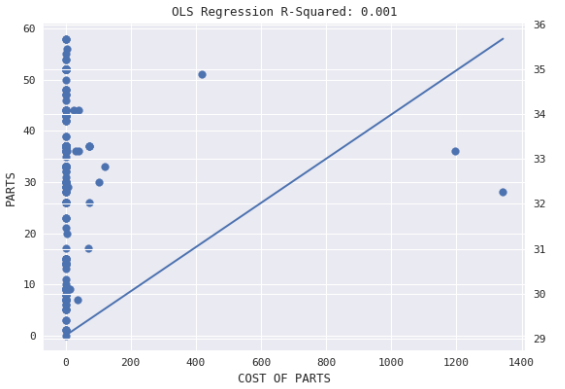

OLS Regression R-Squared: 0.002
PARTS
PERSON HOURS


OLS Regression R-Squared: 0.003
ACTION
PERSON HOURS


OLS Regression R-Squared: 0.455
COST OF PARTS
COST OF PERSON HOURS


OLS Regression R-Squared: 0.002
PARTS
COST OF PERSON HOURS


OLS Regression R-Squared: 0.004
ACTION
COST OF PERSON HOURS



Model	Variable 1	Variable 2	R-squared
1	PERSON HOURS	DURATION	0.035
2	COST OF PEARSON HOURS	DURATION	0.041
3	COST OF PARTS	DURATION	0.024
4	PARTS	DURATION	0.001
5	ACTION	DURATION	0.002
6	COST OF PEARSON HOURS	PERSON HOURS	0.981
7	COST OF PARTS	PERSON HOURS	0.439
8	PARTS	PERSON HOURS	0.002
9	ACTION	PERSON HOURS	0.003
10	COST OF PARTS	COST OF PERSON HOURS	0.455
11	PARTS	COST OF PERSON HOURS	0.002
12	ACTION	COST OF PERSON HOURS	0.004
13	PARTS	COST OF PARTS	0.001
14	ACTION	COST OF PARTS	0.000
15	ACTION	PARTS	0.000

Figure 7: Linear regression models among different variables of use case dataset. R-squared metric for each linear regression model.

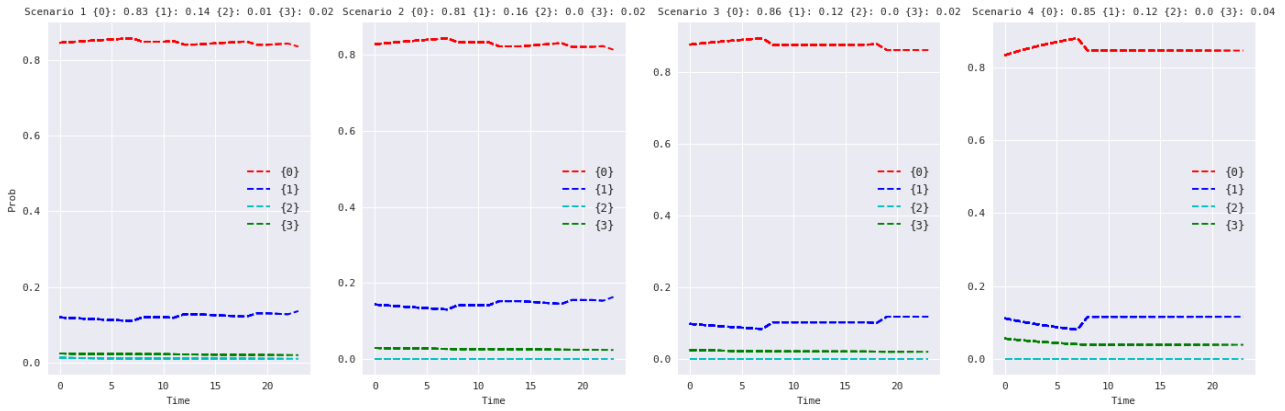
From the calculation of r-squared metric for the 15 linear regression models, the best model is found for the variables COST OF PERSON HOURS and PERSON HOURS (0.981) and the models between COST OF PARTS and PERSON HOURS (0.439) and COST OF PARTS and COST OF PERSON HOURS (0.439) are the second best. The overall outcome is that the approach of linear regression for prediction of future events does not provide reliable predictions and outcomes.

6.1.2.2 Probability model

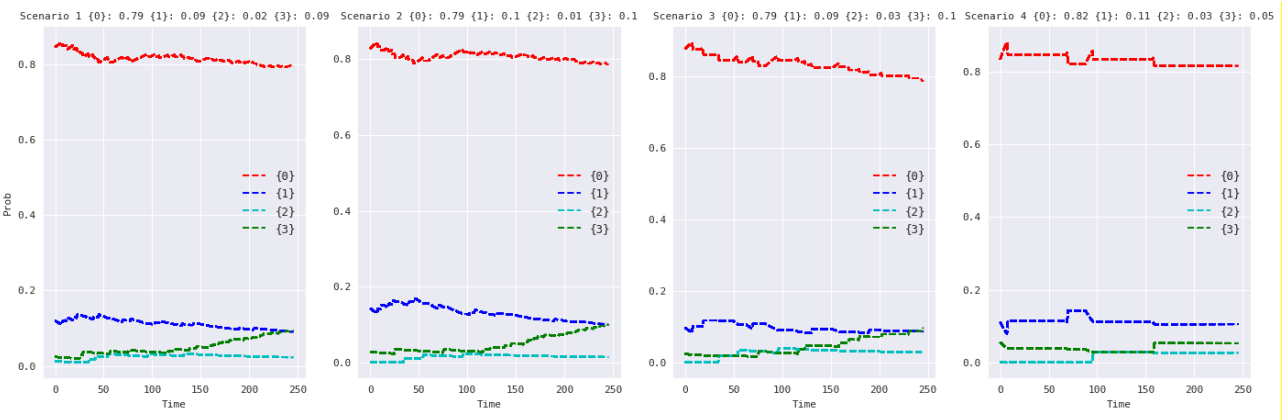
For the calculation of the probabilities of an upcoming event or machine fault (no fault, electrical, hydraulic and mechanical) per day we provide a visual analytics approach where the probabilities are calculated based on scenarios prior to that fault. The concept behind this analytic is to provide reliable calculations based on four scenarios: scenario 1 – provide the probability of an event to happen (no fault, electrical, hydraulic and mechanical) after one (1) day of no-fault, scenario 2 – provide the probability of an event to happen (no fault, electrical, hydraulic and mechanical) after two (2) days of no-fault, scenario 3 – provide the probability of an event to happen (no fault, electrical, hydraulic and mechanical) after five (5) days of no-fault, scenario 4 – provide the probability of an event to happen (no fault, electrical, hydraulic and mechanical) after ten (10) days of no-fault. For example, in figure 8(a), for scenario 3 (event after 5 days of no fault) the probability of no-fault is 0.86, the probability of an electrical fault is 0.12, the probability of a hydraulic fault is 0.0 and the probability of a mechanical fault is 0.02, while for scenario 4 (event after 10 days of no fault) the probability of no-fault is 0.85, the probability of an electrical fault is 0.12, the probability of a hydraulic fault is 0.0 and the probability of a mechanical fault is 0.04. After 220 time moments (days) the probabilities are: for scenario 3 (event after 5 days of no fault) the probability of no-fault is 0.79 the probability of an electrical fault is 0.09, the probability of a hydraulic fault is 0.03 and the probability of a mechanical fault is 0.1, while for scenario 4 (event after 10 days of no fault) the probability of no-fault is 0.82 the probability of an electrical fault is 0.11, the probability of a hydraulic fault is 0.03 and the probability of a mechanical fault is 0.05 (see Figure 8(b)). After 280 time moments (days) the probabilities are: for scenario 3 (event after 5 days of no fault) the probability of no-fault is 0.79 the probability of an electrical fault is 0.1, the probability of a hydraulic fault is 0.02 and the probability of a mechanical fault is 0.09, while for scenario 4 (event after 10 days of no fault) the probability of no-fault is 0.79 the probability of an electrical fault is 0.14, the probability of a hydraulic fault is

0.02 and the probability of a mechanical fault is 0.05 (see Figure 8(c)). The initialization length (parameter) for this visual analytic is set at 100 days and the is functional for both historical data or in real time.

(a)



(b)



(c)

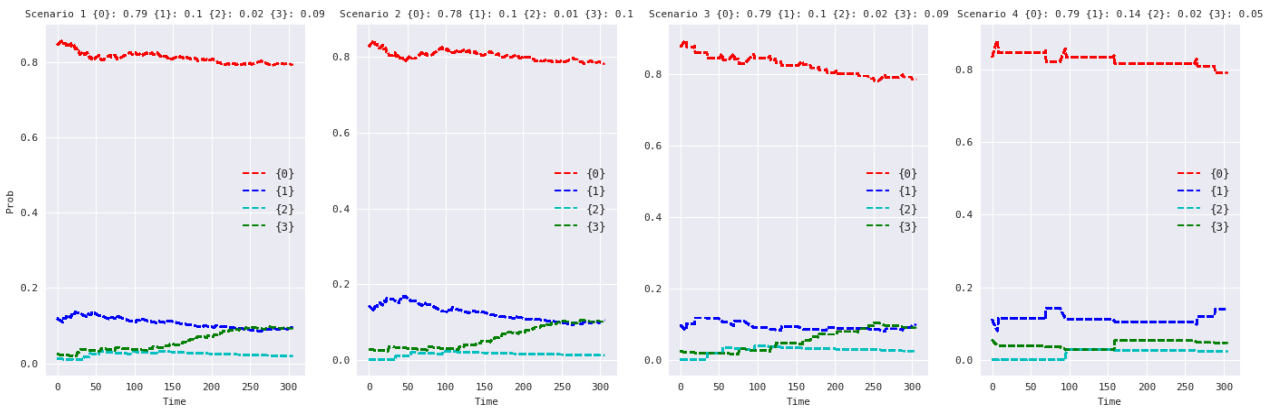


Figure 8: Visualizations of various moments of the probability visual analytic.

6.1.3 Vibration Sensor Deployment

The previous described approach is based on historical data which have been provided by KLEEMANN and they are related to previous faults of the Bossi machine. However, as the Bossi machine has no built-in sensors it was unable to provide live data for further analysis in order to enhance the efficiency of predictions. After a visit at KLEEMANN's shop-floor and discussions with its technicians about the Bossi and its performance before it breaks, the installation of a vibration sensor was considered as the most promising approach in order to get useful data for the prediction of future failures.

Figure 9: Vibration sensor position at Bossi machine

Primary data for Vibration are coming from a LIS3DH accelerometer controlled by ESP-32 Microcontroller/WiFi module, mounted on a Bossi Motor. The communication between the sensor and ESP-32 is done via SPI protocol. The data are 500 samples of accelerations from three axis(x,y,z) at a sampling rate of 5 kHz. The unit of measurement is m/s^2. Restrictions on the ESP32 memory confine us to the number of samples. Furthermore the WiFi buffer size restricts the measurement cycle in one sampling every 10-15 seconds depending on WiFi connectivity and sending the data in groups of three packets, each one corresponding the samples of each axis. Data are processed to a json object format inside the microcontroller and are sent via MQTT to a broker for further processing.

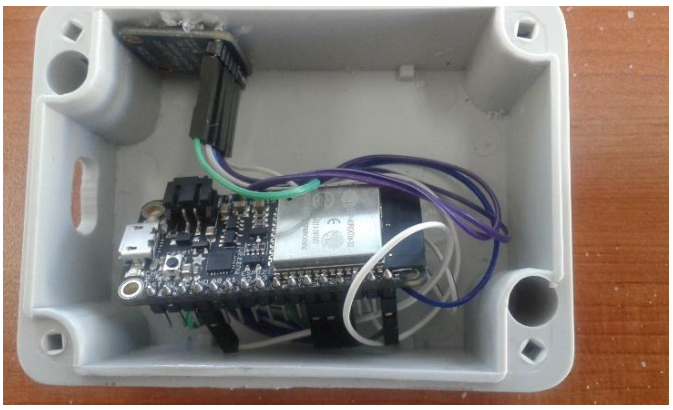


Figure 10: Vibration Sensor

6.2 UC – KLE 3 Scrap Metal and Recyclable Waste Transportation

6.2.1 Fill Level Sensor Deployment

The UC – KLE 3 Scrap Metal and Recyclable Waste Transportation is triggered by a full bin in KLEEMANN's production. In order the Prediction engine to be able to estimate the fill level in the future and propose to a worker the optimal path in order to transport the wastes to a central bin outside of the production line, a fill level sensor for the internal bins has been designed and developed.

For the fill sensor data the VL53L0X micro-Lidar Time of flight Sensor was used to capture raw measurements of distance of the waste heap from the deployment point. The sensor and the communication is controlled by the STM32L053c8t6 low power microcontroller. The microcontroller-sensor communication is carried out through an I2C bus that can serve multiple sensors. The microcontroller also checks for faulty

measurements and repeats the measurement process again and if the measurement is faulty sends an error flag. The wireless communication is carried out via the sx1272mbas LoRa Module for STM.

Figure 11: Recycling bins in KLEEMANN's production line

Data are transferred via the LoRa low power protocol to the LoRa Gateway. The gateway to be used is the LoRank 8. It has to be noted that LoRa protocol allows transmission of only very small packets of data, so this means that only raw measurements of distance are transferred along with the measurement of battery level and id of bins. The gateway that is connected to the internet via Ethernet or Wi-Fi publishes those data on an MQTT topic on the cloud and a listener, connected to the same broker as the gateway, reprocesses the data to convert the raw distance measurement into a fill percentage and derive the json object format to be sent on the destination platform.

6.2.1.1 Slope Statistic Profile (SSP)

For the analysis of fill level sensor data, we will apply the Slope Statistic Profile methodology. Slope Statistic Profile (Vafeiadis, 2011), denoted hereafter as SSP, is a method that detects the single structural break T, denoted hereafter as incident, in a time series using a standard parametric linear trend test, denoted hereafter as t-statistic. The t-statistic is calculated on overlapping sliding data windows of size w with sliding step one, along the time series. By this way we obtain the profile of the t-statistic, denoted as $\{\widetilde{U}_i\}$, for $i = 1 + [\frac{w}{2}], \dots, n - [\frac{w}{2}]$, where $[x]$ is the integer part of x. The form of this profile depends on time series characteristics, i.e. the strength of the autocorrelation, the distribution of the residuals, the strength of the linear trend, as well as, the size of the sliding data window w. The profile of the t-statistic $\{\widetilde{U}_i\}$, exhibits small fluctuations (glitches) due to edge effects of the local data windows and therefore the profile curve is smoothed using a zero-phase filter of a small order, set to about 5% of w. Such a small filter order removes the glitches in $\{\widetilde{U}_i\}$, but maintains its original signature. The smoothed value of $\widetilde{U}_i$ is denoted as U_i and referred to as U-profile. In a short presentation of the method below, we will assume the situation from no trend to a positive trend. Other types of change between no trend and trend can be treated similarly.

The t-statistic for the parametric linear trend test is $t = {\hat{\beta}}/{s(\hat{\beta})} \sim t_{w-2}$, where $\hat{\beta}$ the trend parameters and $s(\hat{\beta})$ is the standard error of $\hat{\beta}$. The null hypothesis of no trend is rejected at the significance level a if $|t| \geq t_{w-2,1-a/2}$.

SSP methodology will be applied on time series of fill level so as to provide notifications and warnings regarding the fill level of the recycle bins. SSP methodology hasn't been tested yet on use case data. The research is ongoing and the results of the application of the method will be presented on the next version of this document.

6.2.2 Recyclable Waste Transportation

A common way to mathematically model and represent road networks, in order to deal with problems such as the shortest path problem, is graphs G that are composed by sets of nodes N and sets of edges E. In graph theory, the shortest path problem is the problem of finding a path between two nodes in a graph such that the sum of the weights of its constituent edges is minimized. There are several works trying to solve this problem, the well-known Dijkstra's algorithm solves the single-source shortest path problem in $O(V^2)$ (worst case computational complexity), while there are various implementations of Dijkstra's algorithm that reduce the computational cost (Ahuja, 1993; Cherkassky,1996; Zhan, 1998).
An extension of the Dijkstra's algorithm is the A* search algorithm (Hart, 1968) which achieves better performance by using heuristics to guide its search. Moreover the Bellman–Ford algorithm (Kenneth, 2003) solves the single-source problem if edge weights may be negative. There are, also, the Floyd–Warshall (Bang-Jensen, 2000) algorithm which solves all pairs' shortest paths and the Johnson's algorithm which solves the same problem, and may be faster than Floyd–Warshall on sparse graphs. Additional algorithms and associated evaluations may be found in Cherkassky, Goldberg & Radzik (Cherkassky, 1996).

We will apply the Dijkstra and the A algorithm so as to find the optimal route for the recyclable bins in industrial environment. The research is ongoing and the results of the application of the method will be presented on the next version of this document.*

6.3 UC – BSL 2 Predictive Maintenance

6.3.1 Methodologies

6.3.1.1 Heatmaps

A **heat map** (or **heatmap**) is a graphical representation of data where the individual values contained in a matrix are represented as colors. The term 'heat map' was originally coined and trademarked by software designer Cormac Kinney in 1991, to describe a 2D display depicting financial market information (US Patent, 1993). Heat maps originated in 2D displays of the values in a data matrix. Larger values were represented by small dark gray or black squares (pixels) and smaller values by lighter squares.

For this use case, we have applied heatmaps to describe the metric of Pearson's correlations among the variables of the dataset.

6.3.1.2 Local Outlier Factor (LOF)

In anomaly detection, the local outlier factor (LOF) is an algorithm proposed by Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and Jörg Sander in 2000 for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbours (Breunig, 2000). The local outlier factor is based on a concept of a local density, where locality is given by k nearest neighbours, whose distance is used to estimate the density. By comparing the local density of an object to the local densities of its neighbours, one can identify regions of similar density, and points that have a substantially lower density than their neighbours. These are considered to be outliers. The local density is estimated by the typical distance at which a point can be "reached" from its neighbours. The definition of "reachability distance" used in LOF is an additional measure to produce more stable results within clusters.

LOF methodology hasn't tested yet on use case data. The research is ongoing and the results of the application of the method will be presented on the next version of this document.

6.3.1.3 Slope Statistic Profile (SSP)

As in the case of the analysis of fill level sensor data, the Slope Statistic Profile methodology will be applied so as to detected deviations (if possible) from a normal condition (no linear trend existence to linear trend, either upward or down ward) (see Section 6.2.1.1)

SSP methodology hasn't tested yet on use case data. The research is ongoing and the results of the application of the method will be presented on the next version of this document.

6.3.2 Application

In this section the application of heatmaps is provided and briefly described below.

6.3.2.1 Heatmaps

For the specified use case, we focus on the early detection of motor failure in the reflow ovens at BSL. A number of 20 fans per oven are measured. The workflow of the ovens is divided on three different conditions: the *wake-up*, the *work under load* and the *cooldown.* There were three available measurements the *set point*, the *present value* and the *output power*. The variable with the most useful information is the *present value* of the real temperature measured by the thermometer. As for the other variables *set point* is steady and *output power* is always zero. Thus, the heatmaps were created based on *present value* variable measurements.

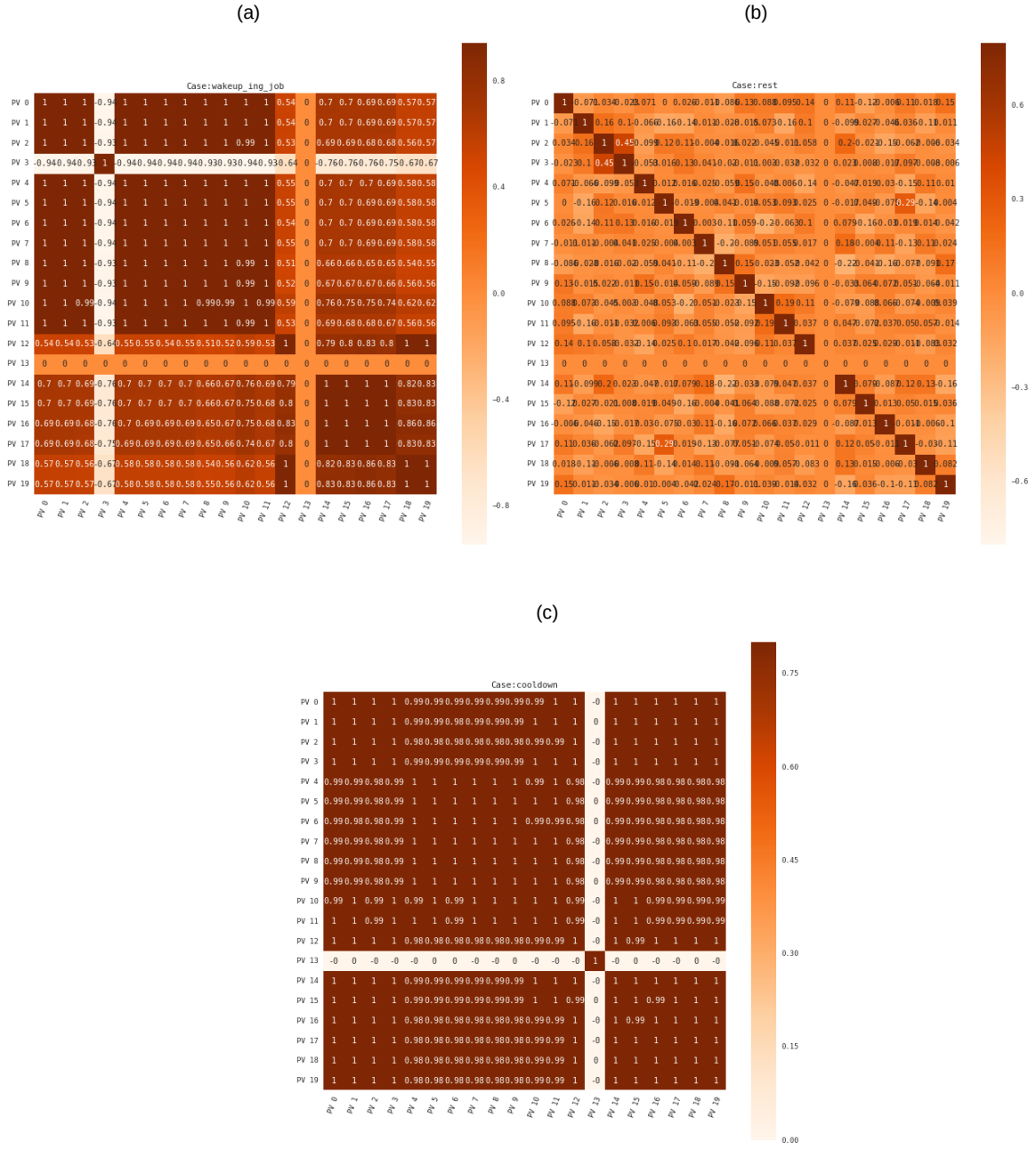


Figure 12: Correlation heatmaps overlay for the different condition of the oven: (a) wake up, (b) work under load and (c) cooldown.

For the first set of the available data the *present value* measurements per condition were: *wake up* (136), *work under load* (131) and *cooldown* (19). Thus, based on the current measurements, the pearson's correlations were calculated among ovens' fans for the different conditions and the correlation heatmap of each condition are given on figure 12. On the corrlation heatmaps, dark red areas show a direct positive corretation, whiter areas show an indirect negative correlation (or anti-correlation), while orange areas indicate no apparent correlation between variables. A significant output of the correlation heatmaps

described on figure 12 is that during *wake up* (see Figure 11(a)) and *cooldown* (see Figure 12(c)) conditions variables are corrlated on an obvious manner as they begin stop functioning concurrently. On the other hand, on the *work under load* condition (see Figure 12(b)), fans work independently.

7 Next Steps

The future work at Task 3.3 Simulation and Forecasting in Production and Logistics will be mainly focused at procedures related to:

- For the UC-BSL-2 Predictive Maintenance we plan to develop and test a visual analytic that shows and provides the correlation heatmaps of the 20 fans per oven in real time, aiming to detect situations that point to a motor failure. Moreover, we plan to apply and test both local outlier factor and slope statistic profile methodologies on the fans measurements aiming to detect anomalies and possible malfunction of the machine.
- For the UC-KLE-1 Maintenance decision support the data coming from the installed vibrometer will be used in order to enhance the efficiency of the predictions.
- Further research and development will be conducted for UC-KLE-2 Delayed Process Step and UC-KLE-3 Scrap Metal and Recyclable Waste Transportation
- Evaluate a possible collaboration of Simulation and forecasting tool with Monitoring Framework of Task 3.1 for some of BSL intra-factory scenarios
- Dockerization of the future developed algorithms in order the prediction results to be available to DSS via DFM component

The results of the previous mentioned procedures of Task 3.3 will be reflected at the next and final version of this report. This work will be presented in D3.5 Computational Modelling, Simulation and Prediction in Production II in M28.

8 Conclusions

In conclusion, this deliverable describes the effort spent from M4 to M18 and represents the current status of T3.3 – Simulation and Forecasting in Production and Logistics of WP3. More precisely, this report documents the results in the part of production as the logistics results are documented in D3.6 - Computational Modelling, Simulation and Prediction of Logistics I which will be submitted in parallel with the current report. The complete work of Task 3.3 related to the production will be presented in D3.5 Computational Modelling, Simulation and Prediction in Production II in M28.

In the first steps of Task 3.3 the research and development were focused on the architecture design, the analysis of existing pilot partners' data, the selection of the proper methodologies for this data, and the development of the aforementioned methodologies for the use cases with the highest priority.

Based on the analysis of the applied methods, the Probability Model for KLEEMANN's maintenance decision support enhance the efficiency of the DSS component by providing the possibility of future faults (mechanical, hydraulic and electrical). The probabilities of fault are available to DSS via the DFM component. A Slope Statistics Profile methodology will provide the estimation of fullness from indoor bins in the KLEEMANN's shop-floor. Based this estimation the simulation and prediction engine will provide the optimal route for the recyclable bins transportation by applying Dijkstra and the A* algorithm. Furthermore, the application of Correlation Heatmaps, Local Outlier Factor and Slope Statistic Profile methodologies provides anomalies' detection for the BSL predictive maintenance scenarios in order to support the decision making process.

Finally, as it is perceived in this document many of the selected and applied methodologies are still in testing phase and their results are under evaluation. As soon as, the evaluation phase will be over, the outcome of the simulation and prediction engine will be more concrete. Moreover, the developed methodologies will be enriched based the live data that will be available from the deployed sensors and new methodologies can be examined and applied based this data as well.

9 List of Figures and Tables

9.1 Figures

9.2 Tables

10 References

(Ahuja, 1993) R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. Network Flows: Theory, Algorithms, and Applications. Prentice Hall, 1993

(Bang-Jensen, 2000) Bang-Jensen, Jørgen; Gutin, Gregory (2000). "Section 2.3.4: The Bellman-Ford-Moore algorithm". Digraphs: Theory, Algorithms and Applications (First ed.). ISBN 978-1-84800-997-4.

(Breunig, 2000) Breunig, M. M., Kriegel, H.-P., Ng, R. T., Sander, J., (2000). LOF: Identifying Density-based Local Outliers (PDF). Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD. pp. 93–104.

(Cherkassky, 1996) Cherkassky, Boris V.; Goldberg, Andrew V.; Radzik, Tomasz (1996). "Shortest paths algorithms: theory and experimental evaluation". Mathematical Programming. Ser. A. 73 (2): 129–174

(Gauss 1809) Gauss, C.F., Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum. (1809)

(Gauss, 1821) Gauss, C.F., Theoria combinationis observationum erroribus minimis obnoxiae. (1821/1823)

(Hart, 1968) Hart, P. E.; Nilsson, N. J.; Raphael, B. (1968). "A Formal Basis for the Heuristic Determination of Minimum Cost Paths". IEEE Transactions on Systems Science and Cybernetics SSC4. 4 (2): 100–107

(Kenneth, 2003) Kenneth H. Rosen (2003). Discrete Mathematics and Its Applications, 5th Edition. Addison Wesley. ISBN 0-07-119881-4.

(Legendre, 1805) Legendre, A.M,. Nouvelles méthodes pour la détermination des orbites des comètes, Firmin Didot, Paris, 1805. “Sur la Méthode des moindres quarrés” appears as an appendix.

(US Patent, 1993) "United States Patent and Trademark Office", 1993-09-01.

(Vafeiadis, 2011) Vafeiadis, T., Bora-Senta, E., Kugiumtzis, D., (2011). ‘Estimation of linear trend onset in time series’. Simulation modelling – Practice and Theory, Vol. 19, Issue 5, p 1384 - 1398.

(Zhan, 1998) F.B. Zhan and C.E. Noon. Shortest Paths Algorithms: An Evaluation Using Real Road Networks. Transportation Science, 32:65–73, 1998.