Data-based fault diagnosis model using a Bayesian causal analysis framework

Thierno M.L. Diallo, Sébastien Henry, Yacine Ouzrout, Abdelaziz Bouras

DATA-BASED FAULT DIAGNOSIS MODEL USING A BAYESIAN CAUSAL ANALYSIS FRAMEWORK

THIERNO M. L. DIALLO[†]
QUARTZ Laboratory, Supmeca - Superior Engineering Institute of Paris, France
thierno.diallo@supmeca.fr

SÉBASTIEN HENRY
DISP laboratory, University of Lyon,
University Lyon 1, France
sebastien.henry@univ-lyon1.fr

YACINE OUZROUT
DISP laboratory, University of Lyon,
University Lyon 2, France
yacine.ouzrout@univ-lyon2.fr

ABDELAZIZ BOURAS
Qatar University, Computer Science and Engineering Department,
College of Engineering,
Doha, Qatar
abdelaziz.bouras@qu.edu.qa

Abstract: This paper provides a comprehensive data-driven diagnosis approach applicable to complex manufacturing industries. The proposed approach is based on the Bayesian network paradigm. Both the implementation of the Bayesian model (the structure and parameters of the network) and the use of the resulting model for diagnosis are presented. The construction of the structure taking into account the issue related to the explosion in the number of variables and the determination of the network's parameters are addressed. A diagnosis procedure using the developed Bayesian framework is proposed. In order to provide the structured data required for the construction and the usage of the diagnosis model, a unitary traceability data model is proposed and its use for forward and backward traceability is explained. Finally, an industrial benchmark – the Tennessee Eastman (TE) process – is utilized to show the ability of the developed framework to make an accurate diagnosis.

[†] Corresponding author at: Supméca - Institut supérieur de mécanique de Paris
3 rue Fernand Hainaut, 93400 Saint-Ouen, France
Tel. : 0033 (0) 6 49 73 95 36
Corresponding author's primary affiliation: DISP laboratory, University of Lyon, University Lyon 1, France.

Keywords: Fault Diagnosis; Continuous Improvement; Manufacturing System; Unitary Traceability; Bayesian Network.

1. Introduction

Due to the continuing growth of the industrial processes instrumentation and to the development of control systems such as Manufacturing Execution System (MES) and automatic data collection tools and technologies (RFID, Data Matrix, etc.), large amounts of data are generated and collected during manufacturing processes [1-5]. Because of the speed with which this data is collected and given the amount of the concerned data and its diversity, it can be labelled as "Big Data" [6]. Indeed, Big Data is generally characterized by 3 dimensions: the size, the variety and the velocity [7-9]. Some authors consider that Big Data are also characterized by other dimensions than volume (size), velocity and variety dimensions. Among the other characteristic dimensions cited, there are [7, 9, 10]: value, veracity, variability, complexity and decay. The size of Big Data depends on the considered sector according to the common sizes of datasets and the existing collection tools and technologies [11]. J. Manyika et *al.* define Big Data as "*dataset whose size is beyond the ability of typical database software tools to capture, store, manage and analyze"* [11]. In the manufacturing sector, diagnosis functions involving several facilities in some cases concern hundreds of parameters with millions of possible records. According to General Electric, which developed Predix, a cloud platform for the "Industrial Internet", the order of magnitude of industrial processes parameters range from thousand, for a factory, to 1 million, on the scale of a company [12]. The amount of data collected by one of our industrial partners is of the same order of magnitude.

All sectors are experiencing this explosion in the amount of collected and stored data. The collection and storage of such data is no longer a technical problem in itself. However, the full exploitation of this data remains a challenge that researchers and industrialists are facing. In many industries, this data is just archived [13, 14]. However, industry is broadly aware of the usefulness of this data and is more and more likely to engage in its exploitation.

- There are analysis methods using only data. These methods often work like a "black box" and the results provided are not always explainable. Furthermore, these methods using only data do not take into account expert knowledge that can simplify the methods or improve the obtained results. The focus of this work lies within the exploitation of data collected both on the process and on products with the integration of expert knowledge. In particular, we are interested in the diagnosis of non-compliances of products using the Bayesian network paradigm.

Several industrial diagnostic methods using Bayesian models have been proposed in the literature. Among these may be mentioned: [15], [16], [17], [18], [19] and [20]. However, most of these works deal with simple networks with a few dozen nodes and the obtained diagnostic models are applied to a device or a limited set of equipment. The diagnosis approaches we

are interested in in this work are global approaches covering the whole production processes for a product. It involves analyzing process and product data across the entire value chain to identify possible causes of a product defect or a performance degradation. When using this type of approach, the number of variables is in the hundreds or more. Under these conditions, the definition of all the causal relationships between these variables cannot be the sole responsibility of expert. Expert knowledge alone is therefore insufficient to completely define the network.

In this work we propose a comprehensive approach to develop diagnosis functions applicable to manufacturing industries. This approach uses process historical data, considered as big data. The data we are interested in is traceability data, especially unitary traceability data. Our causal analysis framework including data pre-processing, network construction, parameters estimation and inference calculation was applied on an industrial case study.
Our contribution is twofold:

- The construction of a diagnosis framework based on the Bayesian network paradigm: the definition of the network's structure and the learning of its parameters from historical data,
- The use of the developed Bayesian framework for diagnosis: a diagnosis procedure using forward and backward traceability processes

The construction and operation of the Bayesian network require ordered data. We have therefore proposed a unitary traceability data model.

The remainder of this paper is organized as follows: industrial diagnostics and traceability are presented in Section 2. The addressed problem and the proposed approach are highlighted in Section 3. Section 4 presents our proposed Bayesian causal model and its uses. The proposed data model and its usage for traceability are described in Section 5. In Section 6, we discuss on how our proposed Bayesian Network model allow to deal with Big Data challenges followed by the case study in section 7. We conclude this work in Section 8.

2. Industrial Diagnostics and Traceability

The distinction between existing diagnostic approaches can be based on different criteria:

- The dynamic of the system to supervise: state space and the evolution of the state,
- The nature of the exploited information: qualitative or quantitative, analytical or heuristic, structural, functional and / or temporal,
- The implementation of the approach: online or offline, centralized or decentralized,
- etc.

According to [21], the form of the process prior information is the most distinctive criteria of diagnostic approaches. Considering this criteria, three families of approaches can be distinguished: Model-based approaches, knowledge-based approaches and data-driven approaches.

Model-based approaches (qualitative or quantitative) employ the system normal or faulty functioning model to generate residuals. A fault is detected from the analysis of this residual. If the symptom analysis permits, we can then diagnose the detected fault. The model might be expressed using analytical equations, automata, petri nets, etc.

In the family of knowledge-based approaches, there are several methods including FMEA, fault tree and expert systems. The main feature of this family of approaches is that knowledge used here is obtained empirically.

The third family consists of process historical data-driven approaches. These approaches apply classification methods on the system heuristic data to achieve detection and diagnosis functions. These process-history-based approaches proceed by pattern recognition that can be achieved in two ways: supervised classification and unsupervised classification. Classifiers can be statistical (e.g. Statistical Process Control; Principal Component Analysis/Partial Least Squares, Independent component analysis, Fisher discriminant analysis, Subspace aided approach, Bayesian classifier. See examples in [22, 23]) or non-statistical (e.g. Template Matching, Neural Networks).

The impossibility of obtaining a reliable model or completed reliable expert knowledge and the difficulty to manage multiple and unobservable faults, model uncertainties, noise, and unknown disturbances make the use of model-based and knowledge-based approaches difficult for complex systems.

Diagnosis approaches based on process historical data have been proposed in several works in the literature. These approaches consist on operating equipment's historical data in order to diagnosis a default. Only equipment parameters are considered for most of the time. The use of such approaches based on historical data is often done when the models of the nominal behaviour and the faulty behaviour of the system are not available. Even when these models exist, the diagnosis of multiple or unobservable faults remains a challenging task. Indeed, the diagnosis function include fault isolation (i.e. locate the fault) and fault identification (i.e. identify the fault type). The diagnosis occurs after fault detection. The detection tasks requires the system nominal behaviour model and the diagnosis requires the model of the faulty behaviour or the fault symptoms. In some cases, one symptom might be the result of two or many different faults and some symptoms or faults might be unobservable. In these cases the use of only process data may not be sufficient. Furthermore, certain faults on product should not be due to the production process but rather to transportation or storage processes. In addition, unobservable or multiple faults which cannot be identified with certainty based solely on process data, may be so by considering their potential effects on products. This is why it is necessary to adopt a comprehensive approach. In our work, the use of product parameter data in addition to

those of the process parameter data for diagnosis of multiple and unobservable faults was addressed.

The diagnosis procedure of data-driven approaches include two steps: data acquisition and pre-processing, and decision making. We address in this work the decision making process by analyzing the collected data. We aim to diagnosis default at the parameters level. We would like also to give explanation of the conclusion of the method to the user. A lot of data-driven fault detection and diagnosis methods have been proposed. Among these methods, there are: Principal Components Analysis (PCA) [24, 25], Clustering [26-29], Neural Networks [4, 30] and Bayesian Networks [17, 31, 32]. We compared the 4 previous methods in order to select the formalism to be used to model our diagnostic approach. Four comparative criteria were used: the consideration of uncertainties, the possibility of using both data and expert knowledge, the possibility for the user to know the existing causal relationships between the different parameters (semantic) and the resources needed to implement the method (time needed for modelling tasks, amount of memory consumed, necessary computing time,…) (see Table 1).

Methods	Consideration of uncertainties	Incorporation of expert knowledge	Semantics	Resources needed
PCA	-	-	+	--
Clustering	-	-	-	-
Neural Network	-	+	-	--
Bayesian Network	+	+	+	---

Table 1: Comparative analysis of data-driven fault detection and diagnosis methods ("-"= NO or NEGATIVE, "+" = YES or POSITIVE)

Among the studied methods, the BN formalism is the only framework that explicitly incorporates uncertainties and allow to exploit both data and expert knowledge. The semantic of the BN makes it possible to understand the causal mechanism linking a symptom to its root cause. BN nonetheless requires more resources compared to the other methods. Bayesian theory has been chosen to set up our causal framework. The exploited data is obtained through a process and product traceability system.

According to ISO 9000: 2015, traceability is the "ability to trace the history, application or location of an object". Depending on the industry or the desired level of precision, the traced object may be a shipment (truck load, vessel, etc.), a logistic unit (pallet, container, etc.), trade item (carton, bag, etc.) or a serialized trade item, the unitary traceability (consumer unit, one product, etc.) [33]. Generally, the traceability unit is a lot of trade items. A lot can be defined from different perspectives. From the perspective of production, a lot is a set of items considered homogeneous and produced in the same process or a series of processes. From a control point of view, a lot is a defined as quantity of a product or material accumulated under conditions considered uniform for sampling purposes [34]. This is the definition of lot of trade items. This level of detail is enough to have an accurate picture of the conditions of production for batch production but not for job production and flow production. For the latter two types of production, lot of trade items might not be homogeneous. In the case of a traceability by lot, when a non-quality requiring a recall for example is detected, the entire lot is recalled even in the case of a production that is not batch-type. Yet as we have explained above, items of the same lot of trade items may not be homogeneous with respect to certain production parameters. Thus, when a defect is detected on an article, it is not certain that all the other items from the same lot of trade items have the same defect. The unitary traceability, however, enables a serialized unique identification at the item level and allows to know accurately the process parameters values of each item. The principle of the unitary traceability is to identify each item individually by allocating to it a serial number. This identifier must be unique and accompany the article throughout its life cycle. The process parameters and life cycle events must be associated with each article through its identifier [35]. The knowledge of manufacturing conditions of each item through the process and product traceability will allow to better control the process and to react more quickly in case of non-compliance detection. The process parameters to trace are raw materials and ingredients used while making up the product, the transformation processes and distribution historical and location of the product after delivery. In addition to these process data, the product features should be recorded. The volume of data to collect (process and product data) and their diversity are very large. This data, which has the Big Data characteristics, requires dedicated resources and management tools. The processing of such data (validation, reconciliation, aggregation, disaggregation, etc.) presents some challenges[36]. The processing of the collected data for industrial diagnostics purposes is our aim in this research work.

3. Problem Statement and Proposed Approach

In this research work, we address product non-quality diagnosis produced by complex and challenging process industries. Let us consider the glass industry for illustrative purpose. The glass production process is mixed: the first steps are continuous and several discrete lines (from 4 to 8) then emerge from the continuous line. The physico-chemical processes that occur during the manufacture are difficult to model. High production rates (production order ~ up to 1 million bottles) and the large number of possible faults (~200) do not allow to detect all product defaults. In some industries, different articles are produced on the discrete lines. However, as these lines come from the same continuous line, the

parameterization of a given line will depend on the production ongoing on the other lines. If, in addition, there are frequent changes of production series, the process qualification became difficult to manage. If we consider 5 lines and 20 types of product, then 1, 860, 480 configurations have to be defined. Under the conditions set out above, the possibilities of using deterministic reliability engineering tools such as FMECA, cause-effect diagram and fault tree or model based diagnosis methods are very limited.

The proposed approach in this work operate process and product traceability data to diagnosis product non-qualities. The aim is to improve responsiveness towards production hazards by detecting and correcting default root causes and managing produced non-compliant product by optimizing the withdrawal of defective products. Our approach consists in two steps and it is sketched in Fig. 1: i) the causal model implementation / update and ii) the causal model use.

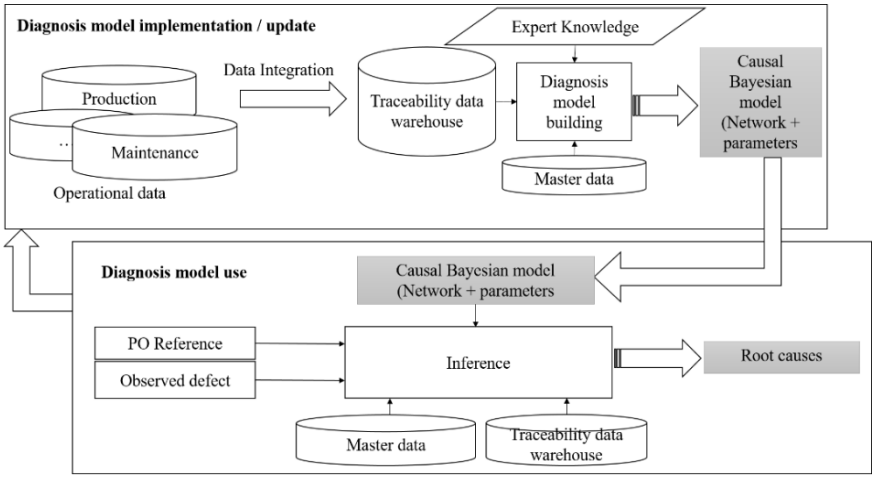

Fig. 1: Our proposed approach

For the first step, process and product historical data from different operational data bases are integrated in a traceability data warehouse. The data structure of this data warehouse should enable unitary backward and forward traceability and process default diagnosis. Then, based on traceability data, master data and expert knowledge, the causal model is implemented. The master data includes product definition (product parameter, manufacturing bill, equipment and material specification, etc.) data and production request (production rule and requirement). In order to integrate business rules and expert knowledge with data in the analysis process, Bayesian network (BN) formalism was chosen for our causal analysis framework. At one side, Bayesian networks allow to combine certain and uncertain knowledge while they allow to exploit both data and expertise on the other side. BNs reduce the combinatorial explosion in Big Data processing and allow a better exploitation of its potential. Thanks to the Bayes' theorem, the computational complexity of the joint probabilities is reduced. The BNs offer the

possibility to integrate the expert knowledge thus making it possible to contextualize the data and improve the knowledge extracted. Bayesian Networks (BN) has been used for fault diagnosis and applied for various range of systems. For example, in [37], BN are applied for fault detection and treatment in automated machines. In [19], BN are applied for diagnosing the most probable cause of fault in the operation of fuel cells. The Bayesian probabilistic framework is suitable to address the uncertainty involved in the diagnosis process and especially in the case of root cause diagnosis for multiple-simultaneous faults [17, 31]. The uncertainty involved in the diagnosis process is managed by the probabilistic modelling and calculation of the Bayesian framework. To set up a Bayesian model, two elements have to be defined: the structure of the network (nodes and arcs) and the network parameters (conditional probabilities distribution). These two issues have been addressed. We have developed an algorithm for Bayesian network structure learning optimized for industrial diagnosis application in previous research [38].
The second step of our proposed approach consists in determining the root causes of a detected non-conformity by making a causal inference using the causal model. The PO (Production Order) reference is used to determine the traceability data to employ for the inference.

In our causal framework, we consider three types of variable: Product parameters, Process characteristics and Control variables.

Product parameters allow to detect nonconformities. These are product non-qualities observed directly or indirectly through their effects. A product nonconformity can be detected within the company which manufactures or manipulates it or by the end-user. The definition of a nonconformity may be vague and imprecise. It can be due to different factors (defaults). In the glass industry for example, a bottle leak (a nonconformity) may be cause by a chipped ring, a deformed ring, a cap default, a contents default, etc. The considered control variables in our framework are materials (raw materials and consumables) properties and machine parameters. They are considered as potential root causes. They are of different types: qualitative or quantitative, continuous or discrete. They can also be measured or estimated and certain or uncertain. A control variable may be Normal or Abnormal according to its value or state. The difficulty here is that the normal state of all the control variable are unknown and this normal state may change according to product type and production condition (configuration, production history, etc.). Our objective will therefore be to determine the control variables that correlate with a given product default. An expert will then analyses and validate the results achieved by the causal model. When a new causal link between control variable and default is confirmed, this new knowledge will enrich the causal model.

The setting up and exploitation of the diagnostic model require structured process and product data.
Nowadays the collection and storage of data is no longer a problem in itself. In contrast, the transformation of data into knowledge still poses difficulties. In order to make collected data usable, actions needs to be conducted upstream on the data to collect, on how to collect

and storage the data. By observing industrial practices, we found that traceability performed internally by different departments (manufacturing, maintenance, etc.) are managed separately and reconciliation are not made between them. Another observation we made is that there is no general rule on what data to collect especially for internal traceability. We have developed a data model for unitary traceability data. The data is aggregated by production order and concerned the whole production and distribution processes. The processes are divided into segments and for each segment, the data related to process and product are collected. This proposed data model allows to know the process parameters of manufacture of every item from historical time series data sets. This data model achieves the aggregation of data scattered throughout and outside the company. Analysis of this data allows to derive contextual and meaningful information.

4. Proposed Causal Bayesian Model

In this section, we describe our proposed probabilistic causal model using Bayesian Networks (BN). BN are graphical models for reasoning under uncertainty [39]. They allow to combine, on the one hand, certain and uncertain knowledge, on the other they allow to exploit both data and expertise. A BN is Directed Acyclic Graph (DAG) represented by the pair (V, E) where V is a set of vertices and E a set of directed edges connecting vertices. It is associated with each node marginal or conditional probability distribution table of the corresponding variable. The main purpose of our model is to determine the control variables that might be responsible for the detected nonconformity.
To set up a Bayesian model, two elements have to be defined: the structure of the network (nodes and arcs) and the network parameters (conditional probabilities distributions).
We first present the construction of the model (structure and parameters) from the prior knowledge and historical data and then we explain the way our model can be used with traceability data.

4.1. Structure of our model

We have proposed [38] a structural building algorithm applied to industrial processes, the Causal Bayesian Networks Building (CNBN) algorithm. The CBNB algorithm differs from conventional algorithms which seeks to identify, for a given set of parameters and a dataset, one Bayesian network among all possible networks that fit best the dataset. CBNB incorporates expert knowledge in order to limit the combinatorial explosion and to enhance the quality of the built network. The CBNB algorithm has two phases: the allocation phase and the causal relationships learning phase. Expert knowledge is first used to allocate the system's variables to the predefined levels of causality and finally the data is employed to determine causal relationships between the system variables.

In the allocation phase, the variables are assigned to different levels of causality. These levels should then be ordered from the root level (consisting of root nodes) to leaf level (composed of leaf nodes). There can be one or more intermediate levels between these two extreme levels. A variable belonging to a given level is likely to directly influence one or more variables belonging to the consecutive lower level. Direct dependency relationships are not possible between nodes of non-consecutive levels. Levels must be homogeneous in

terms of causality such that a variable cannot influence a variable with which it shares the same level of causality. We designate this particular network configuration by the cascade arrangement (c.f. Fig. 2).

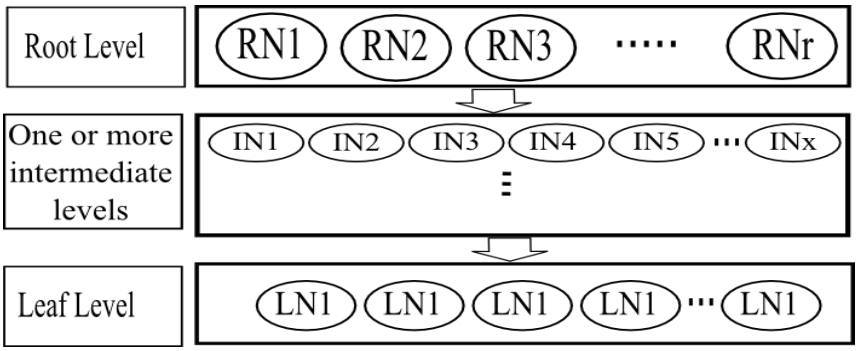

Fig. 2: The cascade arrangement

For causal Bayesian networks applied for industrial diagnostics, variables allocation into homogeneous causal levels and the prioritization of these levels can be set by an expert. We suggest in this case to structure the network in 3 levels: the root level, one intermediate level and the leaf level. Root level composed by control variables, the intermediate level constituted by process characteristics and leaf level formed by product parameters.

In the second phase of structural building algorithm, unitary traceability data is employed to determine causal relationships between the system variables. The CBNB algorithm is a polynomial time learning algorithm *O* (n2), where n << N (N = total number of variables) [38].

Must of the published algorithm allow for expert knowledge to be incorporated. However, two moments when expert knowledge may be introduced have to be distinguished, namely 1) the algorithm development phase and 2) the algorithm use phase (see Fig. 3).

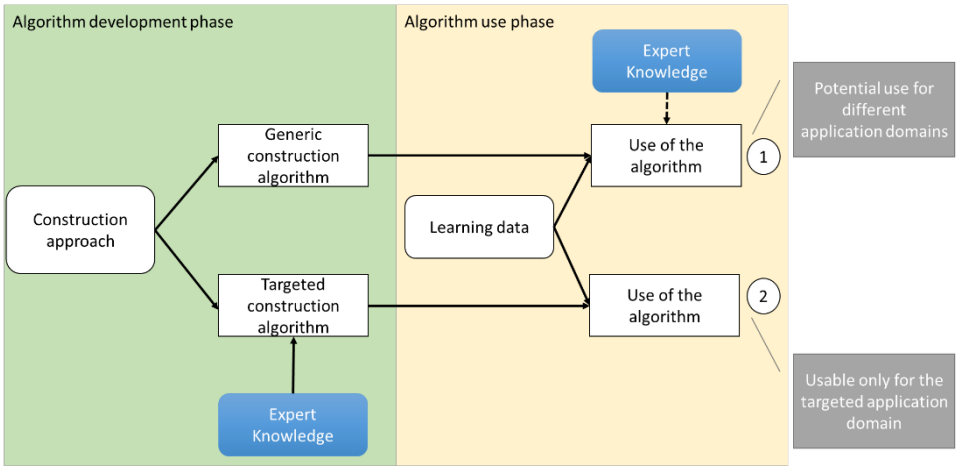

Fig. 3: Different moments when expert knowledge may be introduced: 1) the algorithm development phase and 2) the algorithm use phase

In the first case, the expert knowledge is introduced in the use phase of a generic construction algorithm, on a case-by-case basis. The algorithm is designed to be used in different application domains. For a given domain, the algorithm allows to incorporate the expert knowledge of the domain in question. The advantage of these generic algorithms is that they can address several application domains. However, this approach requires the participation of the expert in each use of the algorithm. Almost all structure learning algorithms in the literature are generic and allow for expert knowledge to be incorporated in the use phase.

For the second case, the expert knowledge is taken into account from the algorithm development phase. Expert knowledge of the concerned domain is employed in conjunction with the rules derived from one the construction approach to define the algorithm. The obtained construction algorithm is dedicated to a specific application domain and it is optimized for this domain. In theory, it cannot be used for another application domain. Our proposed structure building algorithm belong to this second case.

In order to complete the definition of the Bayesian network, we should determine the conditional (or marginal, for root nodes) probability distribution of each nodes.

4.2 Determination of our model's parameters

Let Pa be the process variables vector defined by its d features $Pa = (x_1, x_2, \dots, x_d)$. Pa is assume to be a random vector (in the sense of Bayesian theory) with x_i the i^{th} marginal.

We assume that the considered variables are mutually independent. Thus, the probability law of Pa is the product of the marginal probability of the different x_i.

Two cases can be distinguished:

- First case: the prior (unconditional) distribution of the variable is given. For example, the probability that a sensor is healthy or faulty. The given distribution is considered as the distribution of the corresponding node in the network.
- Second case: the variable is collected as value with uncertainty. This case requires further processing to determine prior distribution.

When the model's parameters have to be estimated, two approaches can be used. The first approach consists in using the frequentist and maximum likelihood estimator. The second alternative is the posterior distributions method also called the Bayesian approach. Posterior estimates are more robust than maximum likelihood estimate and fulfill the regularity conditions of model estimation and inference methods [40].

Control variables or Root Causes.

For each control variable, we divide the value space into two regions: Normal (N) and Abnormal (AN).

Let x_{up} and x_{low} the upper and lower bounds of definition range of parameter P_i.

$$P(P_i = N) = P(x_{low} \leq x \leq x_{up}) = \int_{x_{low}}^{x_{up}} f(x)\, dx$$

and

$$P(P_i = AN) = 1 - P(P_i = N)$$

Product parameters and Process characteristics, intermediate and leaf nodes

For each node belonging to the intermediate or leaf level, its conditional probability table is learned from historical data and background knowledge. These probabilities are updated based on new data or knowledge acquired on the process. These updates may also suggest a change in the structure of the Bayesian network.

4.3. Use of the causal Bayesian model for diagnosis

In the previous two subsections we have addressed the Bayesian network construction. This Bayesian network models the causal relationships of the studied system. Its main use will consist in updating the prior belief and inference calculations (computing marginal probabilities) to find reasons of detected nonconformities. When a nonconformity is detected, this causal analysis model is used to found the root causes. From an evidence of the presence of a nonconformity, we search the parameters which are likely to be its causes. We do this by computing posteriori probability Pr of parameter variables **X** conditioning

on evidence E through our model: Pr(**X** | E). This technique is known as inference or probabilistic reasoning or belief updating [40]. In general, probabilistic inference in Bayesian networks is NP-hard [39, 41]. Despite the potentially large size of this type of graph, this inference problem can be addressed by techniques that are custom tailored to particular inference queries [42]. This is what we did by proposing a diagnostic procedure using the Bayesian network. The diagram below (Fig. 4) illustrates the proposed diagnosis procedure. This procedure uses inference results from the proposed causal Bayesian model.

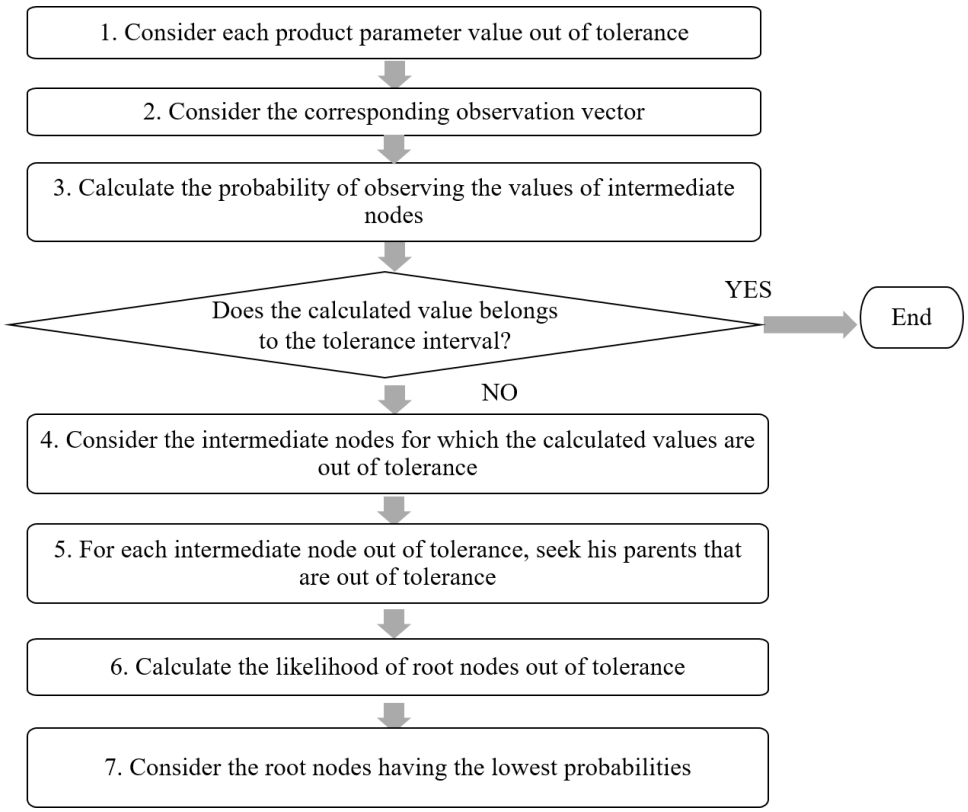

Fig. 4. Proposed diagnosis procedure using the developed Bayesian framework

(1) Consider each product parameter value out of tolerance: We assume that the product fault is manifested by an abnormal change in one or more of its parameters. An abnormal value of a product parameter will serve as symptom to detect the dysfunction of the process.

(2) Consider the corresponding observation vector: In order to analyse this value of the product parameter which is out of tolerance, we consider the other observed values of the other parameters (product and process) belonging to the same observation vector.

(3) Calculate the probability of observing the values of intermediate nodes: This is the conditional probability of observing the values of the parent nodes of the node corresponding to the product parameter. Two cases can be distinguished:

- If all obtained values belong to the tolerance interval, then the parent nodes in the network cannot account for this symptom. Two cases are possible: either other parameters not taken into account in the network are the cause of this symptom or the observed values are incorrect. The procedure with the network in this case ends here.
- One or more of the obtained values do not belong to the tolerance interval, the procedure continues in this case.

(4) Consider the intermediate nodes for which the calculated values are out of tolerance: We seek now to explain the deviation of the values of intermediate nodes.

(5) For each intermediate node out of tolerance, seek his parents that are out of tolerance: As in step 3, we look for the root nodes out of tolerance.

(6) Calculate the likelihood of root nodes out of tolerance: This is conditional probabilities of root nodes out of tolerance by conditioning on the value of the observed parameter.

(7) Consider the root nodes having the lowest probabilities: The process parameters of which the conditional probabilities are lowest are more likely to be abnormal. They should therefore be analyzed in priority to seek an explanation for the observed symptoms.

The use of the proposed diagnostic procedure is justified even when the nominal values of all variables are known and that all these variables are observable, which is rarely the case. Indeed, even in this ideal case, it is difficult to understand all existing causal relationships between variables. These relationships are often due to physical laws that are not modelled or difficult to calculate.

A concrete example of this diagnosis approach is provided in Section 7.

Once the root causes behind the nonconformity have been determined, a search in the traceability data is conducted to determine the duration of abnormal operation and then the other items likely to be noncompliant.

4.4. Determination of other noncompliant items

Once the detected default root causes are identified, a search in the traceability data can determine the duration of the abnormal operation period. The other products likely to be non-compliant are those manufactured during this period. They are identified by a forward and backward traceability process. We divide the manufacturing process into process segments. Each process segment receives inputs (raw materials, intermediate products,

energy, etc.) and other ancillary inputs (consumables, etc.). Intermediate products _IP or final products _FP (for the last segment) are produced by each segment (see Fig. 5).

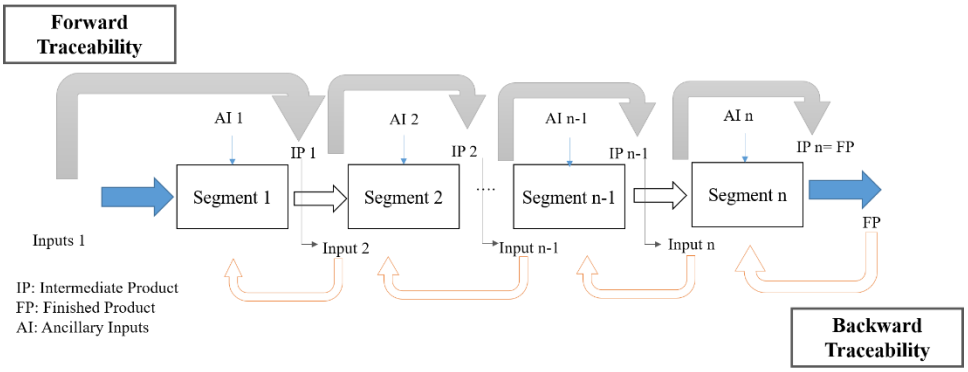

Fig. 5: Forward and backward traceability

To ensure end-to-end traceability, the intermediate product of a given segment which is used as input by the following segment has to be identified identically by the two segments. Two possibilities could be considered. The first and simplest one is to assign an identifier to the output. But in some situations this is not possible. Typically, this is when it is impossible to assign an identifier to the output or to read the assigned identifier. The second option consists in using the time parameter. For this, you need the release time of the intermediate product of the segment that produced it, the date of entry in the segment that uses it and the time taken between the two segments. If the traceability system allow to meet the conditions set out above, it will be possible for a given segment, to determine raw materials lots used, actual process parameters values and list of produced items at a given point in time.

Backward traceability: Starting from a finished product, the process parameters and inputs of the segment n are determined based on its identifier. The intermediate product used to produce this finished product is also identified. The data related to the segment n-1 to produce the intermediate product used by the segment n is considered. This procedure is repeated up to ultimate raw materials.

Forward traceability: The production chain is browsed through from raw materials to finished product to determine items containing a given material or produced in specific conditions.

Traceability in continuous process industries is more challenging than for discrete processes. In order to identify the root cause of the detected fault and accurately determine the other articles likely to have the same defect, it is necessary to uniquely identify all the finished or semi-finished products and to know the date of passage to each process segment and the value of the process and product parameters at that moment for each finished or semi-finished product. We proposed for our glassmaker partner a breakdown of its production line into 6 process segments: Batch mixing, Melting, Delivery of molten glass,

Molding, Annealing and Coating. We then identified together the process parameters to be collected for each process segment (see Fig. 6).

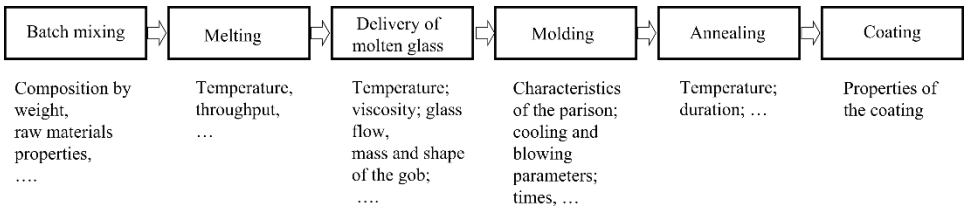

Fig. 6: Main process segments of the glass process and some examples of process parameters collected for each segment

The construction and the different usages of the diagnostic model described above require structured unitary traceability data.

5. Proposed Unitary Traceability Data Model

Two types of traceability can be distinguished: tracking or forward traceability and tracing or backward traceability. Forward traceability is used to determine, for example, finished products containing a particular ingredient or having undergone a specific process. Backward traceability offers the possibility to identify suppliers and processes involved in producing a particular article [43]. As the aggregated level of traceability is not enough to have an accurate picture of the conditions of production for job production and flow production, we consider unitary traceability. This unitary traceability enables a serialized unique identification at the item level and allows to know accurately the process parameters values of each item.

5.1. Related works

A few data models dedicated to the unitary traceability have been published. Jansen-Vullers et al. [44] and Khabbazi et al.[45] propose traceability data models with some restrictions in terms of actual material and process data registration and unitary traceability capability. Indeed, the data model proposed in [44] is more suitable for material traceability, i.e. to determine the actual composition of produced goods in batch production. The traceability model developed in [45] is a lot-based level manufacturing proprieties traceability data model. The data model proposed in [43] has interesting features for unitary traceability but should be supplemented with additional features and data to achieve the purposes of this research work.

Almost all of the analysed data models are dedicated to the management of the supply chain. The application areas targeted are often agriculture and livestock breeding sectors. For those who are dedicated to the manufacturing or the chemical industry, it is most often about material traceability where interest is focused on materials used in the process. The machine parameters or variables related to the environment are not clearly addressed.

5.2. Proposed data model for traceability

The properties that a good traceability system should have are [35]:

i) Ingredients and raw materials must be grouped into units having similar properties (notion of "Traceable Resource Units"),
ii) Unique Identifiers must be assigned to these units,
iii) Product and process properties must be recorded and either directly or indirectly linked to these identifiers and
iv) An access mechanism to these properties must be defined.

Our proposed data model strives to meet the above conditions [46].

The data needed for traceability is managed by disconnected transactional systems. Root cause search is very tedious in these conditions. Our aim is to integrate all the necessary data for causal analysis to determine root causes and to facilitate data exchange. In particular, we aim to determine precisely the related data to each manufactured item for each step of its manufacturing and distribution processes. For example, the data model should allow to answer questions like: what are the characteristics of the inputs used to produce a given item? And what were the values of the process parameters during its manufacture? Or what are the products that contain a given ingredient or produced in given conditions? In order to answer these specific questions, the data model of the warehouse has to be designed so that it enable a unitary traceability.

IEC 62264 [47] standard provides objects models and attributes of manufacturing operations. The GS1 EPC (Electronic Product Code) Global standards [48] allow end-to-end product traceability along a supply chain. The proposed item-based traceability data model (see Fig.7) integrates models of IEC 62264 and EPC Global standards.

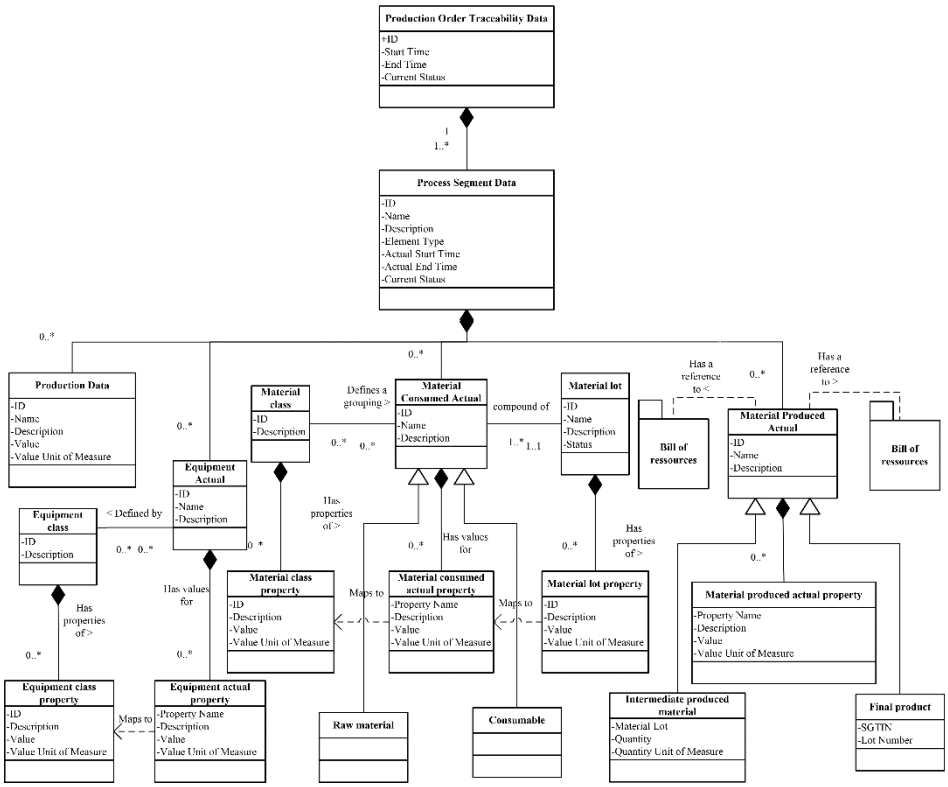

Fig. 7: Unitary traceability data model based on IEC 62264 and GS1 EPCIS standards

In our proposed data model, the traceability data are organized by production order. We opted for the functional decomposition to describe the manufacturing process. The manufacturing process is broke down into functional segments called process segments. The production order data is made up of data related to different process segments. For each process segment, production data, material consumed actual and material produced actual are recorded. This data model is to be implemented by the traceability data warehouse for the decision support system. The historized data will serve as the basis for causal analysis in particular. Each process segment realize an observable or measurable added value. Thus, normal or abnormal operation of a segment can be estimated by the level of achievement of the added value. All equipment and material parameters required for the execution of the stage of the process corresponding to the considered segment should be recorded. The characteristics of the output (final or intermediate product) or any other parameter which enables to qualify the segment should also be recorded. Parameters should be recorded at regular time intervals. The process inputs and outputs must have a serialized unique identification (at lot or item level). The proposed model allows to know for each item, the process parameters of its manufacture.

6. Our Proposed Bayesian Network Model and Big Data Challenges

The data referred to in this research work is obtained through a process and product traceability system for glass making process. As mentioned earlier, the volume of data to collect (process and product data) and their diversity make that we consider them as Big Data. This data are very varied (process parameters, life cycle events, the product features, etc.) and is collected from different systems and databases (machines parameters databases, operators recording interface, etc.). For one bottle, about 1000 process parameters can be collected in the manufacturing phase and several hundred logistic events may occur during its life cycle. According to the European Container Glass Federation (FEVE), 75.9 billion units of glass packaging were produced in Europe in 2016 by its 60 corporate members belonging to approximately 20 independent corporate groups, averaging nearly 3.8 billion units by corporate [49].

The processing of Big Data (validation, reconciliation, aggregation, disaggregation, etc.) presents some challenges. To overcome these challenges, one should address the different aspects of Big Data, the most challenging of which are the size, variety and velocity.

Our proposed model in this paper address mainly the variety and size aspects:

- Variety aspect: The set up and usage of the diagnosis model require a structured process and product data. We have proposed a unitary traceability data model in order to classified process and product historical data collected from different operational data bases. However, due to the volume and complex structure of data to be archived and data processing requirements, their handling by traditional relational database management systems is impracticable. Another approach is therefore needed. NoSQL is one of the most popular proposed approaches to handle Big Data. Considering the Big Data modeling, NoSQL proposed 4 types of databases [50] : Key-value database, Document-oriented database, Wide-column (or column-family) database and Graph database (For further details about these types of database, the reader can refer to [51]).

  In the main database, we propose to store the minimum data allowing on the one hand to know the production order of an article from its serial number and, on the other hand, to retrieve the corresponding process and product data necessary for the diagnostic procedure in case of non-compliance detection.

  The transactional data will therefore remain in the operational databases and master data of Production Orders (PO) will be stored in the main database. When a given PO is concerned by the diagnostic procedure, the corresponding transactional data will be retrieved from operational databases using the proposed data model shown in Figure 7. Key-value database was adapted to store the PO master data. The proposed model is presented in Figure 8.

Production Order (PO)	
Key	Value
PO_ID1	Start Date : End Date : Produced Articles :
PO_ID2	Start Date : End Date : Produced Articles :
…	…

Fig. 8: Key-value database for the Production Order (PO) master data

- Size aspect: Thanks to the Bayes' theorem, BNs contributes to the reduction of the combinatorial explosion in Big Data processing. Our contribution regarding this aspect concerns the construction of the Bayesian model (definition of structure and parameters) and the use of the obtained model for diagnosis. Both structure learning and inference are NP-hard problems [39, 41, 52].
    - Structure: In the definition of the BN structure, we can act on 2 levers to face the size aspect of Big Data Challenges: the structure learning algorithm by to building scalable learning algorithm and the leaning process by designing a distributed learning process. The proposed CBNB algorithm is a polynomial time learning algorithm. It complexity is *O* (n2), where n << N (N = total number of variables). By way of comparison, the complexity of the MMHC algorithm[53] which is able to scale up to thousands of variables, is *O* (N^2. PC^{l+1}), where PC is the largest set of parents and children over all variables in N and l is the maximum size of the conditioning subsets.
      When a structure of a BN is defined, it is used several times to make inference before a possible update.

    - Inference: we have proposed a custom tailored inference process to achieve the diagnostic procedure using the system Bayesian network model (see 4.3 and 4.4). The diagnostic procedure can be performed in a batch-oriented processing approach. Consider the proposed diagnosis procedure (Fig. 4.). From step 3 to step 7, all the calculation and analysis can executed in parallel by selecting the appropriate data for each node based on MapReduce [54] and splitting data in a distributed computing environments.

7. Case Study and Proof of Concept

Our contribution in this work was developing a data-based diagnosis model applicable to complex manufacturing industries with large amounts of data considered as "Big Data". In order to validate our proposed approach, 2 elements must be proven: the ability of the developed framework to make an accurate diagnosis and its capacity to deal with "Big Data".

In Section 6 (Proposed Bayesian Network Model and Big Data Challenges), we have shown how we address challenges raised by "Big data", especially variety and size challenges.

To verify the accuracy of the diagnosis realized by our model, we need a case study for which we know the real cause of each default tested. Since the real causal links between process parameters and faults are unknown for the real system we are addressing by our model, we based our decision on a well-established benchmark process, the Tennessee Eastman (TE) process for which we know the true default for each dataset. This help us to verify and validate our approach and algorithms. This benchmark is small compared to the system for which the proposed diagnosis model is intended. The choice of this benchmark is justified by the fact that it allows us to assess the capacity of our model to realize a correct diagnosis. As we know the true default for each dataset, we can thus compare the diagnosis provided by our model and the real root cause. The TE process is a real process widely accepted as benchmark for fault detection and diagnosis methods. A complete description of this process can be found in [55].

The process can be decomposed into five process segments:

1. The reactor,
2. The condenser,
3. The compressor,
4. The separator, and
5. The stripper.

The process leads to two products (G and H) and a by-product F from four reactants (A, C, D and E) and an inert, a non-reactant (B). The flow diagram of the process is depicted in Fig. 9.

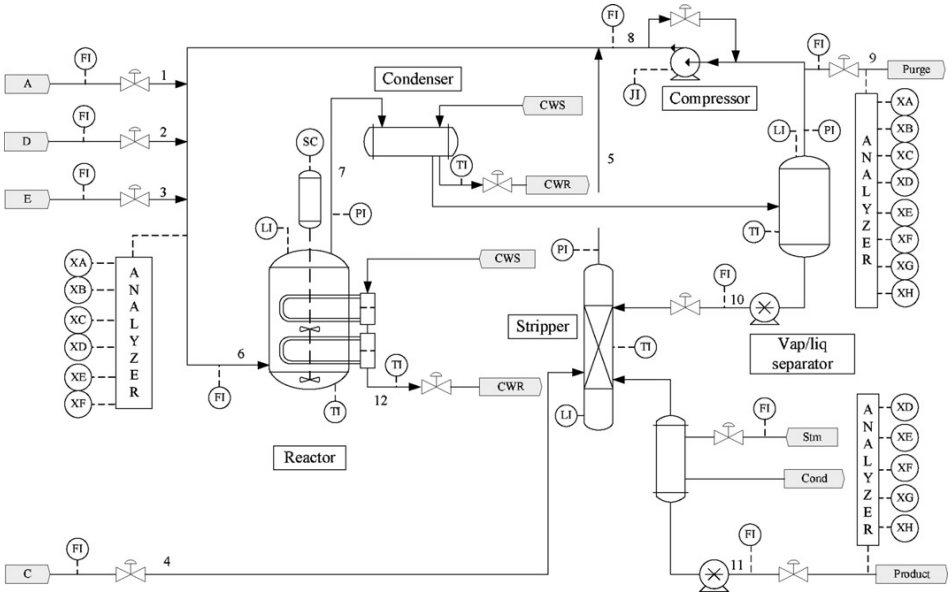

Fig. 9: The Tennessee Eastman process

41 variables was collected from the process including 22 continuous process measurements and 19 sampled process measurements (see Tables 5 and 6 of [55]). 21 process faults was defined (see Table 5 in Appendix). These 41 observed process parameters will be considered for the construction of the diagnosis model.

The dataset used in this case study is generated from a simulation described in [56]. This dataset was downloaded from [57]. It is composed of:

- 22 training data sets (1set for the normal operation and 21sets for each of the 21 process faults) collected during 24 operation hours and
- 22 test data sets corresponding to 48 operation hours. For the test datasets, faults were introduced after 8 simulation hours.

For all the collected data, the sampling time was 3 min.

The Bayesian model structure for this process was learned based on the 22 training sets. The 480 rows for the 41 considered columns from each training data file were concatenated in one training dataset. That makes 10,560 rows for 41 columns corresponding to 22 operation days. However, the parameters of the Bayesian model were learned solely from the normal operation training data set.

This case study has been carried out using the bnlearn package for R. The model building (structure and parameters) and usage (inferences) were performed with this package.

The tests were conducted on PC Intel (R) Core (TM) i5-3340 M CPU @2.70GHz 2.7 GHz, 4Go RAM running Windows 7 Professional.

7.1. The TE process diagnosis model structure

We apply our diagnosis approach on the TE process in order to detect and diagnosis any deviation from quality objectives set. This will be done by monitoring stream 11 (product). The first step of the approach is the allocation of the system's variables to the different levels of causality. This is where the expert knowledge of the system is integrated in order to reduce the complexity of the learning algorithm.

We consider observable variables whose values over time are collected (Tables 4 and 5 of [55]). The construction and use phases of the system may suggest to exclude certain variables or on the contrary to include additional variables. Indeed, CBNB algorithm requires that nodes belonging to the same level should not be dependent. When two or more variables belonging to the same level are dependent, in this case we retain only one among them in order to respect the constraint on causal homogeneity of nodes belonging to the same level. In the use phase, when we cannot find a defect root causes, one explanation could be that the model is not complete and some variables are missing. It will be necessary in this case to collect additional variables. Among collected variables in this case study:

(1) The control variables XMEAS1, XMEAS2, XMEAS3, XMEAS4, XMEAS5, XMEAS9, XMEAS14 and XMEAS17 form the root nodes (Table 2).

These variables are taken from the 12 manipulated variables (Table 3 of [55]). Among these 12 variables, we exclude:

- Unobserved variables: XMV (11) and XMV (12)
- Variables that do not have any impact on the characteristics of the finished product: XMV (6), XMV (9).

For some manipulated variables of which the manipulated variable is not observed, we considered an observed value that can be considered as a direct consequence of that manipulated variable: XMEAS(9) for XMV (10) and XMEAS(5) for XMV (5).

(2) The variables which characterize the production system form intermediate layer (Table 3). These are : XMEAS (7), XMEAS (10), XMEAS (13), XMEAS (16), XMEAS (20), XMEAS (21), XMEAS (22), XMEAS (23), XMEAS (24), XMEAS (25), XMEAS (26), XMEAS (27), XMEAS (28), XMEAS (29), XMEAS (30), XMEAS (31), XMEAS (32), XMEAS (33), XMEAS (34), XMEAS (35) and XMEAS (36).

(3) Finally, product analysis measurements (XMEAS37– XMEAS41) are considered as leaf nodes (Table 4).

Table 2. Root nodes (manipulated variables) (#8)

Process segment	Parameter name	Parameter code
input (material)	A feed (stream 1)	XMEAS(1)
	D feed (stream 2)	XMEAS(2)
	E feed (stream 3)	XMEAS(3)
	A and C feed	XMEAS(4)
Compressor	Recycle flow (stream 8)	XMEAS(5)
Reactor	Reactor temperature	XMEAS(9)
Separator	Separator underflow	XMEAS(14)
Stripper	Stripper underflow	XMEAS(17)

Table 3. Intermediate nodes (system features) (#22)

Process segment	Parameter name	Parameter code
Reactor	Reactor pressure	XMEAS(7)
	Reactor cooling water outlet temperature	XMEAS(21)
Compressor	Purge rate (stream 9)	XMEAS(10)
	Compressor work	XMEAS(20)
Separator	Product separator level	XMEAS(12)
	Product separator pressure	XMEAS(13)
	Separator cooling water outlet temperature	XMEAS(22)

Stripper	Stripper pressure	XMEAS(16)
Reactor feed	Component A	XMEAS(23)
	Component B	XMEAS(24)
	Component C	XMEAS(25)
	Component D	XMEAS(26)
	Component E	XMEAS(27)
	Component F	XMEAS(28)
Purge gas	Component A	XMEAS(29)
	Component B	XMEAS(30)
	Component C	XMEAS(31)
	Component D	XMEAS(32)
	Component E	XMEAS(33)
	Component F	XMEAS(34)
	Component G	XMEAS(35)
	Component H	XMEAS(36)

Table 4: Leaf nodes (finished product characteristics) (#5)

Parameter name	Parameter code
Component D	XMEAS(37)
Component E	XMEAS(38)
Component F	XMEAS(39)

Component G	XMEAS(40)
Component H	XMEAS(41)

A total of 35 variables was used to build the diagnosis Bayesian model. These 35 variables are divided into 3 groups according to the data model proposed in Section 4: 26 machine/process parameters (XMEAS(5), XMEAS(9), XMEAS(14), XMEAS(17), XMEAS(7), XMEAS(21), XMEAS(10), XMEAS(20), XMEAS(12), XMEAS(13), XMEAS(22), XMEAS(16), XMEAS(23), XMEAS(24), XMEAS(25), XMEAS(26), XMEAS(27), XMEAS(28), XMEAS(29), XMEAS(30), XMEAS(31), XMEAS(32), XMEAS(33), XMEAS(34), XMEAS(35), XMEAS(36)), 4 consumed material parameters (XMEAS(1), XMEAS(2), XMEAS(3), XMEAS(4)) and 5 produced material parameters (XMEAS(37), XMEAS(38), XMEAS(39), XMEAS(40) and XMEAS(41)).

The 1st step of the CBNB algorithm (the allocation phase) yield the following variable repartition (Fig.10)

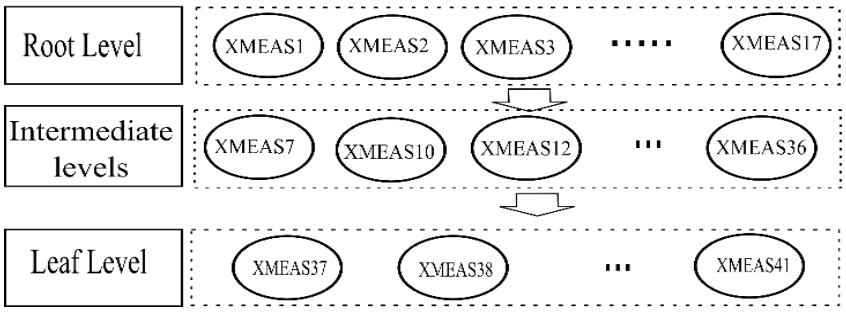

Fig. 10: Allocation of TE process variables to 3 levels of causality

The causals relationships (network's edges) are leaned from the dataset composed by all the 22 provided datasets using the second phase of the CBNB algorithm. The obtained network is depicted in Fig. 11.

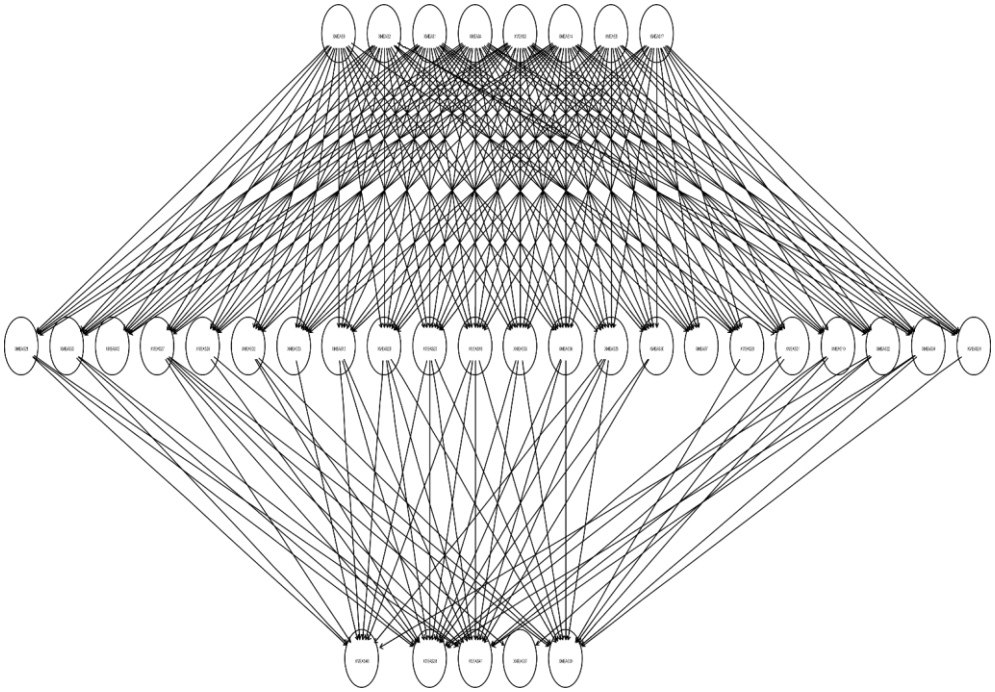

Fig. 11: Structure of the causal model for the TE process

Some lessons can be drawn from this graph. For example, process parameter XMEAS12 (Separator level) and XMEAS7 (Reactor pressure) do not have any impact on products characteristics. Product parameter XMEAS37 (Component D) is impacted solely by the process segments outputs characteristics XMEAS27 (Component E) and XMEAS22 (Separator cooling water outlet temperature). XMEAS24 (Component B) impact only parameter XMEAS39 (Component F) .We can also verify expert knowledge. For example, in the TEP process, it is known that component G is highly sensitive to temperature because the reaction to produce G has a higher activation energy [55]. To check that, we verify whether there is a path between process parameters XMEAS9 (Reactor temperature) and product parameter XMEAS40 (Component G). The following are the finding obtained.

```
> path (TEP_Model_Structure, from="XMEAS9", to="XMEAS40")

[1] TRUE
```

In order to represent quantitatively the behaviour of the variables and their relative dependencies, we need to specify their parameters (probability distributions).

7.2. The TE process diagnosis model parameters

The network parameters were learned from the normal operation training dataset. The bnlearn function *bn.fit* was used to fit the parameters of the network. Two alternatives are possible: the maximum likelihood estimates and Bayesian Posterior estimates. At the time of this study, only the maximum likelihood estimator were implemented by bn.fit function for continuous data. Thus, the network parameters were learned using the maximum likelihood estimates.

Below we present the repartition of three variables as an illustration (Fig.12).

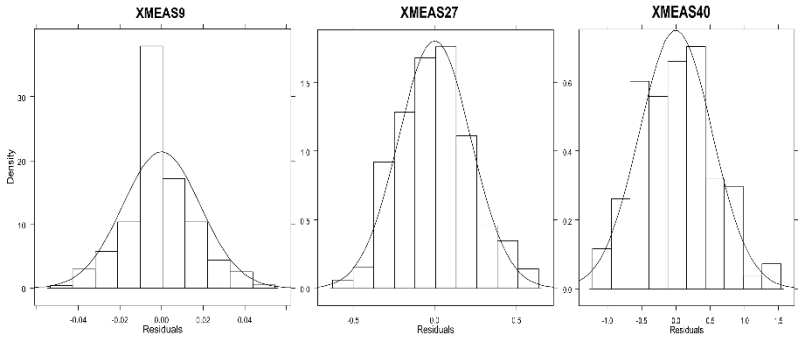

Fig. 12: The repartition graph of XMEAS9 (root node), XMEAS27 (intermediate node) and XMEAS40 (leaf node)

The Gaussian distribution parameters (mean and standard deviation) of all the model variables can be found in Appendix (Table 6 -8).

As we can notice in the Table 6 (Appendix), the mean value of root nodes, i.e. nodes without any parent are given by a real value and intermediate and leaf nodes mean values are expressed as a function of parent nodes. For XMEAS37 for example, the mean value is: 4.545812e-02 - 4.089572e-05 * XMEAS22 + -1.338949e-03 * XMEAS27. By replacing the parent variables by their base case value, the obtained result is close to the base case value of variable XMEAS37. In general, the mean value of the leaned distribution are practically all equal to the base case value of the variable.

7.3. The model usage (inference)

Once the definition of the Bayesian network is completed (structure and parameters), several uses are conceivable. It can be used both for detection and diagnosis purposes. Below are some of potential uses for the obtained Bayesian model.

For each observation corresponding to one vector in the dataset, we can calculate the density value for the observed value for each variable of interest. The variable whose density is very low compared to the base case density are considered doubtful. Thresholds should be defined. One or many observed values can be tested to see whether they are plausible and consistent with normal operation. The use in diagnosis mode would consist,

for example, in determining abnormal process parameter (root causes) when we have an evidence about the nonconformity of a product parameter. This latter case, which is the focus of this work, is illustrated below by an example.

Let consider dataset "d06_te" corresponding to testing dataset for process fault IDV (6), A feed loss. The product parameters during this experience are depicted in Fig.13.

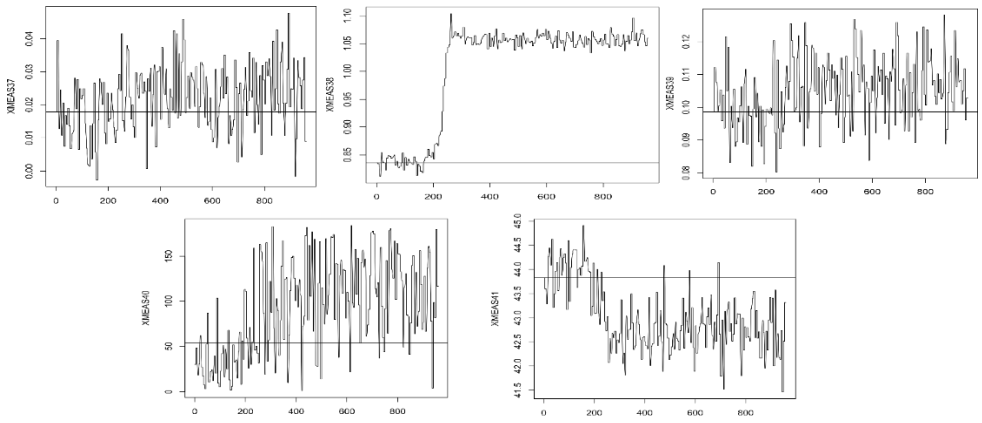

Fig. 13: Product parameters monitoring with introduction of fault IDV (6)

It can be observed that parameters XMEAS38, XMEAS40 and XMEAS41 diverge from base case value after a certain period of time. Parameter XMEAS38 reaches a maximum value of 1.1039 for observation vectors from 261 to 265. Likewise, XMEAS40 reaches a maximum value of 55.754 for observation vectors from 616 to 620. Finally, XMEAS41 move downwards to 41.462 for observation vectors from 946 to 950.

Let look in particular at one observation vector of this abnormal operation periods, observation vector n°616 for example. For this vector, the observed value is XMEAS40 = 55.754 mol%. However, the base case value is 53.724 mol%. The calculated value from the model is 52.519. This value is within the tolerance interval of XMEAS40 = 53.724 $\pm$ 0.531 mol%. It can be conclude that the parents of XMEAS40 in the model are not responsible for this nonconformity.

Let consider now observation vector n°948, another abnormal vector. XMEAS41 = 41.462 mol% for this observation (base case value is 43.828). The calculated value for this parameter from the model is 43.124. This value does not belong to the tolerance interval of XMEAS41 = 43.828 ± 0.485mol%. One or more of XMEAS41 parents could therefore be abnormal. The parents of XMEAS41 in the graph are: XMEAS21, XMEAS10, XMEAS20, XMEAS13, XMEAS22, XMEAS16, XMEAS25, XMEAS27, XMEAS29, XMEAS30, XMEAS32, XMEAS33, XMEAS34, XMEAS35 and XMEAS36.

We calculate their values using the model with observed value. By reasoning as above, it can found that XMEAS21, XMEAS10, XMEAS16, XMEAS25 and XMEAS29 are outside of tolerance.

The root nodes impacting these intermediate nodes are XMEAS1, XMEAS2, XMEAS3, XMEAS4, XMEAS5, XMEAS9, XMEAS14 and XMEAS17

Among these variables, the following are out of tolerance: XMEAS1, XMEAS2, XMEAS3, XMEAS4, XMEAS9 and XMEAS17.

In order to enhance the diagnosis, let calculate the conditional probability of these variable. In R, this is done by employing the cpquery function.

- cpquery (TEP_Model_Network, ((0-0.02855132<=**XMEAS1**)&(**XMEAS1**<=0+0.02855132)), evidence =list(XMEAS41 = 41.462), method = "lw") = **0**
- cpquery(TEP_Model_Network, ((3595.5-32.03159<=**XMEAS2**)&(**XMEAS2**<=3595.5+32.03159)), evidence =list(XMEAS41 = 41.462), method = "lw") = **0.1612637**
- cpquery(TEP_Model_Network, ((4172.5-31.72456<=**XMEAS3**)&(**XMEAS3**<=4172.5+31.72456)), evidence =list(XMEAS41 = 41.462), method = "lw")=**0**
- cpquery(TEP_Model_Network, ((9.7249-0.07651388<=**XMEAS4**)&(**XMEAS4**<=9.7249+0.07651388)), evidence =list(XMEAS41 = 41.462), method = "lw")= **5.532716e-05**
- cpquery(TEP_Model_Network, ((120.45-0.01865429<=**XMEAS9**)&(**XMEAS9**<=120.45+0.01865429)), evidence =list(XMEAS41 = 41.462), method = "lw")=**0.04486187**
- cpquery(TEP_Model_Network, ((22.209-0.6268055<=**XMEAS17**)&(**XMEAS17**<=22.209+0.6268055)), evidence =list(XMEAS41 = 41.462), method = "lw")=**0.4286973**

The parameters for whose conditional probability are not null may be disregarded. The remaining parameters will be considered for further investigation by expert. In this case, the parameters to analyze for root causes search will be XMEAS1 and XMEAS3.

According to these results, the root causes of the detected defect are parameters XMEAS1 and XMEAS3. This diagnosis should be confirmed by an expert who will analyse these two parameters. The proposed diagnostic procedure allows to identify the most likely parameters to cause the default. It directs the expert to the most likely causes and facilitates his work.

As we know the process default in this case, we can confirm the presence of A feed loss (XMEAS1) and its consequences on the other parameters. The presence of XMEAS3 among potential cause of the deviation of the characteristic is due to the fact that the feed E loss (XMEAS3) also causes the deviation of parameter H (XMEAS41). Indeed, the realisation of H involves A and E (see equation of product 2 from [55]).

Data-based methods are effective in fault detection even if they produce relatively more false alarms compared to the other methods. Yin et al [58] compared basic data-driven methods for fault diagnosis and process monitoring including different variants of PCA (Principal Component Analysis) and PLS (Partial Least Squares) applied to the Tennessee Eastman benchmark. Authors considered 2 comparison criteria: fault detection rate (FDR) and false alarm rate (FAR). All tested methods in this study obtained a FDR of 100% for IDV (6). Obtained FAR range from 1.5% to nearly 20%.

In addition to the high rate of false alarms, the other weak point of these basic data-based methods is that they do not allow to diagnose detected faults especially in case of multiple defaults [59].

8. Conclusion

In this paper, we address the challenging problem of diagnosis in complex industrial systems. A data-driven approach using industrial big data is proposed. In order to perform accurate diagnoses for job production and flow production, unitary traceability data is considered. First, a diagnosis framework based on Bayesian theory is developed using data and expert knowledge. The Bayesian Network formalism explicitly incorporates uncertainties and allow to exploit both data and expert knowledge. The semantic of the Bayesian Network makes it possible to understand the causal mechanism linking a symptom to its root cause. The definition of both structure and parameters of the Bayesian network is described. Then, a data model for the traceability data warehouse allowing forward and backward traceability is proposed. Finally, we validate our approach on an industrial benchmark, the Tennessee Eastman (TE) process. We have demonstrated through this study the ability of our proposed approach to identify precisely the root causes of a detected product nonconformity. The proposed approach is able to diagnosis multiple faults. The causal Bayesian model learns the causal relationships between symptoms and potential causes from the historical data of the system. In some cases, a symptom may be due to several possible causes (simultaneous or multiple faults). In this case, depending on the configuration (values taken by the different parameters), a given cause may be more likely than another for the symptom. For a given symptom, our diagnosis model is capable of providing all the potential causes and corresponding probabilities. These probabilities indicate the most probable causes for the observed symptom. This allows to prioritize the potential causes and check first the most likely causes.

Acknowledgments.

This work was supported by Bpifrance (French organization for innovation support and funding, Ministry for Economy, Finance and Industry, and Ministry for higher education and research) through the Traçaverre Project.

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions which helped us to improve the manuscript.

Appendix

Table 5: Process faults

Fault number	Process variable	Type
IDV(1)	A/C feed ratio, B composition constant (stream 4)	Step
IDV(2)	B composition, A/C ratio constant (stream 4)	Step
IDV(3)	D feed temperature (stream 2)	Step
IDV(4)	Reactor cooling water inlet temperature	Step
IDV(5)	Condenser cooling water inlet temperature	Step
IDV(6)	A feed loss (stream 1)	Step
IDV(7)	C header pressure loss-reduced availability (stream 4)	Step
IDV(8)	A, B, and C feed composition (stream 4)	Random variation

IDV(9)	D feed temperature (stream 2)	Random variation
IDV(10)	C feed temperature (stream 4)	Random variation
IDV(11)	Reactor cooling water inlet temperature	Random variation
IDV(12)	Condenser cooling water inlet temperature	Random variation
IDV(13)	Reaction kinetics	Slow drift
IDV(14)	Reactor cooling water valve	Sticking
IDV(15)	Condenser cooling water valve	Sticking
IDV(16)	Unknown	Unknown
IDV(17)	Unknown	Unknown
IDV(18)	Unknown	Unknown
IDV(19)	Unknown	Unknown
IDV(20)	Unknown	Unknown
IDV(21)	The valve fixed at steady state position	Constant position

Mean and standard deviation of the Gaussian distribution of the TEP diagnosis model leaned from normal operation dataset.

Table 6. Gaussian distribution Parameters of root nodes (process parameters)

Parameter code	Means	Standard deviation
XMEAS(1)	0.2511377	0.02855132

XMEAS(2)	3663.538	32.03159
XMEAS(3)	4511.517	31.72456
XMEAS(4)	9.344306	0.07651388
XMEAS(5)	26.90779	0.2086639
XMEAS(9)	120.3994	0.01865429
XMEAS(14)	25.11987	1.063972
XMEAS(17)	22.90933	0.6268055

Table 7. Gaussian distribution Parameters of intermediate nodes (process default)

Parameter code	Parameters		
	Means		Standard deviation
	Parents nodes	Regression coefficients	
XMEAS(7)	(Intercept)	2425.08854340	5.112456
	XMEAS1	9.23829205	
	XMEAS2	0.00922803	
	XMEAS3	0.03049424	
	XMEAS4	11.59211820	
	XMEAS5	0.29106837	
	XMEAS14	-0.50676467	

	XMEAS17	0.13897440	
XMEAS(21)	(Intercept)	2.465749e+02	0.0995569
	XMEAS1	-4.894351e-01	
	XMEAS2	-1.712182e-03	
	XMEAS3	-3.321795e-04	
	XMEAS4	-5.455908e-02	
	XMEAS9	-1.192435e+00	
XMEAS(10)	(Intercept)	3.492251e-01	0.01166236
	XMEAS1	-3.682417e-02	
	XMEAS3	-9.837729e-06	
	XMEAS4	6.247700e-03	
	XMEAS14	1.713339e-04	
	XMEAS17	-9.029568e-04	
XMEAS(20)	(Intercept)	408.284757608	1.196784
	XMEAS1	-5.453612067	
	XMEAS2	0.002043821	
	XMEAS3	0.004268956	
	XMEAS4	2.461676907	
	XMEAS5	-0.136894667	
	XMEAS9	-0.928456203	
	XMEAS14	-0.051436622	

	XMEAS17	0.068347922	
XMEAS(12)	(Intercept)	32.039989508	1.032149
	XMEAS1	-1.019372249	
	XMEAS2	-0.001145328	
	XMEAS3	-0.000428836	
	XMEAS4	-0.723191752	
	XMEAS9	0.236495667	
	XMEAS14	0.107251871	
XMEAS(13)	(Intercept)	-396.69580410	5.453112
	XMEAS1	8.26866828	
	XMEAS2	0.00338284	
	XMEAS4	10.80490126	
	XMEAS5	0.23506748	
	XMEAS9	24.24967012	
	XMEAS14	-0.52899604	
	XMEAS17	0.12013555	
XMEAS(22)	(Intercept)	76.3375241805	0.2443161
	XMEAS1	0.4067480710	
	XMEAS3	-0.0006672621	
	XMEAS4	0.1993590984	
	XMEAS5	0.0747790817	

	XMEAS14	-0.0346136959	
	XMEAS17	0.0371872154	
XMEAS(16)	(Intercept)	-121.63627645	4.391278
	XMEAS1	19.58468133	
	XMEAS2	0.01098730	
	XMEAS3	0.02871683	
	XMEAS4	9.75307944	
	XMEAS5	0.75759240	
	XMEAS9	24.44428174	
	XMEAS14	-0.34839997	
	XMEAS17	0.15553447	
XMEAS(23)	(Intercept)	-1.874980e+01	0.281062
	XMEAS1	-7.633864e-01	
	XMEAS2	5.115559e-05	
	XMEAS3	-3.097666e-05	
	XMEAS4	1.546580e-01	
	XMEAS5	-8.685389e-02	
	XMEAS9	4.322661e-01	
	XMEAS14	1.019263e-02	
	XMEAS17	-1.450913e-02	
XMEAS(24)	(Intercept)	1.027756e+01	0.09821371

	XMEAS2	-2.091280e-04	
	XMEAS3	1.813086e-06	
	XMEAS4	-8.259545e-02	
	XMEAS5	-1.641097e-02	
	XMEAS14	7.175020e-03	
	XMEAS17	1.771350e-02	
XMEAS(25)	(Intercept)	1.408961e+02	0.26823
	XMEAS1	1.033585e+00	
	XMEAS3	5.242528e-04	
	XMEAS4	4.835460e-01	
	XMEAS5	7.128087e-02	
	XMEAS9	-1.016782e+00	
	XMEAS14	-1.953087e-02	
	XMEAS17	-2.896350e-02	
XMEAS(26)	(Intercept)	5.469871e+01	0.1022532
	XMEAS1	1.631961e-01	
	XMEAS2	2.788842e-04	
	XMEAS3	8.066697e-05	
	XMEAS4	-2.671454e-02	
	XMEAS9	-4.055407e-01	
	XMEAS14	-6.620819e-03	

XMEAS(27)	(Intercept)	-3.151767e+01	0.2382544
	XMEAS1	1.254106e+00	
	XMEAS2	-1.003885e-03	
	XMEAS3	-3.833192e-04	
	XMEAS4	1.260000e-01	
	XMEAS9	4.474241e-01	
	XMEAS14	1.345904e-02	
XMEAS(28)	(Intercept)	1.672441e+00	0.02526544
	XMEAS1	-1.365968e-02	
	XMEAS2	-1.805459e-05	
	XMEAS3	-3.987391e-05	
	XMEAS4	1.393013e-03	
	XMEAS5	6.135097e-03	
	XMEAS14	-2.088026e-04	
	XMEAS17	2.523619e-03	
XMEAS(29)	(Intercept)	25.2067847543	0.3020689
	XMEAS1	-2.4768731710	
	XMEAS2	0.0002863278	
	XMEAS3	0.0008430303	
	XMEAS4	0.5780019817	
	XMEAS5	-0.0095601034	

	XMEAS9	-0.0050208082	
	XMEAS14	0.0071952330	
	XMEAS17	-0.0504424068	
XMEAS(30)	(Intercept)	18.1941867811	0.100662
	XMEAS1	0.0474859319	
	XMEAS2	-0.0001916875	
	XMEAS3	-0.0004555704	
	XMEAS4	-0.1651388455	
	XMEAS5	-0.0074139390	
	XMEAS14	0.0032661657	
	XMEAS17	0.0015235031	
XMEAS(31)	(Intercept)	27.732340234	0.309863
	XMEAS1	-1.492040831	
	XMEAS3	0.001298182	
	XMEAS5	0.093797982	
	XMEAS9	-0.099409282	
	XMEAS14	-0.004404744	
	XMEAS17	0.015745386	
XMEAS(32)	(Intercept)	1.757842e+01	0.1043275
	XMEAS1	2.090371e-01	
	XMEAS2	9.994101e-07	

	XMEAS3	-2.726365e-05	
	XMEAS4	7.459986e-02	
	XMEAS9	-1.409629e-01	
	XMEAS14	5.894693e-04	
XMEAS(33)	(Intercept)	-0.5025909673	0.3011958
	XMEAS1	0.8810318308	
	XMEAS2	-0.0006493152	
	XMEAS3	-0.0009366954	
	XMEAS4	-0.2205606490	
	XMEAS9	0.2284552384	
	XMEAS14	0.0005434137	
XMEAS(34)	(Intercept)	2.407241e+00	0.02631956
	XMEAS2	-3.103135e-05	
	XMEAS3	1.358183e-05	
	XMEAS4	2.738944e-04	
	XMEAS5	-2.789282e-03	
	XMEAS14	1.102924e-03	
	XMEAS17	-2.225189e-03	
XMEAS(35)	(Intercept)	1.675001e+01	0.05808402
	XMEAS1	2.146840e-02	
	XMEAS2	8.454982e-06	

	XMEAS3	-1.391250e-04	
	XMEAS4	-1.519118e-02	
	XMEAS5	-1.725029e-03	
	XMEAS9	-9.185337e-02	
	XMEAS14	-1.992573e-03	
	XMEAS17	-8.737113e-04	
XMEAS(36)	(Intercept)	6.266419e+00	0.05240731
	XMEAS1	-1.625933e-01	
	XMEAS2	-5.691578e-05	
	XMEAS3	-4.753973e-05	
	XMEAS4	-9.712575e-03	
	XMEAS5	1.604310e-02	
	XMEAS9	-3.267989e-02	
	XMEAS14	-2.242247e-04	
	XMEAS17	4.062584e-03	

Table 8. Gaussian distribution Parameters of leaf nodes (product default)

Parameter code	Parameters		
	Means		Standard deviation
	Parents nodes	Regression coefficients	
XMEAS(37)	(Intercept)	4.545812e-02	0.009044974

	XMEAS22	-4.089572e-05	
	XMEAS27	-1.338949e-03	
XMEAS(38)	(Intercept)	2.1852878471	0.01331829
	XMEAS21	0.0054120463	
	XMEAS10	-0.1577357567	
	XMEAS20	-0.0050086541	
	XMEAS13	0.0009002320	
	XMEAS16	-0.0007448488	
	XMEAS25	-0.0026016985	
	XMEAS26	-0.0059664628	
	XMEAS27	0.0011840856	
	XMEAS29	0.0017007004	
	XMEAS30	-0.0047718423	
	XMEAS31	-0.0015974713	
	XMEAS32	-0.0013161194	
	XMEAS33	0.0057378637	
	XMEAS34	-0.0193990619	
	XMEAS35	-0.0192639590	
	XMEAS36	0.0043213624	
XMEAS(39)	(Intercept)	-0.7718448458	0.009349292
	XMEAS10	0.0467714568	

	XMEAS20	0.0016652672	
	XMEAS16	0.0001156071	
	XMEAS24	0.0025222979	
	XMEAS25	-0.0044161003	
	XMEAS26	-0.0043423563	
	XMEAS27	0.0026853836	
	XMEAS28	0.0341904357	
	XMEAS29	-0.0022045540	
	XMEAS31	-0.0006290372	
	XMEAS32	-0.0072618232	
	XMEAS34	0.0153710122	
	XMEAS35	-0.0015161120	
	XMEAS36	0.0063591951	
XMEAS(40)	(Intercept)	-20.325300092	0.5309854
	XMEAS21	0.186668486	
	XMEAS20	0.099945608	
	XMEAS13	-0.024192384	
	XMEAS22	0.179898668	
	XMEAS16	0.024098572	
	XMEAS23	-0.224784168	
	XMEAS25	-0.025107247	

	XMEAS27	0.240121857	
	XMEAS29	0.014352776	
	XMEAS33	0.008439173	
XMEAS(41)	(Intercept)	55.6219324151	0.4855669
	XMEAS21	-0.0656111342	
	XMEAS10	-1.4304431566	
	XMEAS20	-0.0039228057	
	XMEAS13	-0.0023454676	
	XMEAS22	0.0945523141	
	XMEAS16	-0.0006671525	
	XMEAS25	0.0873325776	
	XMEAS27	-0.3368783795	
	XMEAS29	0.0987340109	
	XMEAS30	0.0584217855	
	XMEAS32	-0.2912412361	
	XMEAS33	0.0408422462	
	XMEAS34	-0.1391569548	
	XMEAS35	-0.4022155282	
	XMEAS36	-0.4093749056	

References

1. Christoph Gröger FN, and Bernhard Mitschang. Data Mining-driven driven Manufacturing Process Optimization. World Congress on Engineering 2012; 2012 July 4 - 6; London.
2. Gröger C, Schlaudraff J, Niedermann F, Mitschang B. Warehousing Manufacturing Data. In: Cuzzocrea A, Dayal U, eds. Data Warehousing and Knowledge Discovery: Springer Berlin Heidelberg, 2012: 142-155.
3. Kumar S. A knowledge based reliability engineering approach to manage product safety and recalls. Expert Systems with Applications 2014;41:5323-5339.
4. Lei Y, Jia F, Lin J, Xing S, Ding SX. An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data. IEEE Transactions on Industrial Electronics 2016;63:3137-3147.
5. PENG Y, KOU G, SHI Y, CHEN Z. A DESCRIPTIVE FRAMEWORK FOR THE FIELD OF DATA MINING AND KNOWLEDGE DISCOVERY. International Journal of Information Technology & Decision Making 2008;07:639-682.
6. Berman JJ. Introduction. In: Berman JJ, ed. Principles of Big Data. Boston: Morgan Kaufmann, 2013: xix-xxvi.
7. Li J, Tao F, Cheng Y, Zhao L. Big data in product lifecycle management. The International Journal of Advanced Manufacturing Technology 2015;81:667-684.
8. Russom P. Big data analytics. TDWI best practices report, fourth quarter 2011:1-35.
9. Lee I. Big data: Dimensions, evolution, impacts, and challenges. Business Horizons 2017;60:293-303.
10. Akhgar B, Saathoff GB, Arabnia HR, Hill R, Staniforth A, Bayerl PS. Application of big data for national security: a practitioner's guide to emerging technologies. UK, USA: Butterworth-Heinemann, 2015.
11. Institute MG, Manyika J, Chui M, et al. Big Data: The Next Frontier for Innovation, Competition, and Productivity: McKinsey Global Institute, 2011.
12. Company GE. The Case for an Industrial Big Data Platform: Laying the Groundwork for the New Industrial Age2013 2013.
13. Dabbas RM, Chen H-N. Mining semiconductor manufacturing data for productivity improvement — an integrated relational database approach. Computers in Industry 2001;45:29-44.
14. Kheder AB, Henry S, Bouras A. Quality improvement of product data exchanged between engineering and production through the integration of dedicated information systems. ASME 2012 11th Biennial Conference on Engineering Systems Design and Analysis; 2012; Nantes, France: American Society of Mechanical Engineers: 483-489.
15. Ramirez JC, Piqueras AS. Learning Bayesian networks for systems diagnosis. Los Alamitos: Ieee Computer Soc, 2006.
16. Weidl G, Madsen AL, Israelson S. Applications of object-oriented Bayesian networks for condition monitoring, root cause analysis and decision support on operation of complex continuous processes. Computers & Chemical Engineering 2005;29:1996-2009.
17. Dey S, Stori JA. A Bayesian network approach to root cause diagnosis of process variations. International Journal of Machine Tools and Manufacture 2005;45:75-91.

18. Przytula KW, Thompson D. Construction of Bayesian networks for diagnostics. Aerospace Conference Proceedings, 2000 IEEE; 2000 2000; Big Sky, MT, USA: 193-200 vol.195.
19. Riascos LA, Cozman FG, Miyagi PE, Simoes MG. Bayesian network supervision on fault tolerant fuel cells. Industry Applications Conference, 2006 41st IAS Annual Meeting Conference Record of the 2006 IEEE; 2006; Tampa, FL, USA: IEEE: 1059-1066.
20. Chen B, Tavner PJ, Feng Y, Song WW, Qiu YN. Bayesian network for wind turbine fault diagnosis. EWEA 2012. Copenhagen, Denmark: European Wind Energy Association, 2012.
21. Venkatasubramanian V, Rengaswamy R, Yin K, Kavuri SN. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. Computers & Chemical Engineering 2003;27:293-311.
22. KOU G, LU Y, PENG Y, SHI Y. EVALUATION OF CLASSIFICATION ALGORITHMS USING MCDM AND RANK CORRELATION. International Journal of Information Technology & Decision Making 2012;11:197-225.
23. Peng Y, Kou G, Wang G, Shi Y. FAMCDM: A fusion approach of MCDM methods to rank multiclass classification algorithms. Omega 2011;39:677-689.
24. Wang S, Cui J. Sensor-fault detection, diagnosis and estimation for centrifugal chiller systems using principal-component analysis method. Applied Energy 2005;82:197-213.
25. Rato TJ, Reis MS. Fault detection in the Tennessee Eastman benchmark process using dynamic principal components analysis based on decorrelated residuals (DPCA-DR). Chemometrics and Intelligent Laboratory Systems 2013;125:101-108.
26. He QP, Qin SJ, Wang J. A new fault diagnosis method using fault directions in Fisher discriminant analysis. AIChE Journal 2005;51:555-571.
27. Zhao Z, Zhang J, Sun Y, Tian H. Fault detection and diagnosis method for batch process based on ELM-based fault feature phase identification. Neural Computing and Applications 2016;27:167-173.
28. Kou G, Peng Y, Wang G. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Inf Sci 2014;275:1-12.
29. PENG Y, KOU G, WANG G, WU W, SHI Y. ENSEMBLE OF SOFTWARE DEFECT PREDICTORS: AN AHP-BASED EVALUATION METHOD. International Journal of Information Technology & Decision Making 2011;10:187-206.
30. Lee W-Y, House JM, Kyong N-H. Subsystem level fault diagnosis of a building's air-handling unit using general regression neural networks. Applied Energy 2004;77:153-170.
31. Cai B, Liu Y, Fan Q, et al. Multi-source information fusion based fault diagnosis of ground-source heat pump using Bayesian network. Applied Energy 2014;114:1-9.
32. Cai B, Zhao Y, Liu H, Xie M. A Data-Driven Fault Diagnosis Methodology in Three-Phase Inverters for PMSM Drive Systems. IEEE Transactions on Power Electronics 2017;32:5590-5600.
33. GS1. GS1 Global Traceability Standard. 2012.
34. ASQ Chemical and Process Industries Division CIC. ISO 9001 : 2000 Guidelines for the Chemical and Process Industries. Milwaukee, Wis.: ASQ Quality Press, 2002.

35. Olsen P, Borit M. How to define traceability. Trends in Food Science & Technology 2013;29:142-150.
36. Buzon L, Bouras A, Ouzrout Y. Knowledge exchange in a supply chain context. Virtual Enterprises and Collaborative Networks: Springer US, 2004: 145-152.
37. Riascos LA, Cozman FG, Miyagi PE. Detection and treatment of faults in automated machines based on Petri nets and Bayesian networks. Industrial Electronics, 2003 ISIE'03 2003 IEEE International Symposium on; 2003; Rio de Janeiro, Brazil, Brazil: IEEE: 729-734.
38. Diallo TML, Henry S, Ouzrout Y. Bayesian Network Building for Diagnosis in Industrial Domain Based on Expert Knowledge and Unitary Traceability Data. IFAC Symposium on Information Control in Manufacturing. Ottawa, Canada2015.
39. Korb KB, Nicholson AE. Bayesian Artificial Intelligence: Taylor & Francis, 2003.
40. Scutari M, Denis JB. Bayesian Networks: With Examples in R: Taylor & Francis, 2014.
41. Pearl J. Causality: Models, Reasoning, and Inference: Cambridge University Press, 2000.
42. Heckerman D. Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery 1997;1:79-119.
43. Wynn MT, Ouyang C, ter Hofstede AHM, Fidge CJ. Data and process requirements for product recall coordination. Computers in Industry 2011;62:776-786.
44. Jansen-Vullers MH, van Dorp CA, Beulens AJM. Managing traceability information in manufacture. International Journal of Information Management 2003;23:395-413.
45. Khabbazi MR, Ismail N, Ismail MY, Mousavi SA. Data Modeling of Traceability Information for Manufacturing Control System. Information Management and Engineering, 2009 ICIME '09 International Conference on; 2009 3-5 April 2009; TBD Xi'an, China: 633-637.
46. Diallo TML, Henry S, Ouzrout Y. Using Unitary Traceability for an Optimal Product Recall. In: Grabot B, Vallespir B, Gomes S, Bouras A, Kiritsis D, eds. Advances in Production Management Systems Innovative and Knowledge-Based Production Management in a Global-Local World: Springer Berlin Heidelberg, 2014: 159-166.
47. ISO/CEI. IEC 62264-2. Enterprise-control system integration -- Part 2: Model object attributes2004: 96.
48. GS1. The GS1 EPCglobal Architecture Framework. GS1, 2013.
49. (FEVE) ECGF. GLASS PACKAGING DEMAND GROWTH : THE MARKET TRUSTS GLASS 2017.
50. Ribeiro A, Silva A, da Silva AR. Data modeling and data analytics: a survey from a big data perspective. Journal of Software Engineering and Applications 2015;8:617.
51. Grolinger K, Higashino WA, Tiwari A, Capretz MA. Data management in cloud environments: NoSQL and NewSQL data stores. Journal of Cloud Computing: advances, systems and applications 2013;2:22.
52. Chickering DM, Geiger D, Heckerman D. Learning Bayesian networks is NP-hard: Technical Report MSR-TR-94-17, Microsoft Research, 1994.
53. Tsamardinos I, Brown L, Aliferis C. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 2006;65:31-78.
54. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM 2008;51:107-113.

55. Downs JJ, Vogel EF. A plant-wide industrial process control problem. Computers & Chemical Engineering 1993;17:245-255.
56. Chiang LH, Braatz RD, Russell EL. Fault Detection and Diagnosis in Industrial Systems: Springer London, 2001.
57. Braatz PRD. Tennessee Eastman Problem Simulation Data [online]. Available at: http://web.mit.edu/braatzgroup/links.html. Accessed 03 march.
58. Yin S, Ding SX, Haghani A, Hao H, Zhang P. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. Journal of Process Control 2012;22:1567-1581.
59. Isermann R. Supervision, fault-detection and fault-diagnosis methods — An introduction. Control Engineering Practice 1997;5:639-652.