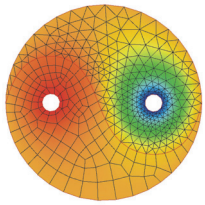
Publishing House
AKAPIT

COMPUTER METHODS IN MATERIALS SCIENCE

Informatyka w Technologii Materiałów

Vol. 7, 2007, No. 2



APPLICATION OF BAYESIAN NETWORK IN THE DIAGNOSIS OF HOT-DIP GALVANISING PROCESS

BARBARA MRZYGŁÓD [1], ANNA ADRIAN [1], STANISŁAWA KLUSKA-NAWARECKA [1,2], ROBERT MARCJAN [3]

[1] AGH UST, Department of Industrial Computer Science
[2] Center of Competence for Advanced Foundry Technology in Cracow
[3] AGH UST, Department of Computer Science

Abstract

This study presents an output of the application of a probabilistic method of inference based on Bayes' rule in the diagnosis of defects formed during hot-dip galvanising process. Bayesian cause-effect network for given group of surface defects and its causes was build. Many factors causing defects was taken into consideration in like: technological parameters, technological nodes and character of cause. The process of creating knowledge representation of the hot-dip galvanising process was disclosed on chosen defect (discontinuity of coating) and two causes (pH fluxing bath and surface contamination) along with a scheme of reasoning in Bayesian network and its implementation in a Norsys Netica packet. The advantages and drawbacks of a probabilistic method of representation of the incomplete and uncertain empirical knowledge were highlighted.

Key words: formalization of knowledge, uncertain knowledge, Bayesian networks, reasoning in a Bayesian network, faults of metal products, hot-dip galvanizing

1. INTRODUCTION

Certain areas of empirical knowledge are presented as data sets. The empirical data are a valuable source of information, providing their interpretation and the related inference are based on the methods which take into account that:

- the data are only a subset in the set of all possible realizations;
- the data are burdened with measuring errors or with a subjective assessment of observations;
- some unknown relationships may occur between the individual quantities;
- the conclusions regarding the selected quantities or relationships are formulated under the conditions of an incomplete and uncertain global information.

The methods of probabilistic calculus and mathematical statistics seems to be more than justified (Russell & Norvig, 1995)

2. PROBABILISTIC METHODS OF THE REPRESENTATION OF INCOMPLETE AND UNCERTAIN KNOWLEDGE

A rich reference literature is available on the calculus of probability and its diversified applications (Cowell, 1999; Plucińska & Pluciński, 2005). In these paper were used such terms as: unconditional probability (*a priori* probability), random variable, probability distribution, joint probability distribution, conditional probability (*a posteriori* probability).

Joint Probability Distribution (*JPD*) is applied when not one but several random variables are ex-

 ISSN 1641-8581

amined at the same time (a set of random variables). JPD is the table of probabilities of all the atomic events, when the term *atomic event* denotes some specific realizations assigned to random variables.

Assuming now that one of the examined random variables is *defect* (Y), while second variable is the *size* of this defect (X) characterized by the values $x_i \in \{large, medium, small\}$, then the resulting atomic events will be, e.g., *large pinholes, medium flaw.* Table 1 gives an example of joint probability distribution JPD (X,Y) for the above mentioned random variables X,Y. Using a JPD table, one can calculate the probability of occurrence of any arbitrary random event. Row *j* and column *i* in the JPD table show the probability of occurrence of an atomic event in the case when the value y_i is adopted by the variable Y, and simultaneously the value x_j is adopted by the variable X, e.g. P(X = *medium*, Y= *flaw*) = 0.09. The last row of Table 1 shows the distribution of variable Y, which is obtained by summing up the values in the fields of the columns; the last column shows the distribution of variable X, which is obtained by summing up the values in the fields of the rows; these are the marginal distributions of variable (X,Y). The sum of all the probabilities (atomic events) is 1, similar as the sum of marginal distributions, which proves that the JPD Table is correct.

The probability of occurrence of an event which consists in sampling a product with *flaw* or with a *small* defect (Y = *flaw* or X = *small*) can also be calculated from the JPD Table by summing up all the values in column Y = *flaw* and in row X = *small*, counting the values in the field of intersection of the column "*flaw"* and the row "*small*" only once (0.1 + 0.06 + 0.11 + 0.03 + 0.2 + 0.09 + 0.11 = 0.7).

Table 1. *Joint probability distribution JPD for the two random variables (X,W) and marginal distributions of variables X and Y.*

	Y= discontinuity	Y= scale	Y= lustreless surface	Y= pinholes	Y= flaw	Distribution of variable X
X= large	0.05	0.01	0.02	0.01	0.11	0.2
X= medium	0.1	0.03	0.07	0.01	0.09	0.3
X= small	0.1	0.06	0.11	0.03	0.20	0.5
Distribution Y	0.25	0.1	0.2	0.05	0.4	1

As we can see, the output will be the same as a probability value of the sum of joint events (X ∪ Y) because P (X ∪ Y) = P(X) + P(Y) – P(X ,Y) = 0.5 + 0.4 - 0.2 = 0.7

Conditional probability (posterior probability):

$$P(A/B) = \frac{P(A,B)}{P(B)} \qquad (1)$$

The conditional probability is used when it is necessary to define the probability of occurrence of an event and we possess some knowledge about other accompanying events. The conditional probability satisfies all axioms of the theory of probability, which means that it possesses the same features as an unconditional probability. The unconditional probability P(A) is, as a matter of fact, a specific case of the conditional probability P(A/) with zero condition, i.e., the condition about whose existence we have as yet no knowledge. The conditional probability can be computed from the JPD Table. For example, using Table 1, the probability of occurrence of *flaw* has been calculated for the detected defect which is *small*.

P(*flaw/small*) = P(*flaw, small*) / P*small*) =

0.2 / 0.5 = 0.4

If there are many random variables, using a JPD Table may be too costly (time consuming), since the dimension of a JPD Table equals number of the variables, while its size is a product of multiplication of the power of each set of the values of these variables.

3. REASONING BASED ON BAYES' THEOREM

In expert systems, the Bayes' theorem is a fundamental for reasoning based on the probabilistic methods (Kluska-Nawarecka, Marcjan & Mrzygłód, 2006). It is applied to determine the probability of occurrence of a hypothesis (H) as a result of the observed symptom (E).

For thus formulated problem, the Bayes' theorem (rule) assumes the form of:

$$P(H/E) = \frac{P(E/H)P(H)}{P(E)} \qquad (2)$$

where:

P(H|E) - the probability that a hypothesis H will be true, if symptom E has occurred;

P(H) - the probability of occurrence of an event which is a hypothesis;

P(E) - the probability of occurrence of an event which is a symptom;

P(E|H) - the probability of occurrence of a symptom, if the event which is a hypothesis has occurred.

In industrial practice there are usually many hypotheses associated with one symptom (e.g. the set of hypotheses {H1,...,Hm} is defined), but these should be the mutually disjoint (i.e. mutually excluding) events, and then the Bayes' rule will be expressed by equation (3)

$$P(H_i / E) = \frac{P(E / H_i)P(H_i)}{\sum_{k=1}^{m} P(E / H_k)P(H_k)} \quad (3)$$

On the other hand, if there are many hypotheses and many symptoms, equation (4) applies, or its reduced form (5). If this is the case, both hypotheses $H_1...H_m$ as well as the symptoms $E_1...E_n$ must be mutually excluding.

$$P(H_i / E_1, E_2, ..., E_n) = \frac{P(E_1, E_2, ..., E_n / H_i)P(H_i)}{\sum_{k=1}^{m} P(E_1, E_2, ..., E_n / H_k)P(H_k)} \quad (4)$$

$$P(H_i / E_1, E_2, ..., E_n) = \frac{P(E_1 / H_i)P(E_2 / H_i)...P(E_n / H_i)P(H_i)}{\sum_{k=1}^{m} P(E_1 / H_k)P(E_2 / H_k)...P(E_n / H_k)P(H_k)} \quad (5)$$

Thus determined Bayes' rule will express a prior probability, if there are conditions for existence of posterior probabilities, which are in many cases easier to derive and calculate.

3. THE SCHEME OF REASONING IN A BAYESIAN NETWORK

The Bayesian network is a graph in which:

- the nodes are random variables,
- (directed) edges represent a relationship of the type: "X has a direct influence on Y",
- every node X has an assigned conditional probability table *(Conditional Probability Table - CPT)* defining what influence on X have its predecessors (parents) in the graph,
- no (directed) cycles can be present.

To make a diagnosis of the surface defects in galvanized products, the following interpretation was adopted:

the symptom is a recorded *defect(s)* on product surface after galvanizing,

the hypotheses correspond to the potential *faults* of occurrence of a specific defect.

The relationships that occur between a subset in the set of defects in galvanised products $\{y_1, .., y_4\}$ and faults of occurrence of these defects is presented in fig. 1. The following parameters were taken into account: the location of the fault (the nodes of production process $\{P_0, ..., P_7\}$), the nature of the fault (a layer of the nodes P_{i_j}) and, finally, the parameters of the galvanising process (the variables X_i in the last column in Fig. 1).

As an example, a procedure of inference was described for an event which was related with the presence of discontinuity observed in coating (defect y_1). The task of the diagnosis was to indicate a (most probable) fault of occurrence of this defect. According to the network of associated relations (Fig.1.), the faults are to be searched in nodes P_0, P_2, P_5 (in reality, the number of the possible faults is much greater).

For more clarity in the presentation of the successive steps of inference it was necessary to further confine the examined area, and therefore in the described example only two potential faults of the coating discontinuity were taken into account. These are: the variable x_1 from the family of parameters P_{0_2} in node P_0, and the variable x_2 from the family P_{5_1} linked to the technological node P_5.

Further in the course of reasoning the following designations were used:

x_1 – surface contaminants out of standard	x_1 – contaminants within standard
x_2 – pH of fluxing bath out of standard	x_2 – pH of bath within standard
y_1 – discontinuity of coating observed	y_1 – no discontinuities observed

The symptom is a discontinuity of coating, which (being a random variable) assumes the value y_1, when the defect has been observed to occur, or the value $\neg y_1$, when absence of this defect has been recorded. A similar procedure was adopted when assigning the values to hypotheses, that is, to the faults of occurrence of a given symptom.

Basing on experts' opinion and using the empirical knowledge available, the following values of (prior) probabilities of occurrence of the faults of the examined defect were adopted:

P(x1)=0.1 and P(x2)=0.15 hence follows that P(¬x1)=0.9 and P(¬x2)=0.85 (rows 1 and 2 in Table 2).

Evaluating the influence of the examined causes on the probability of occurrence of a defect, the fol-

lowing conditional probabilities were assumed *a priori*:

P(y1/¬x1,¬x2)=0; P(y1/x1,¬x2)=0.5 ;
P(y1/¬x1,x2)=0.5; P(y1/x1, x2)=1.

Hence the following relationships were derived:

P(¬y1/¬x1,¬x2)=1; P(¬y1/x1,¬x2)=0.5;
P(¬y1/¬x1,x2)=0.5; P(¬y1/x1, x2)=0.

Table 2. *A compilation of the output of calculations for an example of the inference based on Bayes' theorem.*

1	x1	P(¬x1)=0,9	P(x1)=0,1	P(¬x1)=0,9	P(x1)=0,1	Σ
2	x2	P(¬x2)= 0,85		P(x2)=0,15		
3	P(x1,x2)	0,765	0,085	0,135	0,015	1
4	P(y1/x1,x2) *a priori*	0	0,5	0,5	1	-
5	P(y1,x1,x2)	0	0,0425	0,0675	0,015	
6	P(x1,x2/y1) *Bayes*	0	0,34	0,54	0,12	-

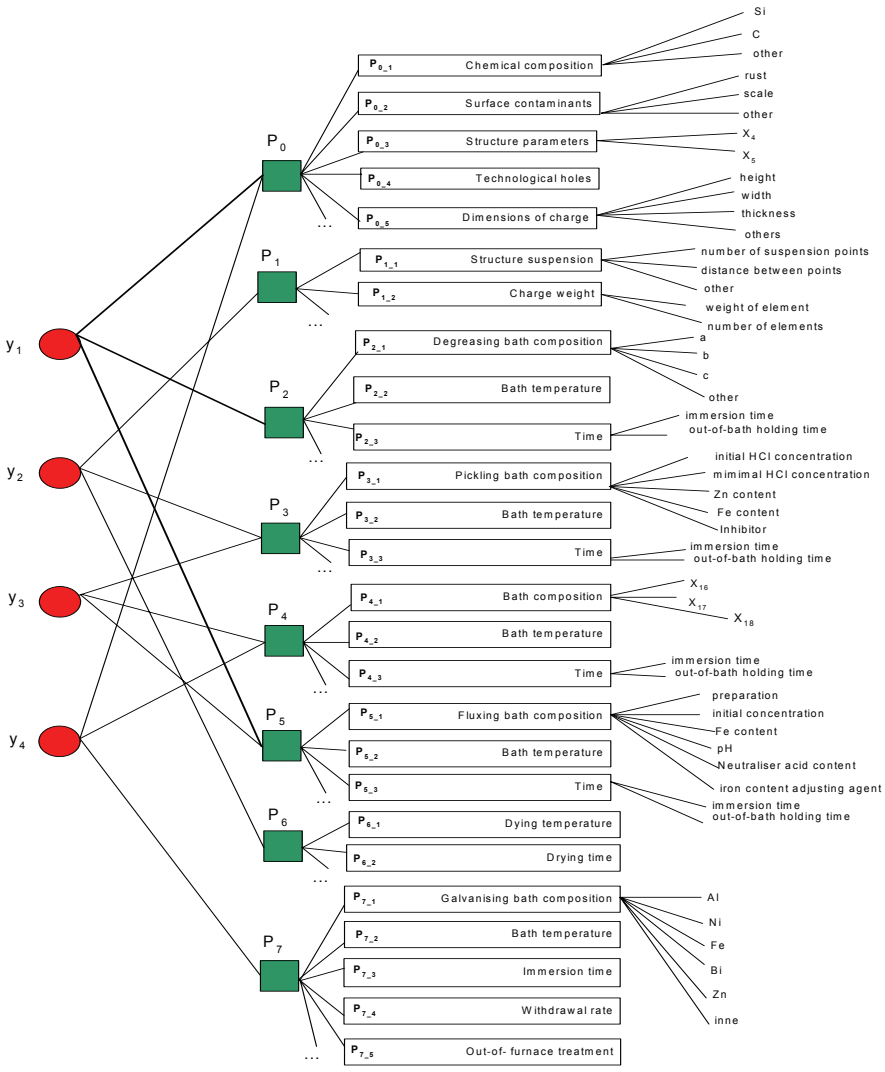


Fig. 1. *Example of a network of causal-resultant relationships between the selected defects (y_1.. y_4), production nodes (P_0..P_7), and technological parameters.*

This set of the conditional probabilities forms a CPT Table (row 4, Table 2). This is the table of the input data to calculate the value of the probability of occurrence of the events examined during inference made in Bayesian network. It is easy to note that the dimension of a CPT Table is smaller than that of a JPD Table and comprises only the prior probabilities, which makes calculations much easier and less time consuming than the calculation of values in a JPD Table.

Table 2 gives values of the probabilities which occur in the inference process based on Bayes' rule. Rows 1,2,4 hold the input data (prior probabilities), while the intermediate results in row 3 hold the probability of occurrence of an event which consists in a simultaneous occurrence of two independent events x1 and x2, while row 5 holds the probability of a simultaneous occurrence of the defect and of both faults analysed here. After summing up the values in row 5, the value of the total probability (equal to 0,125) was obtained. In row 6 of Table 2, a final output of the computations based on Bayes' rule is given. From this output it follows that with 0,54 probability the truth of the hypothesis x2 can be ascertained (the fault of the defect was incorrect pH of the bath), with 0,34 probability the hypothesis x1 is considered true (the fault of the defect was surface contamination), while 0,12 means the probability that both hypotheses are true (that is - both faults are responsible for the occurrence of defect).

In the reasoning presented here a rather bold assumption has been made (for the sake of clarity). Namely, it has beeen assumed that the two faults under consideration are the only possible faults of occurrence of the defect. This is expressed by the probability value equal to 0 in the first column, row 6, Table 2.

4. IMPLEMENTATION

As a drive for implementation of the example of a Bayesian network created for the

process of hot-dip galvanising, a Norsys Netica packet was used (Norsys Software Corp., 2004). A graphical representation of the (causal-resultant) network from figure 1 after implementation in the packet is shown in figure 2. Comparing both networks one can see that all the nodes of the network in figure 2 hold all the values of the examined variable along with the respective probabilities of occurrence (expressed in %), while the network in Figure 1 gives only the names of the variables and their relationships.

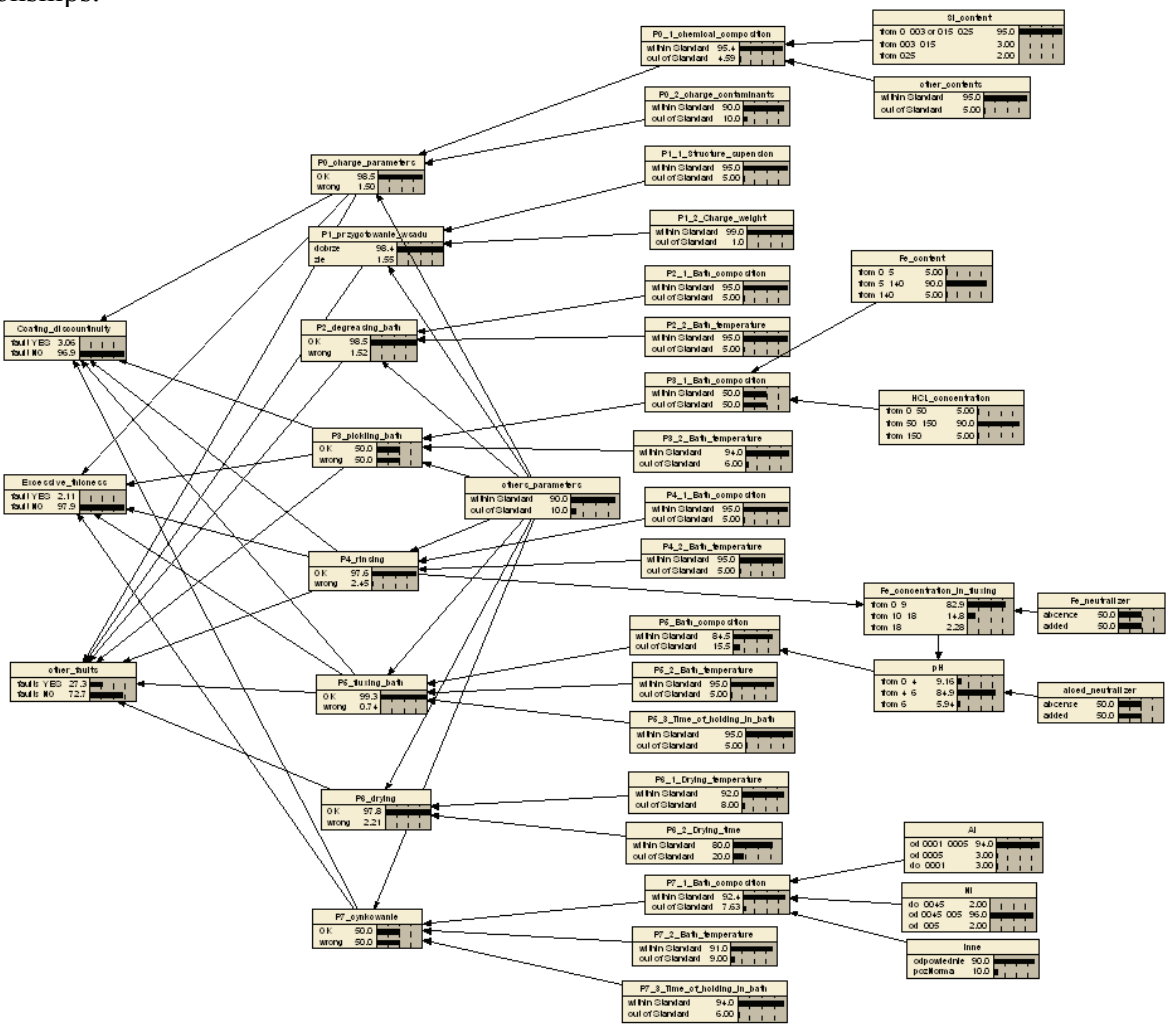
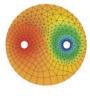


Fig.2. *Graphical form of a network of causal-resultant relationships from Figure 1 as implemented in the Netica packet module.*

For example, the variable named *defects*, assuming the linguistic values from a set comprising the names of possible defects (e.g. discontinuity of coating, thickness of coating, etc.), is in the network represented by a random variable which in all probability will take the value YES (the presence of this specific defect has been ascertained) or NO (figure 3). The technological parameters are the most specified faults of the occurrence of defect. The values of most of them are determined quite precisely in the form of numerical sets (intervals) (e.g. content, temperature, time, concentration). This is illustrated in figure 4, wherefrom it follows that in 5 % of all cases, the concentration of HCL is below 50 [%].

Depending now on whether in a given family of the technological parameters (P_{i_j}) all of the examined parameters (X_i) can satisfy the requirements of a standard, the intermediate node P_{i_j} may assume the values of either *within-Standard* or *out-of-Standard*. On the other hand, the value of the random variable represented by P_i will be influenced by the values of the individual components P_{i_j}. The variable P_i takes values from the set {right, wrong}, as shown in Figure 5.

Nodes P_i and P_{i_j} comprise the conditional probabilities defined in a CPT Table. An example of this table for node P_3_pickling is presented in figure 6. The probabilities accepted *a priori* were determined from the empirical data and experts' (process engineers in a galvanising shop - in this case) knowledge. Figure 7 shows a fragment of the table comprising percent share of different faults in the occurrence of a specific defect.



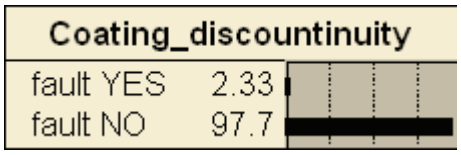


Fig. 3. *Node representing the random variable defect named discontinuity of coating.*

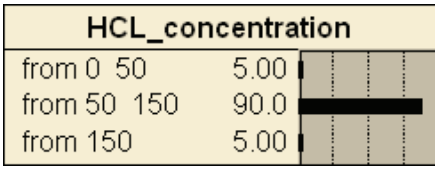


Fig. 4. *Node representing the random variable fault named HCl_concentration.*

for the computation of probabilities enables both forward and backward reasoning, that is, making diagnosis about the faults when the event of occurrence of a specific defect has already been acknowledged. Some of the variables are therefore event variables. These are the variables whose exact values (e.g. taken from measurements or observations) are known. The remaining variables in the network are the query variables, for which the conditional probability is computed in respect of the event variables. The choice of the inference direction is left at the discretion of the packet user when he is introducing the input data. If, on entry, the algorithm receives information that the probability of occurrence of a defect(s) is 1, and some values of the technological process parameters are given, the algorithm

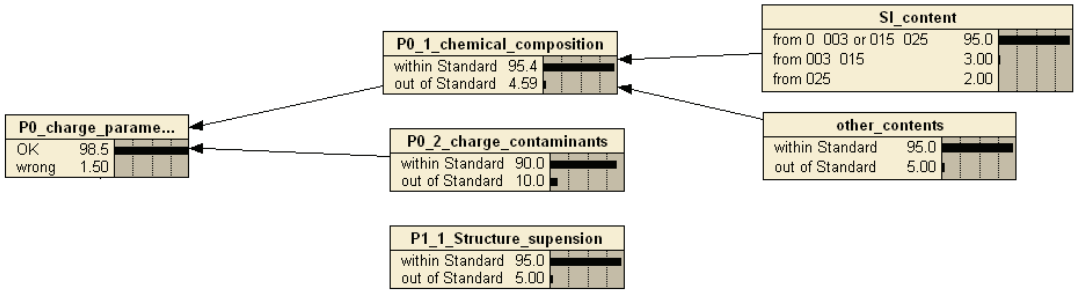
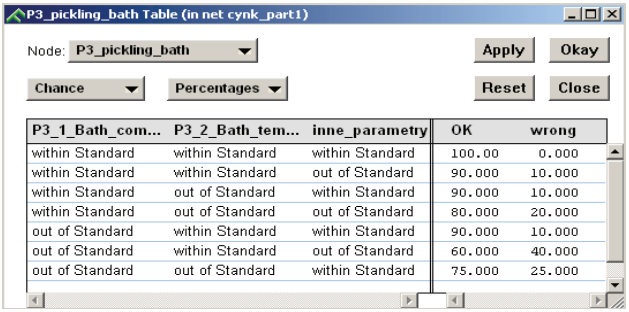


Fig. 5. *Node representing the intermediate random variable Pi and components influencing the value of this node.*



P3_1_Bath_com...	P3_2_Bath_tem...	inne_parametry	OK	wrong
within Standard	within Standard	within Standard	100.00	0.000
within Standard	within Standard	out of Standard	90.000	10.000
within Standard	out of Standard	within Standard	90.000	10.000
within Standard	out of Standard	out of Standard	80.000	20.000
out of Standard	within Standard	within Standard	90.000	10.000
out of Standard	within Standard	out of Standard	60.000	40.000
out of Standard	out of Standard	within Standard	75.000	25.000

Fig. 6.. *A CPT Table of conditional probabilities for a node with the random variable P3_pickling [%].*

is capable of computing the probability of occurrence of the individual faults of the indicated defect. This is an example of the diagnostic reasoning. Quality control of a technological process uses forward reasoning, which means that, basing on the recorded real values of the technological parameters, one can compute the probability of occurence of some specific defects in final products. This information can be used in preventive measures taken to avoid the occurrence of these defects (current adjustments).

fault	cause	probability
Coating discontinuityi	wrong chemical composition of galvanised	1
	contaminants non - removable during surface	80
	technological holes ill selected or located	6
	wrong suspension of structure	1
	degreasing bath composition	1
	degreasing bath temperature	1
	degreasing time	1
	pickling bath composition	1
	pickling temperature	1
	pickling time	1
	fluxing composition	1
	fluxing temperature	1
	fluxing time	1
	after - fluxing drying time	1
	galvanising bath composition	1

Fig. 7. *A fragment of the table stating the share of various faults in occurrence of a chosen defect.*

5. SUMMARY

An advantage of the Bayesian network as applied in this study is the fact that an algorithm used

The Bayesian network has also some drawbacks. Probably the most important one is the fact that it is necessary to possess some knowledge about many probabilities, not always easy to estimate. The prior probabilities determined from statistical data (frequencies of occurrence) have to be supported by a suffiently great number of the data representative of a given population, and when they are determined by humans, may these be

the best experts even, an error resulting from subjective evaluation is always possible.

Moreover, in this approach, the computations are based on the use of some formulae (e.g. equations 3-5), which are true only under certain conditions, e.g. when independence or mutual exclusion of events exists, which need not always be true in practice.

This study was done under Ministry of Scientific Research and Information Technology: Project KBN 3 T08C 061 26, Project PBZ-KBN-114/T08/2004.

REFERENCES

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J., 1999, Probabilistic networks and expert systems, in: *Statistics for Engineering and Information Science*, Springer, New York .

Kluska-Nawarecka, S., Marcjan, R., Mrzygłód, B., 2006, Ocena możliwości implementacji wybranych metod reprezentacji wiedzy dla potrzeb diagnostyki wad powierzchni wyrobów metalowych, *Proc. 13th Conf. Informatyka w Technologii Metali - KomPlasTech*, eds, Szeliga, D., Pietrzyk, M., Kusiak, J., Szczawnica, 23-30 (in Polish).

Norsys Software Corp., 2004, http://www.norsys.com/

Plucińska A., Pluciński T, 2005, *Probabilistyka. Rachunek prawdopodobieństwa. Statystyka matematyczna. Procesy stochastyczne*, Wydawnictwo Naukowo-Techniczne, Warszawa, (in Polish).

Russell S.J., Norvig P., 1995, *Artificial Intelligence: a modern approach*, Prentice Hall, Englewood Cliffs.

ZASTOSOWANIE SIECI BAYESA W DIAGNOSTYCE PROCESU CYNKOWANIA OGNIOWEGO

Streszczenie

W artykule zaprezentowano wyniki zastosowania probabilistycznej metody wnioskowania, opartej na regule Bayesa, w diagnostyce wad powstających w procesie cynkowania ogniowego. Pokazano proces tworzenia reprezentacji wiedzy dotyczącej procesu cynkowania. Zbudowano sieć przyczynowo-skutkową dla wybranej grupy wad powierzchni wyrobów ocynkowanych i przyczyn ich powstawania. Analizując przyczyny uwzględniano miejsce ich występowania, charakter (rodzaj) i reprezentujące je parametry technologiczne. Schemat wnioskowania w sieciach Bayesa pokazano na przykładzie wybranej wady (nieciągłość powierzchni) i dwóch przyczyn (pH kąpieli topnikującej oraz zanieczyszczenie powierzchni). Do implementacji tej sieci Bayesa wykorzystano pakiet Netica firmy Norsys. Wskazano na zalety i wady probabilistycznej metody reprezentacji niepełnej i niepewnej wiedzy empirycznej.

Submitted: October 13, 2006
Submitted in a revised form: December 4, 2006
Accepted: December 12, 2006