Dynamic Bayesian Networks for classification of rail defects

Abdeljabbar Ben Salem
Laboratory of New Technologies
INRETS
2, Av du G[al] Malleret Joinville 94114 Arcueil
France
bensalem@inrets.fr

Laurent Bouillaut
Laboratory of New Technologies
INRETS
2, Av du G[al] Malleret Joinville 94114 Arcueil
France
bouillaut@inrets.fr

Patrice Aknin
Laboratory of New Technologies
INRETS
2, Av du G[al] Malleret Joinville 94114 Arcueil
France
aknin@inrets.fr

Philippe Weber
Centre de Recherche en Automatique de Nancy
Université Nancy 1
2, rue Jean Lamour
54519 Vandoeuvre-les-Nancy
France
weber@esstin.uhp-nancy.fr

Abstract – **This paper deals with the problem of classification of sequential events that occur one after the other and when the different prior transition probabilities can be approached with the help of a labelled database. Dynamic Bayesian Networks (DBN) are employed to formalise such complex dynamic process through a compact representation. DBN allow simulating these processes, taking into account information about time detection. DBN modelling were tested, until the third order, on an experimental problem that concerns the classification of rail defects.**

I. INTRODUCTION

The classification problem can be considered with different points of view according to the fact that the observation is a link in a sequential series or not: simple inference of posterior class membership probabilities, detection problem, time series prediction, pattern recognition problem.

The diagnosis problem is often a sequential process where several default detections occur ones after the others (see figure 1). In order to increase the present decision relevancy, the past decisions and their elapsed times will be taken into account in this research work.

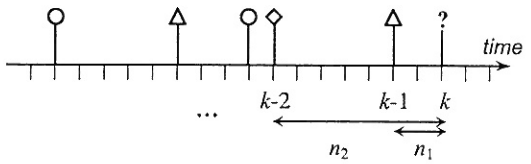

Fig. 1. Sequence of events

The aim of this paper is to use DBN as an equivalent model to the MC [1]. The problems considered here are those whose dynamics can be modelled as stochastic processes and where the decision maker's actions influence the system behaviour. Past system states and the applied action jointly determine the probability distribution over the next states [12].

In connection with our specific application, three DBN modelling Input-Output Hidden Markov Model (IO-HMM) [2] are compared in terms of good detection rate as well as confusion rate. Section 2 introduces briefly theoretical aspects of tools.

Whatever the adopted technique, a preliminary phase of transition probability modelling must be achieved. This can be realized thanks to prior knowledge about the transition possibilities (in physical sense) and/or thanks to a labelled database from which it is possible to "learn" the system behaviour. The second approach has been adopted in section 3 and the modelling is considered with the help of 1, 2 or 3 slices of time. A theoretical comparison based on error calculation between models is presented in section IV. Complete results are discussed in section V that presents our specific application.

II. THEORETICAL DEFINITIONS

A. Bayesian Networks

Bayesian Networks (BN) are causal probabilistic networks based on the graphs theory. They are commonly defined as Directed Acyclic Graph and are used to represent uncertain knowledge [3]. Figure 2 introduces an example of BN, modelling a simple system, characterised by two variables, X_i and X_j. In this article, only the case of discrete variables is considered with d possible states (X_i =1, 2, ..., d).

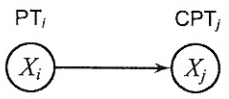

Fig. 2. A basic BN

Nodes represent the various variables of the system (defined over all its possible states) and the connecting arrows indicate the causality between these variables. Root nodes (nodes having only descendants) are described by a probability table (PT) composed of the class belonging probabilities $P(X_i)$ over the d possible value of X_i. Children nodes (node having one parent at least) are described by conditional probability tables (CPT), composed of $P(X_j \mid X_i)$ over each value of X_j knowing the value of X_i.

B. Markov Chains

Markov Chains (MC) are well known and commonly used. Let $X_0, X_1, \ldots X_k$ be a sequence of random variables, taking their values in a same finite discrete space E={1,...,d}. The sequence is an m[th] order MC if the following property is verified [5].

$$P(X_k = i_k \mid X_{k-1} = i_{k-1}, \ldots, X_{k-m} = i_{k-m}, \ldots, X_0 = i_0) =$$

$$P(X_k = i_k \mid X_{k-1} = i_{k-1}, \ldots, X_{k-m} = i_{k-m}) = \pi_{i_k, i_{k-1}, \ldots, i_{k-m}}$$

Note that the case of homogeneous MC is discussing here; the stationarity implies that transition probability of one step transition does not change as k increases [6]. Many other definitions and properties can describe MC. They can be found in many former publications such as [5] [7]. In this work, only first and second order MC are considered. Despite the similarities between MC and BN, we point out that the set E must be the same for all X_k in the MC case, while it can change in BN case.

III. MODELLING THE DYNAMICS

The aim of this study is to determine the most relevant approach for strongly structured sequence of defaults using DBN. As indicated in section 1, the causality is the main notion that the modelling phase must be considered. According to each specific application, few slices of time are necessary. In our case, only one or two slices of time will be used.

It is important to notice that not only the nature of the past decisions infer the present decision, but also the durations between the present and the past time detections. Furthermore, the model is represented by IO-HMM if the distribution over the states of the exogenous variable is known. The difference between standard HMM or IO-HMM is that HMM represents the distribution of $P(X_k)$ whereas the IO-HMM represent the conditional distribution $P(X_k|U^T_k)$ given the input sequence $U_k^T = (u_1, ..., u_m)$. Where U_k represents the exogenous constraint (Input) with states $\{X_1, X_2, ..., X_m\}$ [4].

A. Considering the one last detection

An extended version of BN, including temporal information, provides the Dynamic Bayesian Networks (DBN). In this version, the probability distribution change over the time [8] [9], and the inference process is made of a bayesian term and a time feedback term. Therefore, it is possible to calculate the distribution of X at the present time knowing the distribution of X in the past time.

The first model (Mod1) uses the one last detection. X_{k-1} and X_k correspond to the same variable considering at two different times, characterised by three states. Note that the conditional probability tables depend on the n_1 duration between the k^{th} and the k-1^{th} detections. Figures 3a and 3b introduce this modelling, respectively, by a bayesian network and a Markov chain.

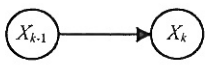

Fig. 3a. Mod1 modelled by a Bayesian Network

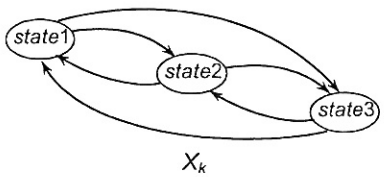

Fig. 3b.Mod1 modelled by a Markov Chain

In fact, these two approaches are perfectly equivalent [4]. Indeed, the CPT describing X_k is exactly the transition probability matrix used in the Markov chain. As we can see in the two previous figures, the markovian representation can quickly be difficult to represent and less understandable. This is the reason why, for all this study, our models are represented by using of BN rather more than MC.

Finally, the process under study is characterised by the various CPT, changing according to different rules (u_n). So one representation of this problem can be done as the following IO-HMM [10]:

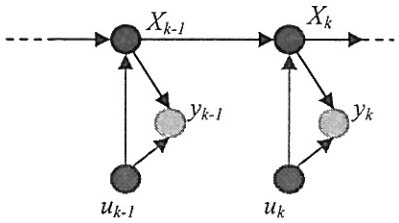

Fig. 4. Un-rollup DBN modelling IOHMM for Mod1

(u_k) and (y_k) representing, respectively, the inputs and the outputs of this IO-HMM.

Another solution, which keeps a compact network form, is based on iterative inferences [4]. This solution is used in the following. Indeed, it is possible to compute the probability distribution of any variable X_i at time step k+1 based on the probabilities corresponding to time step k. The probability distributions at time step k+2... are computed using successive inferences. Then, a network called 2-TBN [11], [14] can be defined with two slices and is introduced in figure 5.

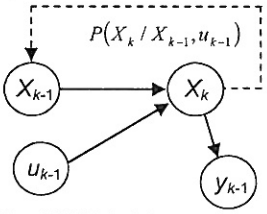

Fig. 5. 2-TBN modelling IOHMM with temporal node X_k and exogenous observations u_k.

B. Considering the two last detections

The second model considers that the nature of the k^{th} detection depends, not only, on the last one but on the next to last one too. For this model, two different cases are considered. First, let us consider that the k-1^{th} and the k-2^{th} detections are dependant variables. Then, the un-rollup DBN modelling the equivalent IOHMM of this model (Mod2[1]) is introduced in the following figure:

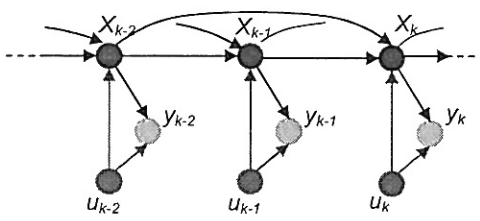

Fig. 6. Un-rollup DBN modelling IOHMM for Mod2[1]

As in section III A, an another representation of Mod2[1] is define: The 3-TBN, introduced in the figure 7.

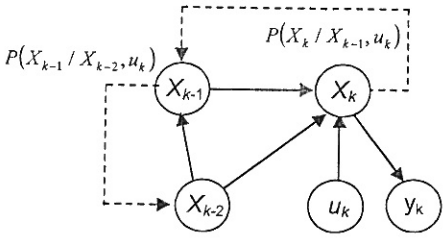

Fig. 7. 3-TBN modelling IOHMM for Mod2[1].

In our second approach considering two slices of time, it is assumed that k-1[th] and k-2[th] are independent. One of the aims of this hypothesis was to compare the influence of the link between X_{k1} and X_{k-2} (as done precisely in the next section). Figures 8 and 9 introduce respectively the un-rollup DBN and the 3-TBN modeling IO-HMM for Mod2[2].

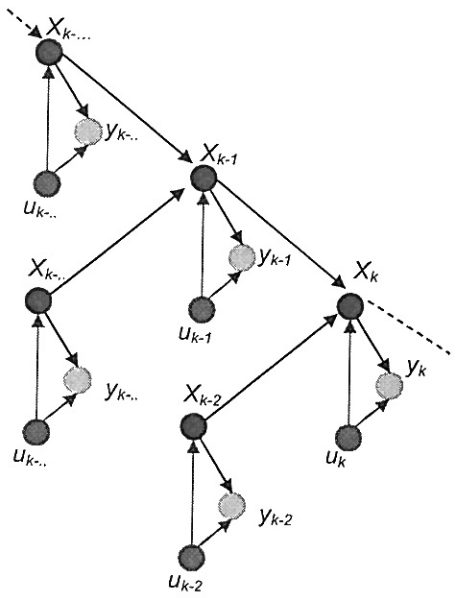

Fig. 8. Un-rollup DBN modelling IOHMM for Mod2[2]

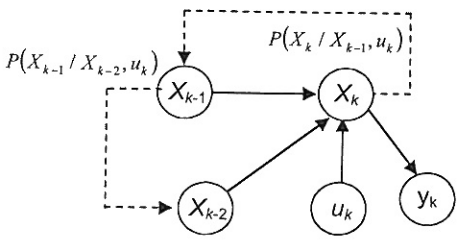

Fig. 9. 3-TBN modelling IOHMM for Mod2[2].

IV. SENSITIVITY OF MOD2 MODELS

A somewhat surprising outcome of the experimental comparison of BN modelling (introduced in section V) is that Mod2[2] gives better results. In fact, contrary to this modelling, Mod2[1] uses more information about past (notably $P(X_{k-1}/X_{k-2})$) to predict the state X_k. This section investigates the sensitivity to errors of the learning phase of Bayesian modelling. Now, according to the Bayesian theory, the likelihood is:

$$P(x_n = k) = \sum_{i,j} P(X_n = k, X_{n-1} = i, X_{n-2} = j) \quad (1)$$

Using the product rule, (1) becomes:

$$P(x_n = k) = \sum_{i,j} P(X_{n-2} = j)\ P(X_{n-1} = i \mid X_{n-2} = j)... \\ ...P(X_n = k \mid X_{n-1} = i, X_{n-2} = j) \quad (2)$$

In Mod2[2], X_{n-1} and X_{n-2} are independent:

$$P(X_{n-1} = i \mid X_{n-2} = j) = P(X_{n-1} = i)$$

So, in that case, equation (2) can be written as:

$$P(x_n = k) = \sum_{i,j} P(X_{n-2} = j) P(X_{n-1} = i)... \\ ...P(X_n = k \mid X_{n-1} = i, X_{n-2} = j) \quad (3)$$

The difference between rules (2) and (3) is that $P(X_{n-1}=i|X_{n-2}=j)$ is obtained by learning whereas $P(X_{n-1}=i)$ is the results of calculations. We will, therefore, get interested in the influence of the learning phase on inference algorithms of Mod2[1] and Mod2[2].

Assuming that conditional probabilities obtained after the learning phase are biased as following:

$$\begin{cases} \hat{P}(X_{n-1} = i \mid X_{n-2} = j) = P(X_{n-1} = i \mid X_{n-2} = j) + e_{ij} \\ \hat{P}(X_n = k \mid X_{n-1} = i, X_{n-2} = j) \\ \qquad = P(X_n = k \mid X_{n-1} = i, X_{n-2} = j) + E_{ijk} \end{cases} \quad (4)$$

A. Sensitivity of Mod2[2] model

In Mod2[2], only E_{ijk} is present because of the independency between X_{n-2} and X_{n-1}. So, using equation (3), we obtain:

$$\hat{P}(X_2 = k) = \sum_{i,j} P(X_{n-2} = j)\ P(X_{n-1} = i)\left[E_{ijk} + \right. \\ \left. P(X_n = k \mid X_{n-1} = i, X_{n-2} = j)\right] \quad (5)$$

So the global calculation error is defined as:

$$E_1 = \hat{P}(X_n = k) - P(X_n = k) \\ = \sum_{i,j} P(X_{n-2} = j)\ P(X_{n-1} = i) E_{ijk} \quad (6)$$

The variation of E_1 in respect to E_{ijk} is given by:

$$\frac{\partial E_1}{\partial E_{ijk}} = \sum_i P(X_{n-1} = i) \sum_j P(X_{n-2} = j) = 1 \quad (7)$$

So the global calculation error E_1 linearly depends on E_{ijk}.

B. Sensitivity of Mod2[1] model

Before determining the error sensitivity to learning in case of Mod2[1], the relationship between errors e_{ij} and E_{ijk} should be determined. Probabilities $P(X_{n-1} = i \mid X_{n-2} = j)$ and $P(X_{n-1} = i)$ are linked by the relation:

$$\hat{P}(X_n = k \mid X_{n-1} = i) = \sum_j \hat{P}(X_n = k, X_{n-2} = j \mid X_{n-1} = i) \quad (8)$$

Using the product rule, this equation becomes:

$$\begin{aligned} &\hat{P}(X_n = k \mid X_{n-1} = i) = \\ &\quad \sum_j P(X_{n-2} = j)\, \hat{P}(X_n = k \mid X_{n-1} = i, X_{n-2} = j) \\ &= \sum_j P(X_{n-2} = j)\, (P(X_n = k \mid X_{n-1} = i, X_{n-2} = j) + E_{ijk}) \\ &= P(X_n = k \mid X_{n-1} = i) + \sum_j P(X_{n-2} = j)\, E_{ijk} \end{aligned} \quad (9)$$

Equations (4) and (9) lead to:

$$e_{ik} = \sum_j P(X_{n-2} = j)\, E_{ijk} \quad (10)$$

Using equations (2) and (4), the following equation is obtained:

$$\begin{aligned} \hat{P}(x_n = k) &= \sum_{i,j} P(X_{n-2} = j)\, \hat{P}(X_{n-1} = i \mid X_{n-2} = j) \ldots \\ &\quad \ldots\ \hat{P}(X_n = k \mid X_{n-1} = i, X_{n-2} = j) \\ &= \sum_{i,j} P(X_{n-2} = j)\ (P(X_{n-1} = i \mid X_{n-2} = j) + e_{ij}) \ldots \\ &\quad (P(X_n = k \mid X_{n-1} = i, X_{n-2} = j) + E_{ijk}) \end{aligned} \quad (11)$$

Using this equation and the relation (10):

$$\begin{aligned} E_2 &= \hat{P}(X_k = i) - P(X_k = i) \\ &= \sum_{i,j} P(X_{n-2} = j)\, \big[P(X_n = k \mid X_{n-1} = i, X_{n-2} = j) \ldots \\ &\quad \ldots \sum_l P(X_{n-2} = l) E_{ilj} + P(X_{n-1} = i \mid X_{n-2} = j) E_{ijk} \ldots \\ &\quad \ldots + E_{ijk} \sum_l P(X_{n-2} = l) E_{ilj}\ \big] \end{aligned} \quad (12)$$

So $\frac{\partial E_2}{\partial E_{ijk}}$ depends on E_{ijk} (because of square term on E_{ijk}).

To conclude, Mod2[1] is therefore much more sensitive to learning errors than Mod2[2]. This theoretical calculation will be confirmed by our application introduced in the section V.

V. APPLICATION TO RAIL DEFECT CLASSIFICATION

A. Context

The infrastructure maintenance is an important field of interest for the railway operators. Particularly, the rail integrity is a critical subject for train control as well as maintenance. For the both, a specific eddy current sensor has been developed to detect any electromagnetic characteristic variation of the rail [12]. Therefore, the real defects are detected (as broken rail or shellings) but singular points too. These expected points are welded joints (WJ), fishplatted joints (FJ) and switch joints (SJ) that join together rail lengths.

The distinction between real defects and singular points is obviously an essential task that must be achieved by the decision system. Its first implementation has been successfully tested in July 2002, on the Paris metro network [13]. Major defects - as splitted rails - have been correctly localized. On the other hand, the minor defect detection rate has been estimated about 72%. It is due to a less energy level of this kind of defect that are mistaken with some singular joints.

This section presents new developments about the improvement of the minor defect detection rate by using an additional pattern recognition module. At the contrary of the previous decision module where only sensor measurements are used to infer the classification decision [13], prior knowledge about the railway track structure are introduced into this additional module. In fact, the sequence of the three possible joints FJ, WJ and SJ (d=3) are greatly constrained by the track structure and its setting procedures. For example, a WJ joint is often 18 meters away from one another, corresponding to the length of manufactured rails. A FJ joint is surrounded by two WJ joints at 3 meters away…

The SIAM database of RATP Company contains the position of all joints of the Paris metro tracks due to the initial track setting or generated by maintenance operations. A part of the SIAM base corresponding to 8 iron-wheel-on-rail tracks (18387 WJ, 2203 FJ and 288 SJ joints) is used in the following. A statistical study of the database underlines that the use of 3 past detections is sufficient to well represent the information.

The CPT and transition matrix of our models are estimated on the database.

B. Results with one slice of time

Many studies tried to compare Bayesian and Markovian approaches [1]. As noted in section III A, the representations of Mod1 by use of a Markov Chain or a Bayesian network are two equivalent approaches. Our first application to rail defect classification confirms this property. In fact, results introduced in Table 1 were obtained by both MC and BN approaches.

Nature of joint	Classification WJ	FJ	SJ
WJ	**97%**	3%	0%
FJ	52.4%	**47.6%**	0%
SJ	64.9%	35.1%	**0%**

Table 1. Mod1 Confusion table

97% of the WJ are correctly detected and 47.6% of the FJ too. It must be remained that the classification decision has been taken with the only knowledge of the class belonging probabilities of the one past detection and its distance with respect to the present detection.

The observation of SIAM database underlined that switches were very few. Moreover, they are characterized by four successive points. In this section, only two points (k and k-1) are considered. So, this approach can hardly detect the SJ.

C. Results with two slices of time

Nature of joint	Classification WJ	FJ	SJ
WJ	**95.3%**	4.7%	0%
FJ	58.2%	**41.3%**	0.5%
SJ	50.7%	47.5%	**1.8%**

Table 2. $Mod2^1$ confusion table

Nature of joint	Classification WJ	FJ	SJ
WJ	**97.4%**	2.6%	0%
FJ	50.6%	**49.2%**	0.2%
SJ	40.2%	57%	**2.8%**

Table 3. $Mod2^2$ confusion table

As noticed previously and in [1], the property of equivalence between BN with one slice of time and first order MC can be extended to the second order. Nevertheless, two different models for $Mod2^1$ and for $Mod2^2$, are computed (considering both the MC approach and the BN approach) to confirm the property of equivalence for the second order. As expected, each approaches furnished exactly the same classification. These results are summarised in Tables 2 and 3.

Better results were expected in the 2nd order case compared with the 1st order case. The tables 2 and 3 go contrary to our plans, especially in the $Mod2^1$ case. The main explanation lies in the fact that the SIAM database groups together old tracks and new ones. Their structures can be quite different, highlighting track history (frequent maintenance operations with WJ achievements for the old tracks and great evenness of the joint distribution on the new tracks). The learning phase, that estimates CPT, averages all the possible track structures and the second order algorithms are more sensitive to the distance between the mean track and the analyzed one and, therefore, can introduced bias, more significant for second order approaches than for the first order model.

Tables 2 and 3 underline that $Mod2^2$ gives better results than $Mod2^1$. Indeed, as noticed in section III, the difference between their algorithm implementation lies into the use of different probabilities for the estimation of probability tables: $P(X_{k-1})$ is used only in the $Mod2^2$ case and $P(X_{k-1} \mid X_{k-2})$ only in the $Mod2^1$ case. The first probability is obtained by inference whereas the second one is obtained by learning.

$Mod2^2$ is therefore less sensitive to the learning phase and its errors than $Mod2^1$.

Finally, we can note that, on both cases, the detection of SJ is still non existent. This is mainly due to the fact that there are very few SJ in the system. So, the only way to discriminate them should be to detect sequences characterising only SJ. These sequences exist but are characterised by more than three points. So, it is perfectly logical that SJ were badly detected with our first models introduced in section III A and III B.

D. Results with three slices of time

As indicated in previous sections, switches (SJ) are highly structured, involving a three slices of time approach. In order to reduce the complexity of the 4-TBN modelling IO-HMM, recognition masks are introduced to adjust the specific patterns of switches. A previous study underlined that a weak number of masks could discriminate most of switches. They are described by three distances $\{n_1, n_2, n_3\}$ with $n_1 \in \{3, 4\}$, $n_2=5$ and $n_3 \in \{3, 4\}$. An evolution of $Mod2^2$, including these information was therefore developed. This new model is called Mod2+ and the table 4 summarizes its results. The SJ well detection rate reaches 78.5%.

Nature of joint	Classification WJ	FJ	SJ
WJ	**97.4%**	2.6%	0%
FJ	45.8%	**53.7%**	0.5%
SJ	10.4%	11.1%	**78.5%**

Table 4. Mod2+ confusion table

We can note that the well detection rate of SJ reach an acceptable value. Of course, compare with rates commonly met in diagnosis problems, this value can appear as quite weak. In fact, for our specific application and, considering the number of switches and the local classifier which will be used, this model brings significant information to our global classifier. Moreover, computably speaking, it should be unreasonable to view higher order models.

VI. CONCLUSION

In this paper, a pattern recognition process has been introduced for short time sequence classification. Its input is constituted with few past detection observation and the inference algorithm is based on IO-HMM. To simplify representations and implementations, n-TBN modelling IO-HMM are considered. First, second and third order models have been tested.

After introducing theoretical aspects, we considered two different second order models and compare, theoretically, the influence of leaning errors on the quality of their

results. For our specific application, the learning phase is still a sensitive step. We therefore decided to keep the model minimizing the influence of learning errors. Finally, we introduced a simplified three slices of time network, advancement of the previous two slices of time network.

Contrary to expected results, higher order models do not correspond to a really better behaviour. Indeed, it remains some errors at the end of the learning phase, due to certain inconsistently of the database. The 2nd order models seem to be more sensitive to these kinds of errors than 1st order models.

Finally, the Mod2+ model improves the Bayesian results by combining a 2nd order DBN with 3rd order recognition masks that characterize a particular time sequence (SJ).

The success of bayesian networks depends on the learning of conditional probability tables. This phase is essential and the results can drastically change with the choice of database. For our application, a divide-to-win approach will be adopted where the database will be split into several homogeneous subsets.

The final development will concern an expert committee for classification that will associate the BN and a local classifier based on the eddy current signature of each detection.

VI. ACKNOWLEDGEMENT

This work was supported by RATP Company, EST Department.

VII. REFERENCES

[1] S. Padhraic, "Belief networks, hidden Markov models, and Markov random fields: A unifying view", *Pattern Recognition Letters*, Vol. 18, pp. 1261-1268, 1997.

[2] Y. Bengio and P. Frasconi, "Input/Output HMMs for Sequence Processing", *IEEE Transaction on Neural Networks*, Vol. 7, no. 5, September 1995, pp.1231-1249.

[3] F.V. Jensen, *An Introduction to Bayesian Networks*, UCL Press, London, 1996.

[4] P. Weber, L. Jouffe, "Reliability modelling with dynamic bayesian networks", Safe Process, 2003, 7-62.

[5] D. Gamerman, *Markov Chain Monte Carlo, Stochastic simulation for Bayesian inference*, Chapman & Hall, London, 1997.

[6] Y. Bengio, P. Frasconi, "Diffusion of credit in Markovian models", NIPS 7, 1995, pp. 553-560.

[7] R.A. Howard, *Dynamic probabilistic systems: V1, Markov Models, V2, Semi-Markov and Decision Processes*, Wiley, 1971.

[8] K.P. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning", PhD Dissertation, University of California, Berkeley, 2002.

[9] U. Kjaerulff, "Hugin : a computational system for dynamic time-sliced Bayesian networks", International journal of forecasting, Vol. 11, 1995 pp. 89-111.

[10] Y. Bengio, "Markovian models for sequential data", Neural Computing Surveys, Vol. 2, 1999, pp. 129-162.

[11] X. Boyen, D. Koller, "Tractable for complex stochastic processes", in *Proceedings of the 16th annual conference on uncertainty and AI*, 1998, pp. 33-42.

[12] Oukhellou, P. Aknin, J-P Perrin. "Dedicated sensor and classifier of rail head defects for railway systems", Control Engineering Practice 7, 1999, pp.57-61.

[13] M. Bentoumi, P. Aknin, G. Bloch. "On-line defect diagnosis with eddy current probes and specific detection processings". EPJAP, Vol. 23, no.3, 2003, pp.227-233.

[14] P. Weber, P. Munteanu, L. Jouffe, "Dynamic Bayesian Networks modelling the dependability of systems with degradations and exogenous constraints", 11th IFAC Symposium on Information Control Problems in Manufacturing, Salvador, Brasil April 5-7th, 2004.