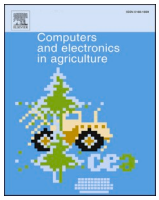
Computers and Electronics in Agriculture 200 (2022) 107167

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



A decision-support system to predict grape berry quality and wine potential for a Chenin vineyard

Nathalie Mejean Perrot [a,*], Alberto Tonda [a], Ilaria Brunetti [c], Hervé Guillemin [b], Bruno Perret [c], Etienne Goulet [d], Laurence Guerin [d], Daniel Picque [c]

[a] Unité MIA-Paris, AgroParisTech, INRAE, Université Paris-Saclay, France
[b] Unité URTAL, INRAE, Unité de recherches en technologie et analyses laitières. Poligny, France
[c] Unité SayFood-Grignon, AgroParisTech, INRAE, Unité de recherches en microbiologie et procédés alimentaires, Université Paris-Saclay, France
[d] IFV Institut Français de la Vigne et du Vin, Unité de VINs, Innovations, Itinéraires, TERroirs et Acteurs, Amboise, France

ARTICLE INFO

Keywords:
Chenin Vineyard
Decision Support System
Dynamic Bayesian Network
Fuzzy Logic
Grape Berry Ripening
Model Coupling
Wine Quality

ABSTRACT

Grape berry ripening is a complex process, and predicting the quality of wine starting from the ripening kinetics of grape berries is a challenging task. To tackle this problem, we present a decision-support system based on coupling expert know-how with probability laws encapsulated in a probabilistic model, a dynamic Bayesian network. The proposed approach predicts the ripening kinetics of grape berries starting from initial measurements and weather conditions, and then exploits the information to evaluate the potential of the wine that will produced from them. The results show that the dynamic Bayesian network predicts the total acidity concentration and the sugar content of the grape berries with a small amount of error (mean of 6% for total acidity concentration, 10% for sugar content) that is considered satisfying by the experts, making it possible to predict the ideal moment for harvesting the grapes up to two weeks in advance. Moreover, feeding the results from the probabilistic model to a fuzzy expert model, the predicted trajectories are compared to an ideal trajectory described by wine experts and formalized mathematically. From this comparison, it is possible to anticipate drifts in wine sensory quality right from the step of grape ripening.

1. Introduction

The ripening of grape berries is a complex process, including physicochemical and biochemical reactions. Still more complex is predicting wine quality from the grape berries maturation kinetics. Reactions during the ripening depend on multiple factors, with weather being the most influential, especially in the last weeks preceding the harvest. Since berries ripeness plays a major role in determining wine potentialities, correctly predicting the ripening process and determining the ideal harvesting date is a significant challenge for the wine industry. While the expert consensus is that choosing an appropriate moment for harvesting has a considerable impact on the final quality of the wine, the exact effect has still not been exactly quantified (Van Leeuwen, 2010). In this particular context, we define the wine potential of the grapes as their capacity of producing a wine of at least acceptable quality, provided that the rest of the processing is performed correctly. In other words, a good wine potential of the grapes is a necessary, but not sufficient condition for obtaining high-quality wine.

The ripeness of grape berries can be evaluated resorting to different measurable quantities, for example their sugar content, the color of their seeds, or their sensorial characteristics. Some of these quantities can be measured exactly resorting to chemical means, while others require an expert evaluation on a symbolic ordinated scale. In the last decade, new sensors have been developed to easily measure grape characteristics such as color, sugar content, and aromatic potentialities, (Ben Ghozlen et al., 2010), (Geraudie et al., 2010). Nevertheless, most of the analyses are in practice still carried out in laboratory, with time-consuming and expensive procedures required for a close monitoring of the grape berries ripeness and never used for on line prediction. Literature reports relatively few contributions attempting to characterize the link between the grape ripeness and a global indicator of wine quality. Interesting studies like (Niimi et al., 2021) have worked on parameters determined by preprocessing techniques for mid-infrared (MIR) spectra of grape berries to model sensory properties of different wines, using Partial Least Squares (PLS) regression models. But the problem is also the cost and time consuming of such approaches, considering the equipment needed

* Corresponding author at: UMR MIA, 16 rue Claude Bernard, F-75231 Paris Cedex 05, France.
E-mail address: nathalie.mejean@inrae.fr (N. Mejean Perrot).

https://doi.org/10.1016/j.compag.2022.107167
Received 3 February 2022; Received in revised form 16 June 2022; Accepted 20 June 2022
Available online 29 July 2022

to acquire the data. Some studies like (Jensen et al., 2008) have linked grape phenolic composition to wine phenolic composition and color of wines using multivariate analysis. Nevertheless it is in this case too specific enough to be used directly for decision help directly at the field.

On the other hand, the development of mathematical models to predict or characterize different mechanisms taking place during winemaking has been widely treated in literature. For example the tool « Epicure » developed by the French Vineyard Institute (IFV) for managing phytosanitary risks (Raynal, et al. 2010), or a work on the prediction of kinetics of fermentation in wine processes like (Goelzer et al., 2009). For the grape ripeness prediction, a model has been developed by (Baudrit, et al., 2015), (Perrot, et al., 2015) linking chemical indicators to weather conditions on Cabernet Franc grape berries. The approach presented in this paper is an extension of this latter work, with several important differences: (i) the target grape berries are for the Chenin wine, and this also reflects on the different structure of the dynamic Bayesian network; (ii) we propose in this paper a semi-automatic method to adapt the DBN algorithm and especially its discretization to a new type of wine (iii) in contrast with the approach outlined in (Perrot, et al., 2015), no expert system is used to integrate the data set, as the available quantity of data is larger in this case; (iv) for the first time, a link between variables related to grape berries ripening and wine quality is provided, thanks to a fuzzy system.

Few other works have been developed in this domain, except PLS approaches (Claverie, et al., 2008) to link climatic and pedologic variables to chemical indicators of ripeness. Nevertheless, those approaches used mathematical classifiers to link two different spaces: climatic and pedologic ones but are not models of prediction of the kinetics of the physicochemical compounds. An interesting study have been developed by (Petropoulos et al., 2017), based on the development of a fuzzy tool linking different physicochemical and sensory parameters of the grape berries to the wine quality. It was nevertheless developed for wine classification and not for decision support during harvesting.

In this work, we present a novel decision-support system able to predict 15 days before the maturation process, considering chemical indicators, by observing the weather conditions. The predicted indicators' values are then used to evaluate the wine potential, according to the winemaker goal. This model relies on probability laws encapsulated in a dynamic Bayesian network formalism learned on available data and a fuzzy expert model based on expertise.

2. Materials and methods

2.1. Experimental data

Experimental data are gathered by the IFV institute from Chenin vineyards located in the French region of "Vallee de la Loire" over several years, with weekly sample collections from July to September before grape harvesting season. The data range from 1989 to 2017 (more precisely 1989, 1995–2001, 2016–2017) with land plots distributed between two geographical places, "Anjou" and "Touraine" (see Fig. 1), for a total of 30 vineyards and between 2 and 5 points by kinetics for each vineyard according to the year of the experiment.

2.2. Variables of the model

The inputs used for the probabilistic model used in proposed approach are weather conditions (see Table 1): Temperature (°C), rainfall (mm) and relative humidity (%) were supplied by Meteo France meteorological stations located near and/or on the vineyards. Insolation (quantity of solar radiation, in hours) was provided by one meteorological station located at Montreuil-Bellay, in the middle of the geographical area of the study.

The outputs of the probabilistic models include both physicochemical and sensory measurements. The physicochemical measurements have been selected after a discussion with wine experts, as those considered essential: sugar content (s) measured in g/l, total acidity concentration (ac) in gH_2SO_4/l, and malic acid concentration (ac_m) in g/l, (Barbeau, 2003), (Riou, 1994). Their variations during a week (defined as the difference between data collected in two subsequent time points) are also considered: variation in sugar content (Var_s), variation in total acidity concentration (Var_ac), and variation in malic acid concentration (Var_ac_m). Each week, a lot of 200 berries of Chenin, with pedicels, were randomly selected from each vineyard according to the method of Vine and Wine French Institute (ITV-France) (Cayla et al., 2002) in order to limit the effects of the grape heterogeneity. With the set of 200 berries of each sampling, a crushing was realized with a blender, then the must was filtered through a Whatman paper filter. Sugars content (g/l) was measured with a refractometer; total acidity

Table 1
Weather conditions, used as inputs of the proposed approach. HR_{min} is the lowest air humidity observe during the week, HR_{max} is the highest air humidity observe during the week, T_{min} is the lowest air temperature observe during the week, T_{max} is the highest air temperature observe during the week Inputs of the model: meteorological conditions.

Variable	Description	Calculation	Unit
HR	Relative Humidity	$mean(\sum_{i=1}^{7} \frac{HRmin_i + HRmax_i}{2})$	%
T	Temperature	$\sum_{i=1}^{7} \frac{Tmin_i + Tmax_i}{2}$	°C
Pl	Rainfall	$\sum_{i=1}^{7} Pl_i$	mm
Ins	Insolation	$\sum_{i=1}^{7} Ins_i$	h

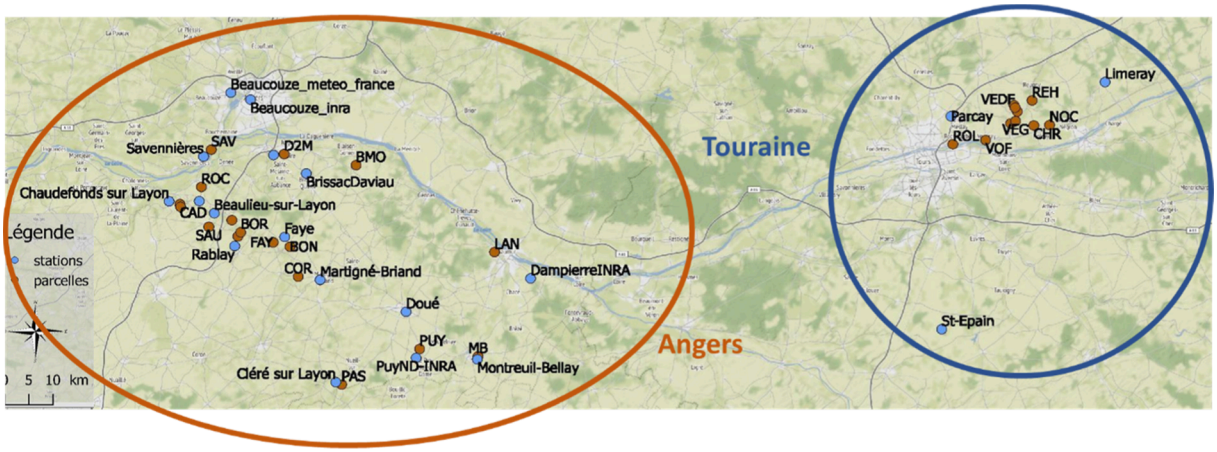


Fig. 1. The geographical distribution of the Chenin vineyards considered in this work.

concentration (g/l eq. H_2SO_4) was measured by the titration method and malic acid (g/l).

2.3. Expert knowledge

Two types of experts were interviewed: 3 scientists and 2 wine-growers working on the two areas considered in this study. Each expert was interviewed during one or two sessions (spanning 2–3 h each). Each of the elicitation sessions was attended by one expert and one or two interviewers. Three hours were allocated for each interview. Notes were taken and audiotapes were recorded. After each session, the tapes were re-played several times to make sure the notes were accurate and complete. To build the interview, adapted methods proposed by (Sicard et al., 2011) were applied. The elicitation process was based on a set of predetermined structured open-ended questions used to direct the interviews. Questions were designed according to techniques based on survey methods with the aim of optimizing the expression of expert knowledge. The objective was to ask clear and simple questions without ambiguities and that did not implicitly direct the expert towards a specific answer. Questions were also asked in such a way to encourage simulation of the expert's situation using the explicitation interview method developed by Vermersh, 2006 (i.e. cognitive interviews, (Moody et al., 1996)). We paid particular attention to context reinstatement as recommended in these methods. This involves having the expert think about and describe their feelings during the episodes being recalled.

Previous works (Perrot et al., 2015) had access to a larger number of experts, but it is important to remark that in the past the aim was to integrate the available data on grape berries ripening with an expert system, while in the current approach the objective is to link grape berries ripening kinetics to wine quality, creating a completely novel model. For this particular task, the availability of domain experts is relatively scarcer.

2.4. Mathematical formulation

The predictive model used in this work is a coupling between two models: a probabilistic graphical model, dynamic Bayesian networks in particular, and a fuzzy expert model. The first one is relevant to formalize mathematically, the implicit information, contained in the available experimental data. With the second one, it is possible to interact with the wine experts of Chenin vineyards and ultimately formalize the explicit knowledge, encoded through expert interviews and know-how.

2.4.1. Introduction to the modeling approaches

The first model used in our approach is a Dynamic Bayesian Network (DBN), a probabilistic graphical model able to describe phenomena developing over time (Jensen and Nielsen, 2010; Pearl, 1988). The structure of a DBN is an oriented graph, representing correlations between variables, which in our case was created by interacting with human experts of the Chenin wine. Once the structure of a DBN is fixed, it is then possible to compute its parameters starting from a training dataset: the parameters are conditional probability tables, assessing the probability for variables taking a specific value, knowing the values of the variables they depend on. For our specific application, the values of the variables need to be discretized. Differently from a classical Bayesian Network, a DBN makes it possible to estimate variable values over several subsequent time steps. In our case, each time step is equivalent to one week in the grapes ripening processes. DBNs have been successfully adopted for several agri-food applications (Baudrit, et al., 2015) (Perrot, et al., 2015).

Coupled to this first model, a fuzzy expert model is set up to mathematically describe the explicit knowledge of the experts concerning the complex link between the maturation of the grape berries and the wine potential.

2.4.2. The DBN algorithm

More formally, a DBN is a graph-based model of a joint multivariate probability distribution, capturing properties of conditional independence between variables. Like a BN, a DBN is a directed acyclical graphs (DAG) where the nodes represent variables, and the missing arcs represent conditional independences between variables. In DBNs in particular, nodes $X(t) = (X_1(t), \cdots, X_n(t))$, represent n discrete random variables, indexed by time t, providing a compact representation of joint probability distribution P for a finite time interval $[1, \tau]$. In other words, the joint probability P can be written as the product of the local probability distribution of each node and its parents, as follows, Equation (1):

$$P(X(1), \cdots, X(\tau)) = \prod_{i=1}^{n} \prod_{t=1}^{\tau} P(X_i(t)|U_i(t)) \quad (1)$$

Where $U_i(.)$ denotes the set of all parents of node $X_i(.)$, and $P(X_i(.)|U_i())$ describes the conditional probability function associated with random variable $X_i(.)$ given the values of $U_i(.)$. $X_i(t)$ is termed "slice", and it represents the set of all variables at time t. This factorization of the joint probability distribution, based on information from the graph, makes it possible to straightforwardly represent large models, and use them for practical applications. In other words, DBNs represent the beliefs of possible trajectories of the variables involved in a dynamic process.

In order to make the problem treatable, DBNs assume the first-order Markov property: the parents of a variable in time slice t must appear in either slice $t-1$ or t. As a consequence, for the first-order homogeneous Markov property, the conditional probabilities are time-invariant, meaning that $P(U(t)) = P(U(2)) \forall t \in (1, \tau)$. In order to fully specify a DBN, we will then need to define the intra-slice topology (within a time slice), the inter-slice topology (between two time slices), as well as the parameters (i.e. conditional probability functions) just for the first two time slices. The structure of a model can be explicitly built on the basis of knowledge available in the literature and parameters can be automatically learned without a priori knowledge on the basis of a dataset, a process termed parameter learning. The techniques for learning DBNs are generally extensions of the techniques for learning BNs. Specialized literature reports several methods to learn the structure or the parameters of a DBN from substantial and/or incomplete data (Geiger & Heckerman, 1997); (Heckerman, 1999). In our work, the topology of the graph is obtained from expert knowledge; for parameter learning, we consider the simplest and most commonly adopted methodology, simply evaluating the co-occurrence rate of values of variables in the training data.

Once a DBN is fully specified, it can be used to estimate marginal probabilities for target variables, through a process also known as Bayesian inference (Equation (2)):

$$P(X(t)/O(t') = o(t')), \forall t' \in [1, \tau] \quad (2)$$

Where X is a set of variables whose values we are interested in predicting, and O is a set of variables whose values are known (for example, in food processing X might be the variables representing the physico-chemical properties of a product and O might be the variables representing the observed environmental conditions). In general, given a way of calculating $P(X(t)|O(t'))$ from the knowledge of $P(X(t')|O(t))$, inference in a DBN is performed using recursive operators and Bayes' theorem, updating the belief state of the DBN as new observations become (Murphy, 2002).

2.4.3. The fuzzy expert algorithm

The second model used in our approach comes from the theory of fuzzy sets. Fuzzy logic was originally proposed by (Zadeh, 1965), and it is an extension of set theory by the replacement of the characteristic function of a set by a membership function whose values range between 0 and 1. Soft transitions between sets are thus obtained and make it possible to represent gradual concepts, as well as the representation and

the inference of linguistic rules stemming from expertise. This type of formalism is particularly adapted for taking human linguistic and reasoning processes into account (Perrot et al., 2006). Fuzzy models can be written in an easy form to understand for an expert, ie linguistic rules. Similarly, an essential fuzzy notion is the fuzzy membership function. A fuzzy set E in universe of discourse U can be defined by Equation (3):

$$\begin{aligned} &E = \{(u, \mu_E(u))u \in E\} \\ &\mu_E{:}U \rightarrow [0,1] \end{aligned} \tag{3}$$

μ_E is thus the membership function of set E, and it represents the set of membership grades ($\mu_E(u)$) of a numerical variable u mapped to a fuzzy set E. This function makes it possible to link real numerical variables to a given linguistic variable. The value of the membership grade is a real number within the interval [0;1], expressing the translation from one space X to another space Y. It was set up with the experts. This notion gives the way to link a numeric variable to a linguistic variable often manipulated by the operators. In fact, fuzzy memberships are used to describe how much an object belongs to a linguistic notion. Going back to an example:

Suppose a value of sugar of 195 g/l belonging to the symbol "not target" with a membership degree of 0.5 and to the symbol "target" with a membership degree of 0.5. It would mean that the maturation will be mitigated for this value of sugar.

$$\mu(x) = \begin{cases} 0 & (x < a_1) \\ \dfrac{x - a_1}{a_2 - a_1} & (a_1 \leqslant x < a_2) \\ \dfrac{a_3 - x}{a_3 - a_2} & (a_2 \leqslant x < a_3) \\ 0 & (a_3 \leqslant x) \end{cases} \tag{4}$$

Membership functions can be expressed through various representations. The representations most widely used are triangular (Equation (4)) for a given triplet series of parameters a_1, a_2, a_3.

On the basis of the definition of the fuzzy subsets, the fuzzy Tnorm, is used in this paper to aggregate the information associating the three input variables of the fuzzy algorithm: the sugar content, the total acidity concentration and the malic acid concentration, to the output: the wine potential. Those input variables are joined by a connector "AND" (a classical mathematical logical interpretation of the join) using the fuzzy Tnorm (Equation (5)).

$$\mathrm{Tnorm}(\mu 0, \mu 1, \cdots, \mu n) = \prod \mu i \tag{5}$$

Where μ_i are the input variables (three in our case study: sugar content, total acidity and malic acid concentrations).

2.4.4. Models evaluation

The DBN previously introduced will be evaluated with a leave-one-out cross-validation (LOOCV), where the model is repeatedly trained on the whole dataset, minus one sample, and the remaining sample is used for testing. The procedure is repeated until each sample has been used for testing. Considering the mean and standard deviation on the results of a LOOCV provides a better estimate of the model's capabilities than just considering a random split of the available data between a training set and a test set (Geisser, 1993).

For the choices made in this study, before training the model, it is necessary to discretize the real-valued variables in the dataset (see subsection 3.1.1). However, in order to evaluate the performance of the model's predictions against the ground truth, the results of the model will have to be converted back into real values. Recalling that the predictions of a DBN model for variable x will consist in a series of probabilities P_i for each possible discrete class $i = 0, 1, \cdots, n_x$ associated with variable x, the predicted outcome can be converted to a real value using the following Equation [6]:

$$x_i{}^{predicted} = \sum_{i=1}^{n_x} \overline{x_i} P_i \tag{6}$$

Where $\overline{x_i}$ is the average value of all samples of variable x that fall under class i.

The first metric used to evaluate the quality of the predictions against the ground truth is the root mean squared error (RMSE), Equation [7]:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i^{predicted} - x_i^{observed}) 2} \tag{7}$$

Where N is the number of predictions considered for target variable $\times$, and $x^{observed}$ indicates its observed value. In this study, we will also use the relative RMSE (RRMSE) that expresses the RMSE as a percentage of the range of observed values for the target variable, and it is thus more informative as an error metric (Equation (8)):

$$RRMSE = \frac{RMSE}{(x_{max}^{observed}) - (x_{min}^{observed})} \times 100 \tag{8}$$

Where $x_{max}^{observed}$ and $x_{min}^{observed}$ are the highest and lowest values observed for variable $\times$, respectively.

For the fuzzy algorithm, the results are evaluated as classification accuracy in percentage, expressing the number of points classified correctly within the range of precision fixed by the experts, over the total amount of points classified.

3. Results

The results presented in this section include both the numerical assessment of the models' predictions, and the trained models themselves. The models are considered an output of this work, as they will be exploited by experts in the field to predict the best moment for harvesting grapes, once they have reached the proper degree of ripeness.

3.1. Description of the Chenin vineyard global model

The Chenin vineyard model is based on a coupling between a DBN expressing the kinetics of ripening of the Chenin grape berries and a Fuzzy expert model expressing the expert knowledge on empirical laws linking kinetics of maturation and wine sensory potentialities. We first built a Chenin Dynamic Bayesian Network (Chenin_DBN), which makes it possible to obtain reliable dynamic predictions of sugar content, total acidity, and malic acid concentrations by measuring air temperature, rainfall, relative humidity, and insolation hours in the three weeks preceding the harvest.

Once we predict the maturation indicators, we evaluate the wine potentialities according to the winemaker expectation, by means of a Chenin Fuzzy expert model (Chenin_FEM), linking the maturity indicators to global wine quality trajectories (see Fig. 2).

3.2. The Chenin_DBN model

The three key elements of the proposed Chenin_DBN model are: (i) its structure, defining the relationships between the problem's variables, (ii) the choice of discretization for each variable, (iii) the parameters, expressed by conditional probability tables (CPTs), which describe how the probabilities of variables assuming a given value change, depending on the values of the variables they depend on. The structure of the network has been defined on the basis of expert knowledge, building upon previous works on different wines that led to the development of the software PREVIMAT (Brousset, 2009; Baudrit et al., 2015).

3.2.1. Network structure of the Chenin_DBN

The first part of the model developed, inspired by previous work on Cabernet-Franc and Gamay wines (Baudrit et al., 2015) predicts

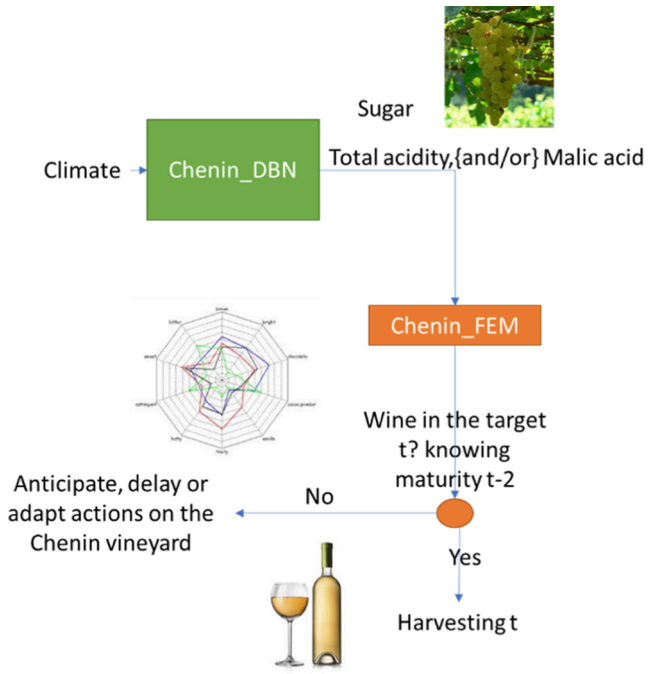


Fig. 2. The Chenin vineyard model developed in this work consists of two coupled models: one for the prediction of the physicochemical variables using a DBN (Chenin_DBN), and another based on a Fuzzy expert theory (Chenin_FEM) dedicated to the prediction of the potentialities of the wine for being in the target required by the winemaker.

physicochemical indicators starting from weather conditions (Fig. 3). For each physicochemical ripeness indicator, the climatic variables playing a key role in their kinetics are selected from expert knowledge and literature. In particular, relative humidity only affects the two acidities (total acidity concentration and malic acid concentration), sunshine influences sugar content, while temperature and rainfall have an impact on the three variables considered: sugar content (s), total acidity concentration (ac), malic acid concentration (ac_m).

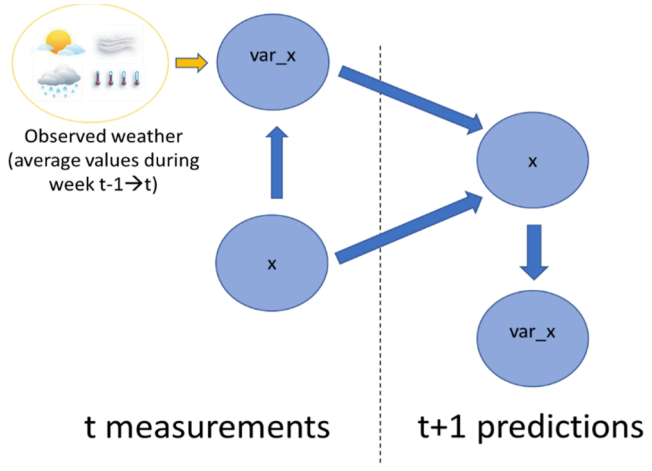


Fig. 3. Representation of the Chenin_DBN, in the form of two generic slices that can be unrolled on several slices representing the different step times. DBNs assume the first-order Markov property, which means that the parents of a variable in time slice t must occur in other slices and the conditional probabilities are time-invariant. The slice representing the time t (t measurements) is concerned at the beginning of the iterations by variables that are measured at time t0. The consecutive slice: time t + 1 is dedicated to predictions. If several slices are added, for example t, t + 1 and t + 2, it starts at t0 with an initialization where variables are measured, followed by two slices predicted t + 1 and t + 2, with t + 2 predicted on the basis of the prediction of t + 1.

As the Chenin_DBN needs to be able to capture dynamical variations of the values over time, to better predict the three variables it is necessary to define new intermediate state variables. In particular, having collected the data related to the last two to five weeks (according to the years of experiment) before the harvest, we can make the assumption that the trajectory in time of each variable is stable and its variation is constant, given identical meteorological conditions. In other words, we consider that a month before the harvest, only alterations in the weather can cause a significant deviation from an established trajectory in time.

More formally, considering each physicochemical variable $x \in$ {ac, ac_m, s} at time t and $t+1$:

$$x(t+1) = var\_x(t+1) + x(t)$$

And consequently.

$$var\_x(t) = x(t) - x(t-1)$$

As already mentioned, the (absolute) value of a variable can be used as an indicator of the current stage of ripening, while its variation, as a function of the climatic variables, will dictate the ripeness trajectory.

Having taken into account the variation of the physicochemical variables over time and the limitation for certain years to two weeks of history, the Chenin_DBN can now be structured over the minimum time steps history of the whole kinetics, which is three time steps $t = \{0,1,2\}$, each one spanning a week before the time of the harvest. At time $t = 0$, the value of each variable is known; for the next two time steps, only the climatic variables are known (observed), while the physicochemical quantities and their variations are predicted by the model. The complete structure of the Chenin_DBN for the physicochemical variables, when considering all variables for the three time steps, is presented in Fig. 4.

3.2.2. Variables discretization

As previously described, to create the CPTs of our Chenin_DBN model, it is necessary to define the discretization of the continuous variables in the problem. In this context, discretizing variable x amounts to finding several intervals $\{[x_1,x_2),[x_2,x_3),\cdots,[x_{n-1},x_n)\}$ of continuous values such that $x_1 < x_2 < \cdots < x_n$, with each interval corresponding to a discrete class.

The definition of such intervals has a considerable impact on the performance of the Chenin_DBN model, so this step is crucial for generalized the approach and to obtain satisfying predictions. Specialized literature reports different established partial solutions for discretization: for example, considering quantiles, nested averages, and amplitudes (Dougherty et al., 1995).

In our case we opted for test the latter solution, since it makes it possible to create intervals of similar amplitude for each category of variables in the problem: climatic variables, grape sensory variables, and variables indicating instantaneous physicochemical quantities, or variations of physicochemical quantities from a time step to the next. The number of classes was chosen by selecting a limited but sufficiently representative number of intervals, following the advice and recommendations of human experts.

For the climatic variables, the following intervals were defined (see Table 1 for the detailed description of the variables):

- $Ins = [[15,30],[30,40],[40,55],[55,60],[60,75]]$
- $Pl = [[0,10],[10,20],[20,30],[30,45],[45,70],[70,100]]$
- $T = [[0,11],[11,15],[15,17],[17,19.5],[19.5,22]]$
- $HR = [[60,70],[70,75],[75,80],[80,90],[90,100]]$

For the physicochemical variables, an interactive semi-automated discretization approach was developed, based on the notion of co-occurrence between variable values and their variations. The methodology is based on a visualization software, EvoGraphDice, coupled with

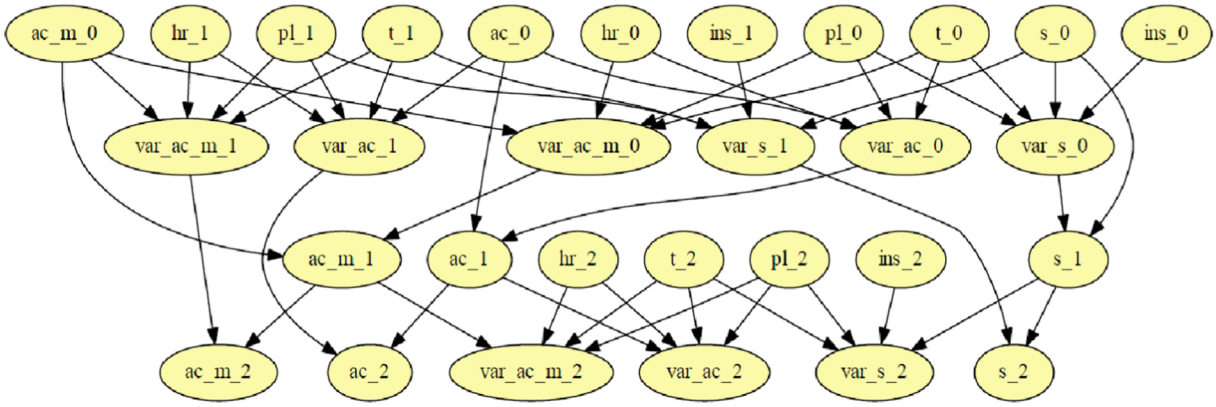


Fig. 4. The unrolled Chenin_DBN model over time steps $t = \{0,1,2\}$. A physicochemical variable $(t+1)$, considered at time step $t+1$ is predicted as a function of its value at the previous time step t and its variation $var\_x(t) = x(t) - x(t-1)$. In turn, $var\_x(t)$ is predicted as a function of the climatic variables at time steps $t-1$ and t. It is interesting to notice that the values of the physicochemical variables are not influenced by each other.

an evolutionary optimization approach (Boukhelifa et al., 2017). Variations of the variable (for example var_s for the sugar content) are fixed by the experts. Diagnostic and decisions of the experts are indeed more based on the physicochemical variations during one week than on the values themselves. It is then easier for them to describe the values in terms of variations. For example, the variation var_s of the variable describing sugar content from one week to the next is fixed on the basis of the expert description. The optimal discretization of the variable itself like the sugar content for example, is then calculated by optimization, to ensure a repartition of the var_s classes of interval as homogeneous as possible, for each sugar interval in the data. Several iterations are performed to reach a good compromise between the values of discretization proposed by the algorithm of optimization and what is considered to be coherent according to the expert evaluation. Experts visualize the results of the optimization at each iteration, and validate or reject the result.

An example of discretization obtained for sugar content is presented in Fig. 5. We can see that classes of s are created by the optimization algorithm after 5 runs of interaction with an expert, with a good result in terms of repartition of the different classes of Var_s for each class of s. This homogeneous repartition is required for a good learning of the probability laws in the DBN model structure.

All the results of the optimization and thus the discretization proposed for the physicochemical variables are presented in Table 2.

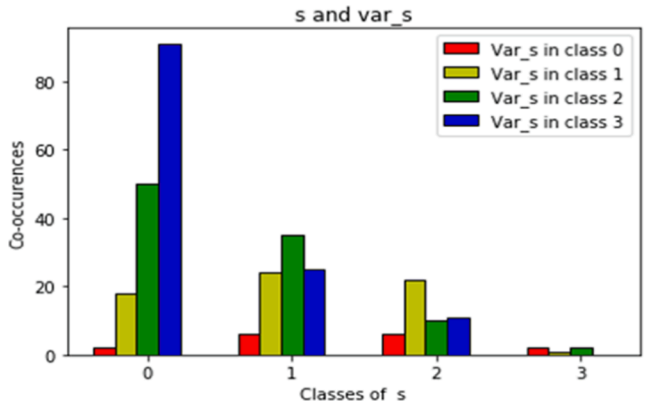


Fig. 5. Result of the discretization of the variable s (classes equivalence: see Table 2) after 5 iterations and expert's interactions, var_s being fixed by the expert. The automatic discretization was used for values ranging from 0 to 210 g/l, generating four classes. Points with values above 210 g/l are rare, but rather than assigning them all to the same class as the algorithm proposed, the experts decided to divide the space evenly in three more classes with an amplitude of 10 g/l, plus one final class for all values above 240 g/l.

Table 2
Discretization of the physicochemical variables, fixed by experts for the variation var_X of variable X, obtained by the experts through interactive optimization.

	Discretization of X	Discretization of Var_X
S (Sugar)	● Class 0 = [∞,156.9] Class 1 = [156.9,182.86] Class 2 = [182.86,201.8] Class 3 = [201.8,210] Class 4 = [210,220] Class 5 = [220,230] Class 6 = [230,240] Class 7 = [240,+∞]	● Class 0 =[0,12], Class 1 =[12,20] Class 2 =[20,35] Class 3 =[35, +∞]
ac (total acidity concentration)	● Class 0 =[− ∞, 5.47] Class 1 =[5.47, 6.33] Class 2 =[33,7.94] Class 3 =[7.94, + ∞]	● Class 0 =[−∞, −1.5] Class 1 =[−1.5, −1] Class 2 =[−1, −0.6] Class 3 =[−0.6, 0]
ac_m (malic acid concentration)	● Class 0 =[−∞, 3.66] Class 1 =[3.66, 4.6] Class 2 =[4.6, 5.68] Class 3 =[5.68, 6.88] Class 4 =[6.88, + ∞]	● Class 0 =[−∞, −2.5] Class 1 =[−2.5, −1.5] Class 2 =[−1.5, −0.75] Class 3 =[−0.75, −0.5] Class 4 =[−0.5, 0]

3.2.3. Network parameters

Once the network structure has been defined, and the variables have been properly discretized, learning the parameters of the Chenin_DBN is a straightforward process. Each node represents a conditional probability table, describing the different probabilities for variable × to assume values $x_1, x_2, \ldots, x_N$ given the values of the parent variables it depends on, $pa(x)$. By reading the training data, the frequency of appearance of values for × together with the values for $pa(x)$ for the same samples, can be directly used as a probability to fill the conditional probability tables, using the classic strategy of maximum likelihood (Redner and Walker, 1984).

3.3. The Chenin fuzzy expert model (Chenin_FEM)

The Chenin_FEM is constituted of a set of membership functions adapted to the Chenin characteristics of variation, and rules of aggregation leading to a quantification of the output of the model.

3.3.1. Membership functions

The membership functions (see Section 2) set up in the Chenin_FEM are presented in Fig. 6 for the three inputs considered: sugar content, total acidity concentration and malic acid concentration.

3.3.2. Rules of aggregation

Two experts were interviewed in this study, to find a link between the ripeness of grape berries and their personal assessment of wine quality. Experts encode their knowledge as a symbolic trajectory towards ideal wine quality, and assess differences in terms of drift from this trajectory. The experts classify grape berry ripeness into three categories (that we named Class 1, 2, and 3, see Table 3), in link with three steps of ripening that in their opinion should appear sequentially at given times during the ripening process. In case of drift from the perceived ideal category, experts usually perform their corrective actions after the harvesting, either by mixing grapes from different batches, or changing the parameters of the wine fermentation. Nevertheless, each expert has their own rules for managing this situation. In this work, we decided to use the rules of the expert with the longest track record in wine production.

The rules used by the experts for the aggregation of the three terms are presented in Table 3 and processed according to the fuzzy Tnorm methodology presented in Section 2.

4. Experimental results

This section describes the experimental results obtained by comparing the complete models developed in Section 3 against the available data.

4.1. Chenin_DBN model predictions for the physicochemical variables

A LOOCV is performed on the dataset. At each iteration the network is trained on the whole dataset minus one sample, and then tested on that sample. We obtained a mean RRMSE for each predicted variable.

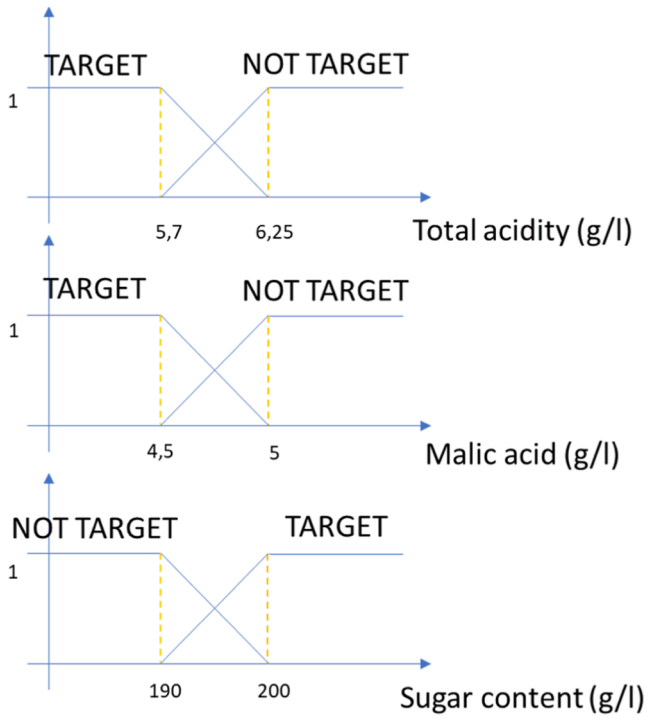


Fig. 6. Membership function and associated targets for the three variables manipulated by the expert to infer the quality of the Chenin vineyard: total acidity concentration, malic acid and sugar content.

Table 3
Expert rules of aggregation for the three input variables of the Chenin vineyard: sugar content (g/l), total acidity concentration (g/l) and malic acid concentration (g/l). The output is the projection of the class of wine inferred by the aggregation, knowing the values of the input variables: Class 3 represents the quality of the berries required to obtain wine that is at the very least of acceptable quality according to the expert. Class 2 means that the maturity of this vineyard plot is not enough to reach an acceptable wine quality, but it can potentially be still corrected by pursuing the ripening of the grapes, postponing the date of the harvest if possible; or, if not possible, by adapting the fermentation process or mixing the berries of different vineyard plots. Class 1 means that the maturity of the berries have, at the time of measurement, not the good potential to produce a good wine. It is generally associated with berries at the very beginning of the ripening step.

Sugar	TARGET		NOT TARGET	
	Malic acid TARGET	Malic acid NOT TARGET	Malic acid TARGET	Malic acid NOT TARGET
Total acidity concentration				
TARGET	3	2	2	Non existant
NOT TARGET	1	1	1	1

The results (Table 4) show that it is possible to predict with good results the total acidity concentration and the sugar content in a range that is satisfying for the experts (10% mean error for sugar, 6% mean error for the total acidity concentration) and so anticipate the maturation date up to 15 days in advance. For the malic acid, it seems to be more complex to have a good prediction with the only variables considered as inputs of the DBN. Other physicochemical variables, that are harder to measure and were not included in this study, such as characteristics of the soil, would probably beneficial towards the aim of obtaining better predictions.

In Fig. 7, the model's predictions of the physicochemical data are compared with the observed values, for each variable and for the two considered time steps 1 and 2. Even if the R^2 value are moderately satisfactory, varying from 0.702 for the sugar to 0.53 for the malic acid, with cumulative errors for two time steps, scatterplots are relatively well aligned around the bisector, and most of the errors of the predicted points are within the range of uncertainty of the measurements defined by the experts. Essentially, 5 vineyards out of 30 were out of range for the predictions and were not considered for the computation of the correlation coefficients. It is a possible effect of the lack of data on those land plots with particular behaviors and an increasing uncertainty. As described above, the proposed approach faces more difficulties for the malic acid variable, but the results are still acceptable.

4.2. Validation of the Chenin_FEM

The Chenin_FEM model is tested in two separate experiments: Firstly, the predictions of the fuzzy expert module are tested with inputs observed in the data; Secondly, the outputs of the DBN presented above are used as inputs of the expert system. In this last case, the uncertainties

Table 4
Results of prediction for the three variables for two time steps: one week (1) or two weeks (2) in advance.

Variable in the unrolled DBN model	RMSE	RRMSE % ac = [3.4,12.5]; ac_m = [1.7,10]; s = [144,271.8]	R2
ac_1 (g/l)	0.536	6	0.668
ac_2 (g/l)	0.648	7	0.583
ac_m_1 (g/l)	0.825	9	0.6673
ac_m_2 (g/l)	0.867	10	0.5302
s_1 (g /l)	11.37	8	0.702
s_2 (g/l)	12.87	10	0.67

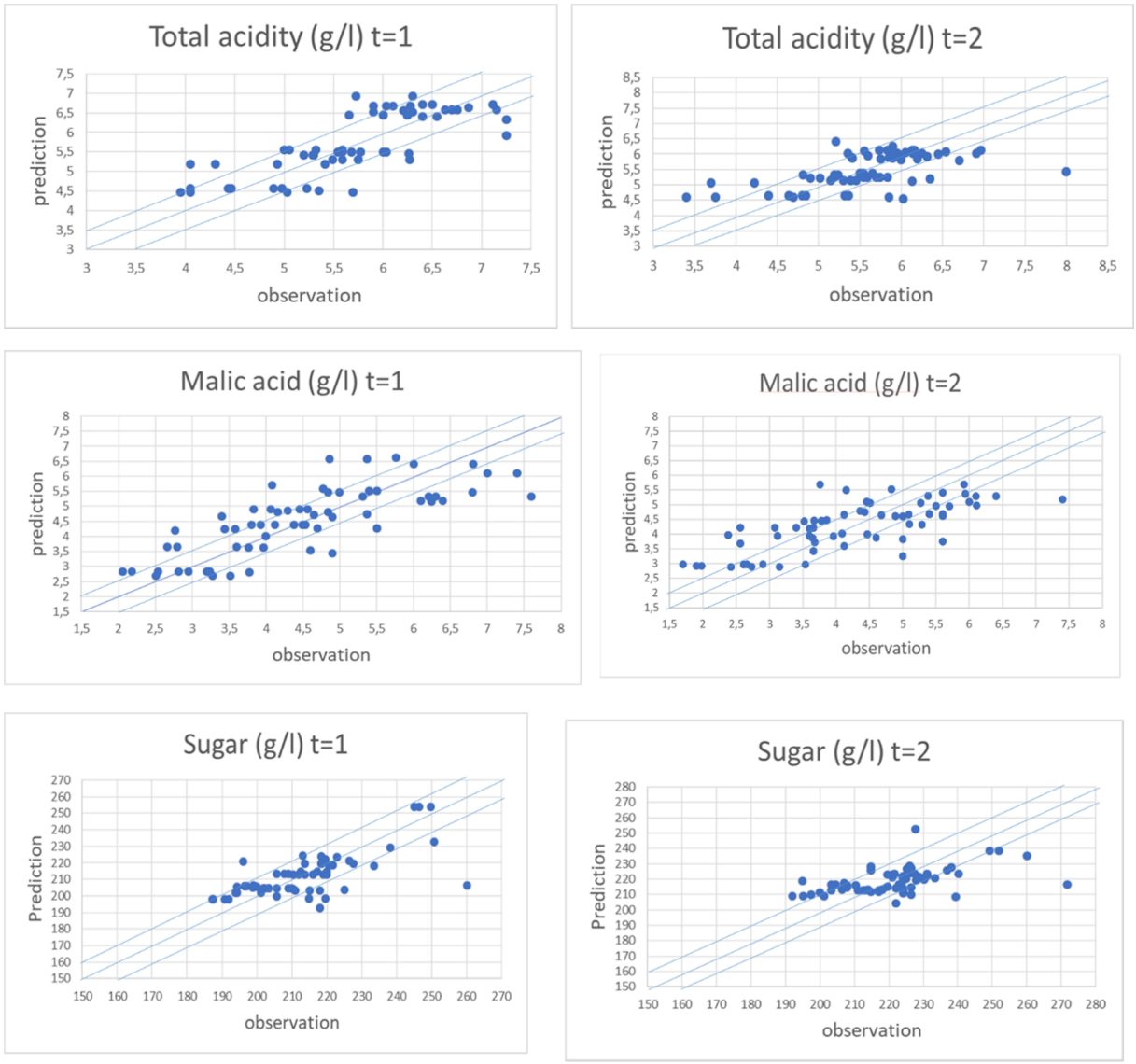


Fig. 7. Scatter plots for the predictions of the physicochemical variables by the Chenin_DBN model, for the two time steps t = 1 and t = 2. The error range considered acceptable by the experts for those variables is represented by the dotted lines in the plots ($\pm 0.5 g/l$ for the total acidity concentration and malic acid concentration, and $\pm 10 g/l$ for the sugar concentration).

pertaining the prediction of the two models are cumulated.

The proposed approach is tested on twenty harvests performed in different batches, equivalent to twenty samples in the database, are tested from 2016 and 2017, at different time steps of maturation, and evaluated by a Chenin sensory expert.

The results are presented in Fig. 8 for the predictions of Chenin_FEM based (1) on measured input data and (2) on the Chenin_DBN predictions. The inputs are the physicochemical values measured or predicted on the berries batch just before the grape harvest: the sugar concentration, the total acidity concentration and the malic acid concentration. The output is the prediction of the wine quality, considering those inputs. Inputs are data measured just before the harvesting of the batch used to produce the wine that is later evaluated by the expert, or classified by the Chenin_FEM. As the output of the Chenin_FEM is a fuzzy value rather than a class, we consider the classification correct if the fuzzy value is within 1.0 of the class assigned by the expert, a classical sensory threshold used by the experts in this experiment. Using this metric, the classification accuracy is 75% for predictions based on measured data, and 60% for respectively predictions based data predicted by the Chenin_DBN model. This result is considered acceptable, as predictions in this uncertain context, trying to establish a global link between grape berries physicochemical measurements and wine quality, are challenging. Errors of classification based on measured data appear more often for Class 2 predictions. Those for classification based on the Chenin_DBN predictions appear more often for Class 1, except one for Class 3. This second class would probably need to be redefined and optimized, if more data were available. The results of classification using data predicted by the Chenin_DBN model are lower than the previous one, which can be explained by a cumulative effect of the error of the two models (Chenin_FEM and Chenin_DBN). It is particularly true for predictions of Class 1.

4.3. Coupling the Chenin_DBN model to the Chenin_FEM: Towards a decision support system for wine quality prediction

Coupling the two models has a strong interest for the winemakers, as together they can be used to take a decision on the best date for harvesting grapes. It is also possible for them to intervene on the vineyards in the earlier steps of ripening to correct the drift from the ideal trajectory. Moreover, the experts' ability to predict the wine potential of

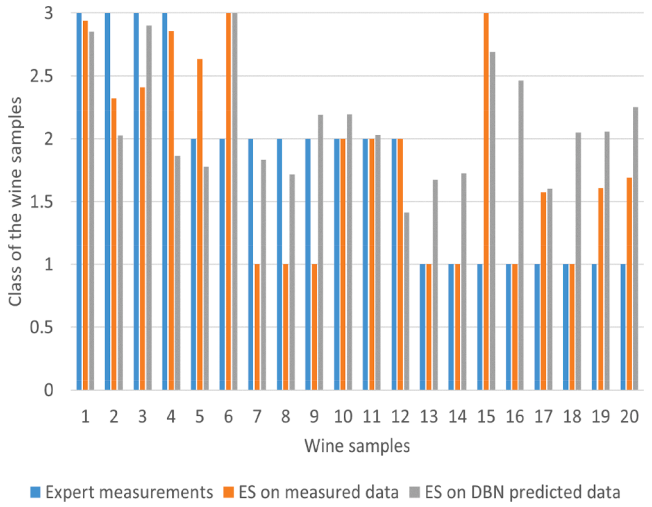


Fig. 8. Predictions of the classes of wine for 20 samples by the Chenin_FEM. The expert evaluations are in blue, the classifications based on inputs measured on the samples are in orange, and the classifications based on physicochemical variable values predicted by the Chenin_DBN are in grey. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the grapes allows them to manage their action directly on the plants or during the fermentation, with a possible prediction of the quality of the wine. Their reasoning in this uncertain space is achieved in terms of deviations from an "ideal" trajectory that they have in mind. This type of reasoning has already been observed in other traditional food production processes (Sicard et al., 2011). We have elicited with the expert this "ideal trajectory" and have compared it to the trajectories of three samples from the available database. Fixed fermentation process conditions were applied. The trajectories of those three batches are shown in Fig. 9, and compared to the ideal expert trajectory. Thus, we can see that the deviation of the batch REH66 is growing all along the weeks, with a maximum of a 2-class difference in week 6. It is important to notice that for this batch, the first measurements were only available at week 4, but previous weeks are very likely classified in Class 1. As predicted, after processing, the grapes of REH66 produced a wine that was judged too acid and not aromatic enough by experts, so unsuitable to be a Chenin wine. In this specific case, corrections could have been applied either before the step of grape ripening or after, by mixing the grapes with other batches. For the batch NOC52, the trajectory is almost the ideal one, but a drift is observed during the last week of the ripening. This was due to unexpected extreme climatic conditions during the week. In this case, the corrective action is generally performed during the fermentation, and the wine potential is recoverable: in this case, the wine resulting from the NOC52 grapes was considered of acceptable quality. Lastly, for the batch CAD57, we can observe that this trajectory is quite the same as the one desired by the expert, and predictably, after fermentation has led to a wine having the ideal characteristics required by the experts for Chenin.

5. Conclusions

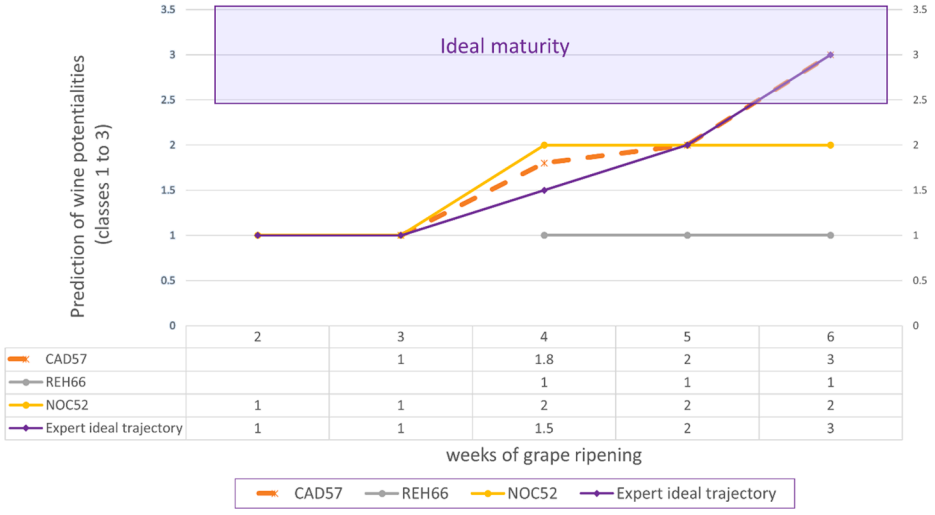
This study shows that it is possible and of valuable interest to propose computing tools able to formalize and capitalize on the knowledge of a food domain, based on data as well as human know-how and expertise. Even if for these complex processes the uncertainty is high, and amount of data is relatively small to describe each ripening kinetics, this combination makes it possible to develop relevant decision support systems based on artificial intelligence.

Interestingly, when compared to (Perrot et al., 2015), the accuracy of the predictive model for acidity is noticeably lower. While the previous work was carried out on a different quality of grapes, and considering less variables, the variability in weather for the data in the 2016–2017 period might also have played a role in the lower performance of the model presented.

Part of the difficulties in obtaining reliable predictive models is likely due to the growing impact of climate change on local weather. Future works will investigate the possibility of integrating climate-weather models into the proposed approach, in order to better take into account the evolution of this global phenomenon.

CRediT authorship contribution statement

Nathalie Mejean: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Alberto Tonda:** Conceptualization, Methodology, Writing – review & editing. **Ilaria Brunetti:** Software, Validation, Investigation, Writing – original draft. **Hervé Guillemin:** Resources, Formal analysis. **Bruno Perret:** Resources, Project administration, Funding acquisition. **Etienne Goulet:** Resources, Project administration, Funding acquisition. **Laurence Guerin:** Resources, Project administration, Funding acquisition. **Daniel Picque:** Resources, Project administration, Funding acquisition.



	2	3	4	5	6
CAD57		1	1.8	2	3
REH66			1	1	1
NOC52	1	1	2	2	2
Expert ideal trajectory	1	1	1.5	2	3

Fig. 9. Ripening trajectories of 3 batches (labeled CAD57, REH66, NOC52) expressed under the form of a prediction by the fuzzy algorithm of classes of potential towards wine. Physicochemical measurements are obtained since the beginning of the maturation step. Predictions of the fuzzy set of expert rules are calculated on these measurements. Measurements and predictions start at week 4 for the batch REH66, because the ripening process of those vineyard plots was judged to be late. Week 6 is the one just before the harvest. The wines produced from these batches were also evaluated by an expert of Chenin. For CAD57 and NOC52 the wine was evaluated to be at a good sensory quality type for a Chenin, better for CAD57 with more aromatic potentialities and a good equilibrium between aroma and acidity than NOC52. For REH66, the wine was evaluated not to be at the level of quality attempted for a Chenin, with too much acidity and not enough aromatic liveliness.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank all the experts and winegrowers of Tour and Angers who have participated to this study. We also thank InterLoire, the Institut Français de la Vigne et du vin – IFV Tours, the Institut Français de la Vigne et du vin – IFV Angers, the Chambre d'Agriculture d'Indre et Loire – Groupement de Développement Viti-Vinicole (GDVV), the Cellule Terroir Viticole, the INRA Angers – Unité Vigne et Vin, the ESA Angers – Laboratoire.

References

Barbeau, G.-B.-H., 2003. Comportement de quatre cépages rouges du Val de Loire en fonction des variables climatiques. J. Int. Sci. Vigne. Vin. 38, 35–40.

Baudrit, C., Perrot, N., Brousset, J.M., Abbal, P., Guillemin, H., Perret, B., Goulet, E., Guerin, L., Barbeau, G., Picque, D., 2015. A probabilistic graphical model for describing the grape berry maturity. Computers and Electronics in Agriculture. Comput. Electron. Agric. 118, 124–135.

Ben Ghozlen, N., Moise, N., Latouche, G., Martinon, V., Mercier, L., Besancon, E., Cerovic, Z., 2010. Assessment of grapevine maturity using a new portable sensor: Non-destructive quantification of anthocyanins. J. des Sciences de la Vigne et du Vin 44, 1–8.

Boukhelifa, N., Tonda, A., Trelea I.-C., Perrot, N., & Lutton, E. (2017). Interactive Knowledge Integration in Modelling for Food Sustainability: Challenges and Prospects. *ACM CHI Workshop on Designing Sustainable Food Systems.*

Brousset, J., 2009. Caractérisation multifactorielle et modélisation de la maturité de baies de Cabernet Franc en moyenne vallée de Loire. Rapport InterLoire.

Claverie, M., Prud'Homme, P., Mongendre, J., Zabollone, E., Raynal, M., Coulon, T., Forget, D. (2008). Modélisation statistique de la qualité en viticulture par la méthode PLS Spline. *VIIe Congrès International des terroirs viticoles.*

Dougherty, J., Kohavi, R., Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In: Machine learning proceedings 1995, pp. 194–202.

Geiger, D., Heckerman, D., 1997. A characterization of the Dirichlet distribution through global and local parameter independence. The Annals Statistics 25, 1344–1369.

Geisser, S., 1993. Predictive Inference. Chapman and Hall, New York.

Geraudie, V., Roger, & J. M., O. H. (2010). Développement d'un appareil permettant de prédire la maturité du raisin par spectroscopie proche infra-rouge(PIR). *Revue Française d'Oenologie*, 240, 2-8.

Goelzer, A., Charnomordic, B., Colombié, S., Fromion, V., Sablayrolles, J., 2009. Simulation and optimization software for alcoholic fermentation in winemaking conditions. Food Control 20 (7), 635–642.

Heckerman, D., 1999. A Tutorial on Learning with Bayesian Networks. MIT Press, Cambridge, MA, USA, pp. 301–354.

Jensen Finn V. and Nielsen Thomas D.. (2010) Bayesian Networks and Decision Graphs, Springer-Verlag. 464p.

Jensen, J.S., Demiray, S., Egebo, M., Meyer, A.S., 2008. Prediction of Wine Color Attributes from the Phenolic Profiles of Red Grapes (Vitis vinifera). J. Agric. Food Chem. 56 (3), 1105–1115.

Moody, J., Blanton, J.E., Augustine, M.A., 1996. Enhancing End-User Mental Models of Computer Systems through the Use of Animation. Proceedings of the 29th annual Hawaii International Conference on System Sciences.

Murphy, K.P., 2002. Dynamic bayesian networks: representation, inference and learning. University of California, Berkeley. PhD Dissertation.

Niimi, J., Liland, K.H., Tomic, O., Jeffery, D.W., Bastian, S.E.P., Boss, P.K., 2021. Prediction of wine sensory properties using mid-infrared spectra of Cabernet Sauvignon and Chardonnay grape berries and wines. Food Chem. 344, 128634.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference. Morgan Kaufmann, San Diego, 552p.

Perrot, N., Baudrit, C., Brousset, J. M., Abbal, P., Guillemin, H., Perret, B., & Picque, D. (2015). A Decision support system coupling fuzzy logic and probabilistic graphical approaches for the agri-food industry: prediction of grape berry maturity. PloS one, 10(7), e0134373.

Perrot, N., Ioannou, I., Allais, I., Curt, C., Hossenlopp, J., Trystram, G., 2006. Fuzzy concepts applied to food product quality control: A review. Fuzzy Sets Syst. 157 (9), 1145–1154.

Petropoulos, S., Karavas, C.S., Balafoutis, A.T., Paraskevopoulos, I., Kallithraka, S., Kotseridis, Y., 2017. Fuzzy logic tool for wine quality classification. Comput. Electron. Agric. 142, 552–562.

Raynal, M., Debord, C., Guittard, S., Vergnes, M., 2010. Epicure, a geographic information decision support system risk assessment of downy and powdery mildew epidemics in Bordeaux vineyards. INRA-ISV, Bordeaux, pp. 144–146.

Redner, R.A., Walker, H.F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. 26 (2), 195–239.

Riou, C., 1994. Le déterminisme climatique de la maturation du raisin: application au zonage de la teneur en sucre dans la Communauté Européenne. Office des Publications Officielles des Communautés Européennes.

Sicard, M., Perrot, N., Leclerc-Perlat, M.N., Baudrit, C., Corrieu, G., 2011. Towards 703 integration of experts skills and instrumental data to control food processes. 704 Application to Camembert-type cheese ripening. Journal of Dairy Science 94 (1), 705 1–13.

Vermersh, P., 2006. L'entretien d'explicitation. ESF, Paris.

Van Leeuwen, C., 2010. Terroir: the effect of the physical environment on vine growth, grape ripening and wine sensory attributes Viticulture and Wine Quality. *« Managing wine quality » Woodhead Publishing Series in Food Science.* Technology Nutrition. Chapter 9 273–315. https://doi.org/10.1533/9781845699284.3.273.

Zadeh, L.A., 1965. Fuzzy Sets. Inf. Control 8 (3), 338–353.