Interactive Machine Learning for Applications in Food Science and Technology

Alberto Tonda, Nadia Boukhelifa, Thomas Chabin, Marc Barnabé,
Benoît Génot, Evelyne Lutton, and Nathalie Perrot

Abstract The apparent simplicity of food processes often hides complex systems, where physical, chemical and living organisms' processes co-exist and interact to create the final product. Data can be plagued by *uncertainty*; *heterogeneity* of available information is likely; *qualitative* and *quantitative* data may also coexist in the same process, from expert perception of food quality to nano-properties of ingredients. In order to obtain reliable models, it then becomes necessary to acquire additional information from external sources. Experts of a domain can provide invaluable insight in products and processes, but this precious knowledge is often available only in the form of intuition and implicit expertise. Including expert insight in a model can be tackled by having humans interacting with a machine learning process, through visualization or via specialists in encoding implicit domain knowledge. In this chapter, three selected case studies in food science portray different success stories of combining machine learning and expert interaction. We show that expert knowledge can be integrated at different stages of the modelling process, either online or offline, to initialize, enrich or guide this process.

1 Introduction

When dealing with meaningful representations of food systems, several important issues have to be considered: data can be plagued by *uncertainty*, particularly when chemical, physical, and biological phenomena concur to define the process; *heterogeneity* of available information is also likely, as a vegetable involved in a process can be characterized by more than 40,000 genes, whereas the quality of the final product can be assessed using just a few sensory features; *qualitative* and *quan-*

A. Tonda, N. Boukhelifa, T. Chabin, M. Barnabé, B. Génot, E. Lutton, N. Perrot
INRA, Université Paris-Saclay, 1 av. Brétignières, 78850, Thiverval-Grignon, France, e-mail: {alberto.tonda,nadia.boukhelifa,thomas.chabin,marc.barnabe,benoit.genot,evelyne.lutton,nathalie.perrot}@inra.fr

titative information, from expert perception of food quality, to nano-properties of ingredients, may also coexist in the same process. Consequently, when applying machine learning to agri-food data, the user has to carefully account for variance, manage heterogeneous data, and be able to include both qualitative and quantitative values in the final model.

As gathering data in food science is an expensive and time-consuming process, available datasets are often sparse and incomplete, which poses a challenge to both human modelling practitioners and machine learning algorithms. This issue has been long acknowledged by the community, and ongoing projects have been approved to tackle it, by defining roadmaps to achieve an e-infrastructure for open science [1], and by fostering cooperation between food scientists and modelling experts [2]. In order to obtain reliable models, it thus becomes necessary to acquire additional information from external sources. Experts in a specific domain can provide invaluable insight into products and processes, but this precious knowledge is often available only in the form of intuition and non-coded expertise. Including expert insight in a model is not a straightforward process, but it can effectively be tackled by having humans interacting with a machine learning process, through visualization, or via specialists in encoding implicit domain knowledge [17].

In the following, three selected case studies portray different ways of combining machine learning with expert interaction, in the domain of food processing:

- first, a model for Camembert cheese ripening is built, encompassing variables from the micro-scale (presence of bacteria and chemical components) to the macro-scale (sensory evaluations), relying upon experts to help design the structure of a dynamic Bayesian network;
- a second dynamic Bayesian network model is constructed to help winemakers assess the appropriate time for harvesting grapes, depending on weather conditions
- a graphical model based on symbolic regression is used to help experts create a model of bacterial production and stabilization.

Interaction with the experts of each specific process is always mediated by visualization, complemented by the use of targeted questionnaires (first case study), fuzzy-logic models (second case study), or human-readable equations (third case study). In all considered cases, oriented graphs are used to provide experts with an intuitive and transparent representation of the model under construction. While the models' inner working, ranging from conditional probability inference to computation of free-form equations, is mostly hidden, users can easily interact with oriented graphs, where arcs represent correlation between variables, and modify connections created by learning algorithms, if they are deemed incorrect. For most users, graphs are familiar representations, and manipulating them is intuitive. When users are dealing with graphs that can be considered small, with less than 50 variables, node-link diagrams are a well suited portrayal, while matrices become more appropriate for larger or denser graphs [14].

[1] eRosa European project, `http://www.erosa.aginfra.eu/`

[2] COST Action CA15118 FoodMC, `http://www.inra.fr/foodmc`

2 Dynamic Bayesian network model for Camembert ripening

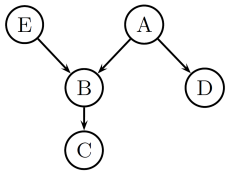
Cheese ripening is a good example of a process that human practitioners can achieve with success but for which several scientific details remain poorly understood. Nevertheless, even for these processes it is possible to create effective models by harnessing knowledge from experts in the domain and coupling it with experimental data. This can be achieved by using an appropriate machine learning framework, that is able to take into account such heterogeneous information. The work presented in [25] shows how the described methodology can be applied to the case of Camembert, a popular French cheese. The desired model goes from micro-scale properties such as concentration of lactose and bacteria, to macro-scale properties such as color and consistency of the crust, with the goal to describe the development of the ripening process, up to the prediction of the current phase of ripening. In Figure 1, a few pictures of the cheese ripening process are reported: experts find it useful to divide the ripening into 4 distinct phases.

Fig. 1 Pictures of Camembert cheese during the ripening process. There are visible changes in the cheese's rind, color, and aroma during the ripening.

The approach used in this experiment is a dynamic Bayesian network (DBN) [18], a variation on a classical Bayesian network [20]. Bayesian networks are probabilistic models widely used to encode knowledge in several different fields: computational biology and bioinformatics (gene regulatory networks, protein structure, gene expression analysis), medicine, document classification, information retrieval, image processing, data fusion, decision support systems, engineering, gaming and law. BNs are directed acyclic graphs, where each node represents a variable in the problem, and links encode correlations between variables. An example of BN is reported in Figure 2.

Like a BN, a DBN is a graph-based model of a joint multivariate probability distribution that captures properties of conditional independence between variables; in the graph, nodes $X_i(t), i = 1,...,N$, represent random variables, indexed by time t. Differently from a regular BN, a DBN is in fact able to encode dependencies between the same variable over multiple instants of time, providing a compact representation of the joint probability distribution P for a finite time interval $[1,\tau]$ defined as follows:

$$P(X(1),...,X(\tau)) = \prod_{i=1}^{N}\prod_{t=1}^{\tau} P(X_i(t)|Pa(X_i)(t)) \quad (1)$$



Node	Parents	Probabilities
A		$P(A=a_1) = 0.99$ $P(A=a_2) = 0.01$
B	A,E	$P(B=b_1 \mid A=a_1, E=e_1) = 0.5$ $P(B=b_2 \mid A=a_1, E=e_1) = 0.5$ $P(B=b_1 \mid A=a_1, E=e_2) = 0.1$ $P(B=b_2 \mid A=a_1, E=e_2) = 0.9$ $P(B=b_1 \mid A=a_2, E=e_1) = 0.4$ $P(B=b_2 \mid A=a_2, E=e_1) = 0.6$ $P(B=b_1 \mid A=a_2, E=e_2) = 0.2$ $P(B=b_2 \mid A=a_2, E=e_2) = 0.8$

Node	Parents	Probabilities
C	B	$P(C=c_1 \mid B=b_1) = 0.3$ $P(C=c_2 \mid B=b_1) = 0.7$ $P(C=c_1 \mid B=b_2) = 0.5$ $P(C=c_2 \mid B=b_2) = 0.5$
D	A	$P(D=d_1 \mid A=a_1) = 0.8$ $P(D=d_2 \mid A=a_1) = 0.2$ $P(D=d_1 \mid A=a_2) = 0.7$ $P(D=d_2 \mid A=a_2) = 0.3$
E		$P(A=e_1) = 0.75$ $P(A=e_2) = 0.25$

Fig. 2 On the left, a directed acyclic graph. On the right, the parameters it is associated with. Together they form a Bayesian network *BN* whose joint probability distribution is $P(BN) = P(A)P(B|A,E)P(C|B)P(D|A)P(E)$.

where $X(t) = X_1(t), ..., X_N(t)$, is called a *slice*, and represents the set of all variables indexed by the same time t. $Pa(X_i)(t)$ denotes the parents of $Xi(t)$. $P(X_i(t)|Pa(X_i)(t))$ denotes the conditional probability function associated with the random variable $X_i(t)$ given $Pa(X_i)(t)$. The joint probability $P(X(1), ..., X(\tau))$ represents the beliefs about possible trajectories of the dynamic process $X(t)$. DBNs are useful tools for combining expert knowledge with data at different levels and length scales. The structure of a model (e.g. the directed graph) can be explicitly built on the basis of expert knowledge, or automatically learned from data by an algorithm [6]. In practice, a combination of the two approaches is commonly used, with a first, automatically-learned structure subsequently corrected by humans, resorting to graphical user interfaces such as BayesiaLab [3] or GeNie [12][4]. Once the structure of a DBN is defined, parameters (i.e. conditional probability functions) can be automatically obtained without a priori knowledge on the basis of a dataset, all through a deterministic machine learning procedure known as *parameter learning*.

In this case study, data is gathered from 6 experiments on the cheese ripening process, each experiment lasting 41 days, with a sampling every day. The information obtained concerns the temperature of the ripening chamber (T, °C), relative humidity (RH, %), and the concentration of lactose (lo, g/kg), lactate (la, g/kg), and the bacteria *Kluyveromyces marxianus* (*Km*, cfu/kg), *Geotrichum candidum* (*Gc*, cfu/kg), *Penicillium camemberti* (*Pc*, cfu/kg), and *Brevibacterium aurantiacum* (*Ba*, cfu/kg). During each experiment, several Camemberts are destroyed to be analyzed, with a considerable economic investment for the producer. At the same time, experts are interviewed to provide additional information. The study involves two groups of experts: 4 cheesemakers with over 15 years of expertise in the industry, and 8 scientists with a track record of over 10 years of research on cheese processes. The questions posed to the experts are carefully constructed in order to elicit expert knowledge, with methods ranging from open-ended questions to focus groups. Values of the variables are discretized in 2 to 12 classes each, depending on expert judgment [2].

[3] `http://www.bayesia.com`

[4] `https://www.bayesfusion.com/`

Following cheesemakers' considerations on the ripening process, the global model is divided into two parts, that are built independently and then linked: **M1** reproduces the temporal links between measured experimental data, simulating how such quantities vary during the ripening process; while **M2** is derived almost entirely from the expert knowledge gathered using questionnaires, and provides a more qualitative assessment between sensory information such as flavor, texture, color, and the ripening phase. Camambert cheesemakers traditionally identify four different phases in the ripening process. Figure 3 shows the final structure of the DBN obtained after the learning process. Variables between **M1** and **M2** are used to link variations in measurable quantities to sensory properties of the cheese.

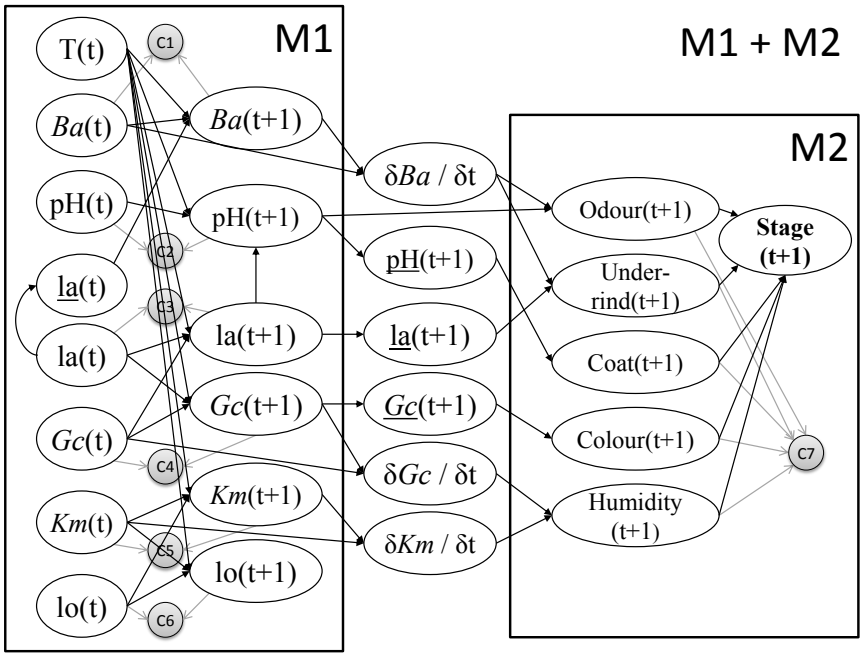


Fig. 3 Final DBN model for the Camembert cheese ripening process. The part denominated M1 represents the variables taken mainly from experimental data, whereas part M2 represents variables derived from expert knowledge and assessment. Grey nodes represent constraints defined by experts. Figure redrawn from [25], with permission from Elsevier.

Figure 4 presents an example of predictions of the dynamics in the process. It is noticeable how the model is able to satisfyingly reproduce the dynamics of variables tied to microbial growth, substrate consumption, and sensory properties, for different temperature conditions. Experts ultimately assessed model simulations resorting to classical two-dimensional plots against test data, and were satisfied with the results.

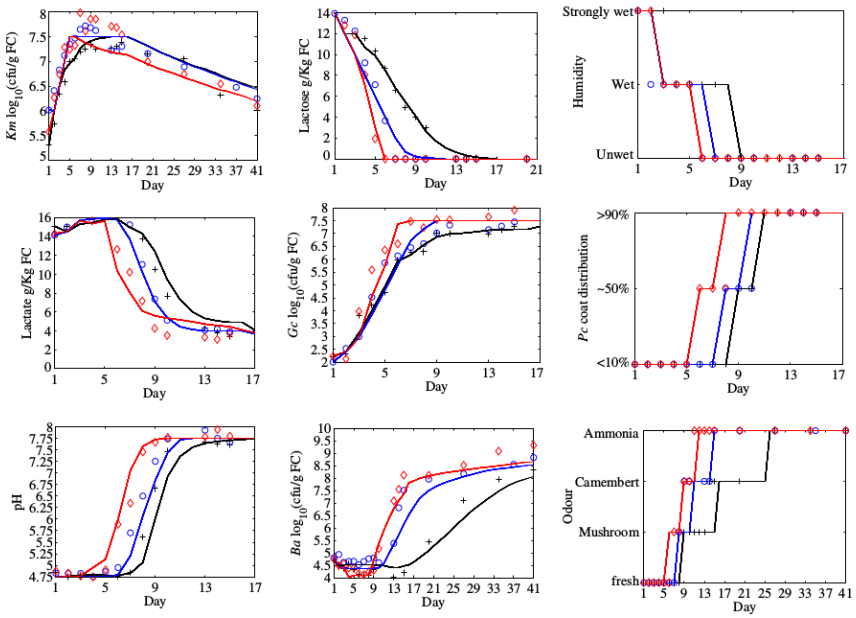


Fig. 4 (color online) Predictions of the Camembert cheese ripening model for the evolutions of **(top row)** microbial growth (*Km*, *Gc*, *Ba* in decimal logarithm scale); **(middle row)** substrate consumption (*lo*, *la*) and **(bottom row)** sensory properties (*RH*, Pc coat and odor). The DBN model's prediction are represented as lines, versus raw data, represented as points, for three different ripening processes, carried out at 8 °C (marked with +), 12 °C (marked with ○) and 16 °C (marked with ◇). Figure reproduced from [25], with permission of Elsevier.

3 Decision-support system for grape maturity prediction

Predicting the right moment to harvest grapes intended for wine production is a task that traditionally is left to specialists in the field. Still, as repercussions of climate change make local weather more unpredictable, experts can use machine learning techniques as a decision support tool, helping them to deal with modified conditions. Such decision support systems are commonly defined as interactive computer-based systems that help organizations in decision-making activities.

In viticulture, some decision support systems are already in use, for example to prevent mildew [23]. Grape berry maturity is analyzed in [10] where the authors built mechanistic models to predict the concentration of sugar in grapes. Other modelling techniques based on spectroscopy predict maturity indicators [13]. These decisions support systems are based solely on experimental data, and do not integrate experts knowledge in order to predict grape maturity. As the human knowledge gained over years of wine production is invaluable and often includes conditions that have not been measured in recent times, it is only sensible to include it as much as possible in the target framework. Expert knowledge handling was already successfully used in the field of viticulture in [7]. Similar to our approach, their model

relies on fuzzy logic but to predict vine development with two indicators, vigor and precocity. In order to predict grape maturity, the innovative work presented in [22] offers a good example of how human expertise can be employed to fill the gaps in experimental data, with the final objective of training a machine learning approach. This study represents the basis of our current work.

For this case study, data related to 66 parcels of land in the Loire Valley is collected over the course of 27 years (1988-2015), for a total of 1,086 data points describing weekly average temperature (T, °C), relative humidity (RH, %), insolation (Ins, hours of sunlight received per day, h/day) and rainfall (Pl, mm). Further data on sugar concentration (S, g/l) and acidity (Ac, g/l Eq H_2SO_4) of the grapes are collected every week, when 200 berries of Cabernet-Franc randomly sampled from the parcels are crushed with a blender and subsequently analyzed. It is important to notice again how obtaining data is an expensive and time-consuming process, and it has to be integrated by expert knowledge, in order to improve the knowledge base eventually used for modelling. For this case study, human expertise is collected through a synthesis of the available literature and industrial reports, performed by 4 scientists and 5 winegrowers working in the areas considered in the study.

As for the previous case study, a Dynamic Bayesian Network proves particularly suited for this application, as such technique makes it possible to employ qualitative and quantitative variables, at different scales, in the same model. The network is designed with the help of the experts, through a trial-and-error process that includes several steps of structure visualization, correction, and analysis of the predictions, initially presented in [1]: the resulting structure is shown in Figure 7 (top). In this particular case study, even with an established structure, computing the parameters of each node is not trivial. Following experts' assessment, in fact, input is discretized into 8 to 15 classes for sugar, acidity, sugar variation, acidity variation, insolation, pluviometry, humidity and temperature. This discretization, featuring a relatively high number of classes when compared to more traditional applications of BNs, leads to conditional probability tables with a considerable amount of combinations: so many, that some of these combinations are not present in experimental data, and thus probabilities for these cases cannot be straightforwardly learned; resorting to experimental data for parameter learning, only, would leave too many gaps. A possible solution to the issue is to resort to experts again, formalizing their knowledge of the process through fuzzy logic mathematical functions.

Fuzzy logic [30] is an extension of the binary logic, where a set is defined by its membership function. A value, x, belongs to a fuzzy set with a membership degree μ_L, with $0 \leq \mu_L(x) \leq 1$, see Figure 5. If we take L a set of *Low* insolation, the membership degree $\mu_L(x)$ of a given insolation value x can be defined as the level up to which insolation x should be considered as *Low*.

Fuzzy sets for the four meteorological variables are then used to build 46 linguistic rules, e.g. *if insolation and pluviometry are Low, then the sugar increase is high*. Each rule is associated by the experts to one of the four classes of meteorological condition, see Figure 6, and is activated according to the activation degree of each rule which define the class. Each class of meteorological condition corresponds to a certain variation of sugar and acidity for one day. The sum of variations on 7 days

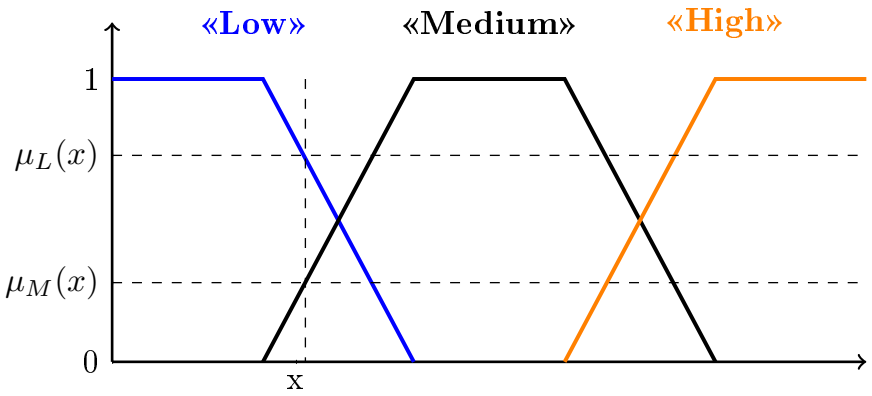


Fig. 5 Example of three fuzzy sets *Low*, *Medium*, *High*, with $\mu_L(x)$: the membership degree in the *Low* fuzzy set and $\mu_M(x)$: the membership degree in the *Medium* fuzzy set.

is performed to produce global variation over the week. This variation of sugar or acidity is added as an input to the DBN.

Fig. 6 Definition of classes related to meteorological conditions defined into four classes for sugar and acidity concentration evolution expressed in g/L. Class index 0: Bad climatic conditions; Class index 1: Not favorable climatic conditions; Class index 2: Standard climatic conditions; Class index 3: exceptional climatic conditions.

The fuzzy logic model is created to produce data for combinations of input variables associated to equiprobability in the probability tables of the DBN; equiprobability, in turn, is associated to combination of input variables never observed in experimental data.

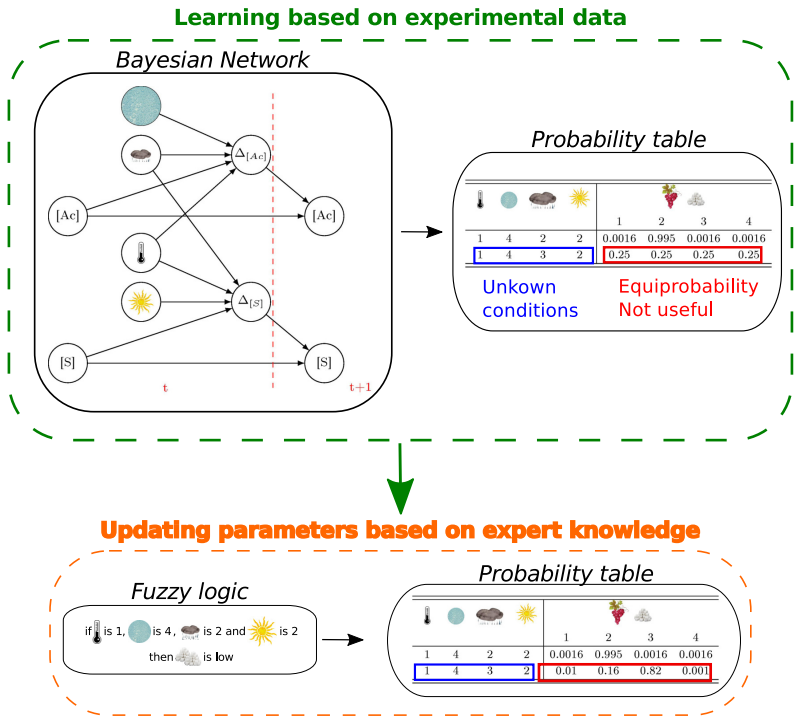


Fig. 7 Proposed framework for the prediction of acidity and sugar content in grapes. (top) Structure of the DBN designed for the prediction of acidity (Ac) and sugar content (S) of grapes. (bottom) Parameters of DBN are updated with data produced by expert knowledge, making it possible to learn robust conditional probability tables for the nodes.

The complete structure of the framework, including the coupling fuzzy logic-DBN is shown in Figure 7. The first step (top) corresponds to the DBN learning based on experimental data. This step allows to produce probability table necessary to perform global predictions. However, some combinations of variable are absent from experimental data. For these specific cases, a probability table is updated using a fuzzy model (bottom). A simulated database is created in variable ranges of interest and variations of sugar and acidity can be produced. These data are included in parallel to experimental data and make it possible to define probabilities in any meteorological conditions necessary.

In order to evaluate the benefit of adding human expertise, the predictions were successively performed with the DBN model, the fuzzy model, and then with combined DBN-fuzzy models, see Figure 8. Best results are obtained by learning from both experimental data and expert knowledge. The resulting model is able to obtain satisfactory predictions, showing good R^2 values (a statistical measure of how close the data are to the fitted regression line) [26] for both sugar content and acidity, with $R^2_S = 0.85$ and $R^2_{Ac} = 0.83$, respectively. In comparison, the DBN model alone obtains $R^2_S = 0.80$ and $R^2_{Ac} = 0.74$ and the expert model alone obtains $R^2_S = 0.81$ and

$R^2_{Ac} = 0.83$. Errors of predictions are shown in Figure 8. We can see that at extremes values, the influence of the coupling DBN-Fuzzy approach is visible with significant improvement.

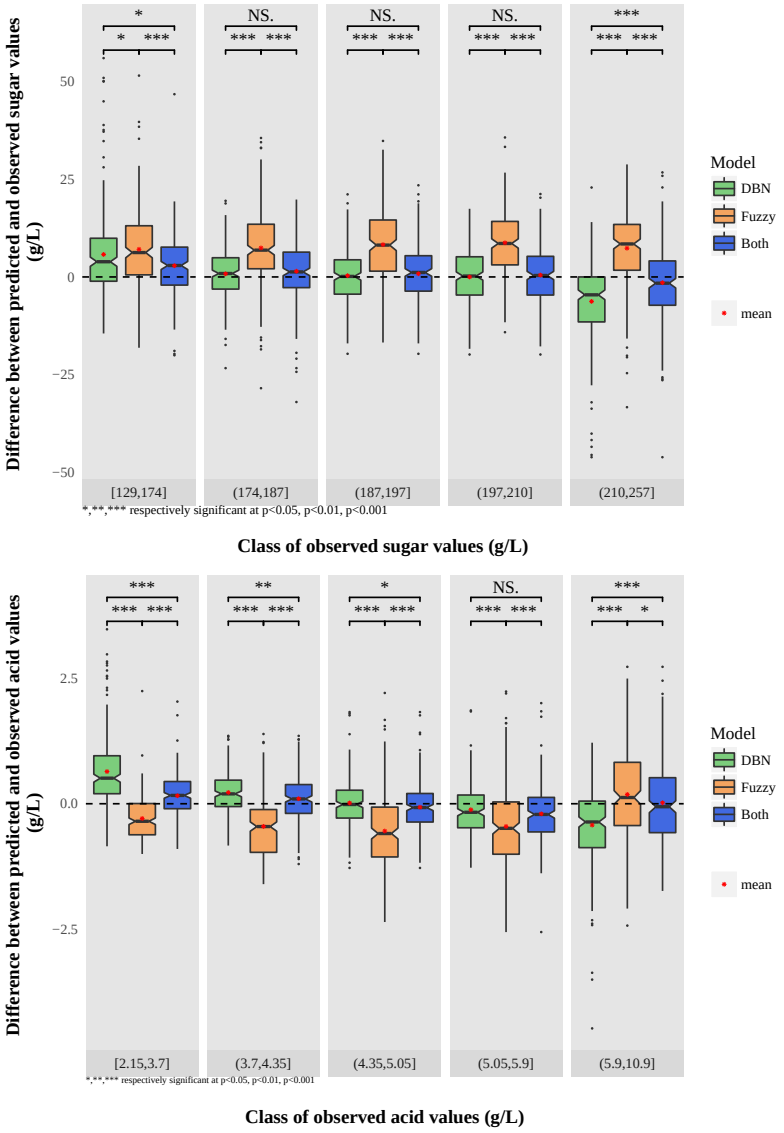


Fig. 8 Prediction error according to class of values, for sugar (top) and acidity (bottom). For each class, the error is reported for the DBN model (green), the fuzzy model (orange) and combined model with DBN and fuzzy method (blue). The combined model clearly obtains the best results.

In the current context of climate change, exceptional meteorological conditions are expected to become more frequent. Learning process performed on experimental data of past years, only, risk to be unsatisfactory. The building of fuzzy models to integrate DBNs offers the possibility to enlarge the range of possible meteorological conditions and make the model more flexible and more robust.

4 Interactive symbolic regression modelling for bacterial production and stabilization

Concentrates of lactic acid bacteria are widely used in the food industry for products such as yogurt, cheese, fermented meat, vegetables and fruit beverages. The quality of bacterial starters, defined by the viability and acidification activity of the cells, depends on numerous control parameters across the different steps of the production and stabilization process, summarized in Figure 9 and described in more details by Champagne and al. [5]. The bacterias levels of resistance to the processes is also dependent to the biochemical and biophysical properties and organization of their membrane [28, 29] which in turn is determined by the genomic expression of the bacteria itself. For these reasons, modelling the bacteria resistance to the process is a complex problem due to many possible non-linear dependencies between the different length scales and steps of the process. In addition, no models are available for several sub-parts of the process, and even those that can be found in literature [19] are often too simple to be included in a wider framework.

One successful approach in modelling complex processes is to stack smaller models such that predictions are propagated between multiple layers formed by these sub-components [9, 8]. In such cases, typically, rich datasets and vast amounts of knowledge are available to describe the stacked components and their interactions. When little data is available, and prior knowledge is limited, mathematical regression techniques can be used to model these complex systems [21]. However, a multitude of candidate models can be obtained through these techniques. Deciding which of these models is the best with respect to the study domain and problem at hand, may be carried out automatically based on a fitness criteria, or delegated to domain experts [16]. While the former is efficient but can result in models that do not capture the reality of the underlying system, the former may be grounded albeit time-consuming. Similarly to Turkay et a. [27], our approach uses mathematical regression to generate candidate solutions. However, we combine automatic evaluation of candidate models, with expert evaluations to ensure both model robustness and validity.

The dataset in this case study concerns the full process of bacteria production and stabilization, with 49 variables measured at 4 different steps (fermentation, freezing, and storage) and at 4 different fermentation conditions (22 °C and 30 °C, with the fermentation stopped at the beginning of the stationary growth phase and 6 hours later). The variables consists of transcriptomics, composition of fatty acid membrane, acidification activity and viability [28]. Such a large number of variables

requires peculiar methods to deal with. Using machine learning capacity to provide automatic modelling enable us to find possible dependencies.

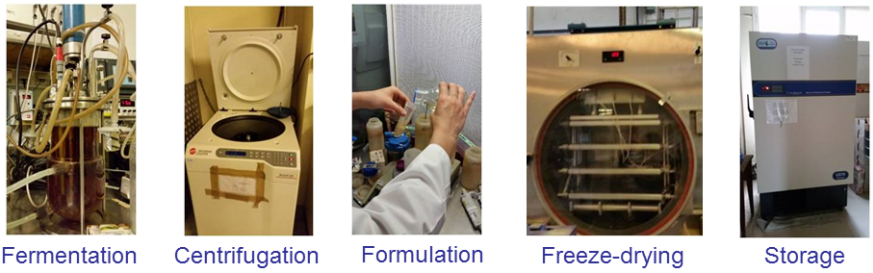


Fig. 9 Steps of the freeze-drying process. Control parameters at every point in the process chain can influence the quality of bacterial starters.

From a vast number of possible dependencies between the measured variables, an automatic methodology can identify the most relevant ones, and combine them to obtain a global model. The main problem of this approach is that the number of variables is far superior to the number of samples in the dataset. The key idea is to remember that experts possess invaluable process knowledge that can considerably improve the robustness of the global model. While formalizing this often-implicit knowledge is not trivial, experts' insights can be effectively included in the modelling process by resorting to interactive approaches. To achieve these objectives, we proposed *LIDeOGraM* (Life science Interactive Development of Graph-based Models), a semi-supervised model learning framework, based on regression analysis [4, 3]. LIDeOGraM is able to obtain free-form equations for each variable in the process, as a function of all other variables. Each equation, describing a sub-part of the global process, can be considered a local model. Such models should fit the experimental data, and at the same time be deemed plausible by the experts. However, when using an automatic technique without expert guidelines, these two goals are often incompatible: it is always possible to find a polynomial equation that perfectly fits the data points, for example with a complex equation featuring as many parameters as data points available but such an equation could overfit the dataset, failing to represent the underlying relationship between the variables, and ultimately poorly predict the unseen data.

To avoid this issue, every variable in LIDeoGraM is associated with a set of candidate equations, obtained through symbolic regression [15]. Eureqa[5] [24], a commercial software specialized in symbolic regression, is able to obtain a set of possible equations for every variable in a given dataset. A local model can thus be associated to each variable by selecting one of the equations in the set. Symbolic regression makes it possible to effectively search the vast space of all possible mathematical expressions, taking into account both the fitting of the equation and

[5] `http://nutonian.com/products/eureqa/`

its complexity – indeed, more complex equations tend to be overfitted, while simpler ones are often unable to characterize the data. A collection of local models will then constitute the base for a global model, built using an evolutionary optimization algorithm [11] that stochastically searches the space of all sets of local models for the one that best fits the global dataset. To evaluate a candidate global model, the input nodes are set to known experimentally-measured values, and the errors in the prediction are averaged over all nodes, thus obtaining a global error, that the evolutionary algorithm aims to minimize.

Human experts are then involved in the modelling process, via a graphical user interface, showing a node-link graph visualization of the global model, where each node represents a variable, and each link marks a possible dependency between two variables. This interface allows experts to visualize the results from Eureqa, contribute with their knowledge, and finally lead the search for an efficient global model.

For this objective, two views are available. The **Local model view** shows an overall qualitative view of the equation sets given by Eureqa for each variable. This view enables nodes with no satisfactory equation in terms of fitting and/or complexity to be easily spotted. The **Global model view** shows the predictive capability of the current global model, for each variable. This view enables users to rapidly assess which variables in the global model are poorly predicted, but also which ones may be responsible for the poor predictions of their dependent nodes.

LIDeoGraM has several ways to add expert knowledge. First, it is possible to attribute a category to each variable, and specify the available dependencies between categories for the symbolic regression. A category of nodes can represent a step in the process, or a scale of information. This interface is presented in Figure 10.

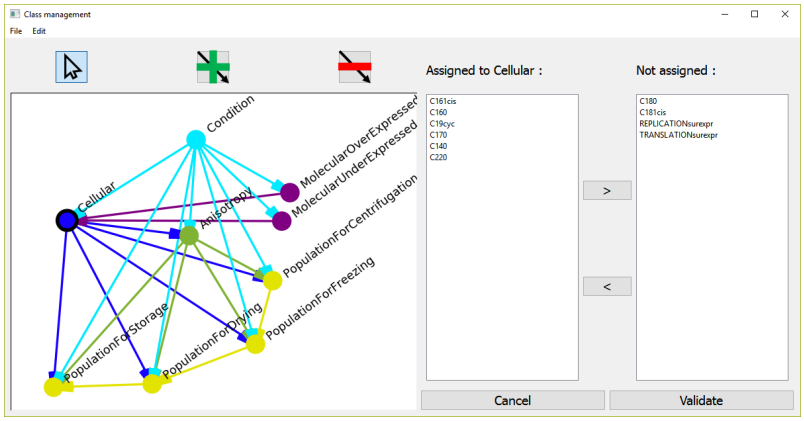


Fig. 10 Screenshot of the interface allowing to choose the authorized links between the defined classes. A link between two classes means that all variables associated to the parent class can be used in the equations for all variables associated to the child class. The displayed graph represents the selected constraints chosen for the presented results.

After obtaining a set of equations for every node, experts can then filter it by specifying that certain kinds of node-to-node dependencies are not allowed. Experts can then manually add new equations in the set of candidate local models for a node, and eventually restart the search for a global model after putting all their constraints in place. With LIDeOGraM, it is possible to learn global models for the production and stabilization of bacteria. Such models can then be used to better understand how to preserve the quality of the culture during the process, foster the emergence of new hypotheses, and design new experiments, whose data could in turn be used to further improve the global model. These functionalities are demonstrated in Figure 11.

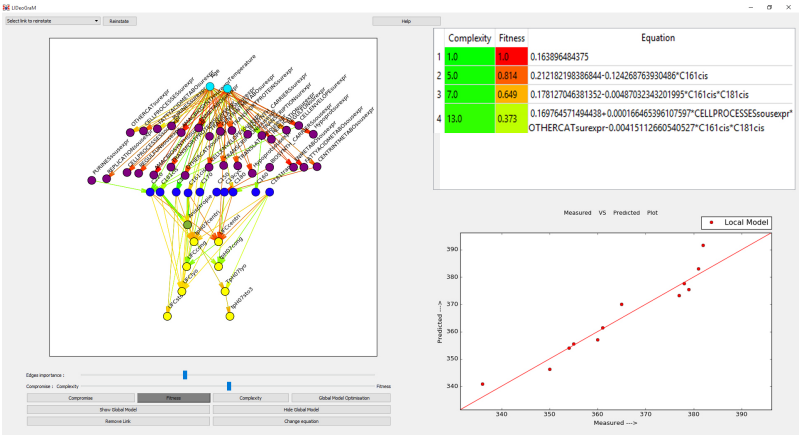


Fig. 11 Screenshot of LIDeOGraM. The left side shows a graphical model representing the mean fitness of the local models obtained by symbolic regression. The top-right part is the list of equations proposed by Eureqa for the selected node, and the bottom-right part shows a plot of the measured versus predicted data associated to the selected equation.

Results obtained for the previously described dataset [28] are presented in Figure 12.

In a preliminary experiment on the presented framework, a user with 20 years of experience on freeze-drying process is able to inject their knowledge into the optimization process. Out of a total of 232 equations generated for the local models, the expert deletes 5 equations, and 2 nodes, removing in turn 14 more equations in which the 2 deleted variables are involved. The expert then restarted symbolic regression on 3 nodes, obtaining 12 new equations. At the end of this process, the global optimization results are better than without the expertise, with the average error computed on all nodes being 0.801, using only the automatic approach, and 0.787 combining the automatic approach with expert interaction. Figure 13 shows the evolution of the mean error per node, for both the automatic and the combined approaches. The results are still not completely satisfactory, as the prediction error for some of the nodes remains large, but the positive influence of the expert on the machine learning process is already substantial. In future works, more data points

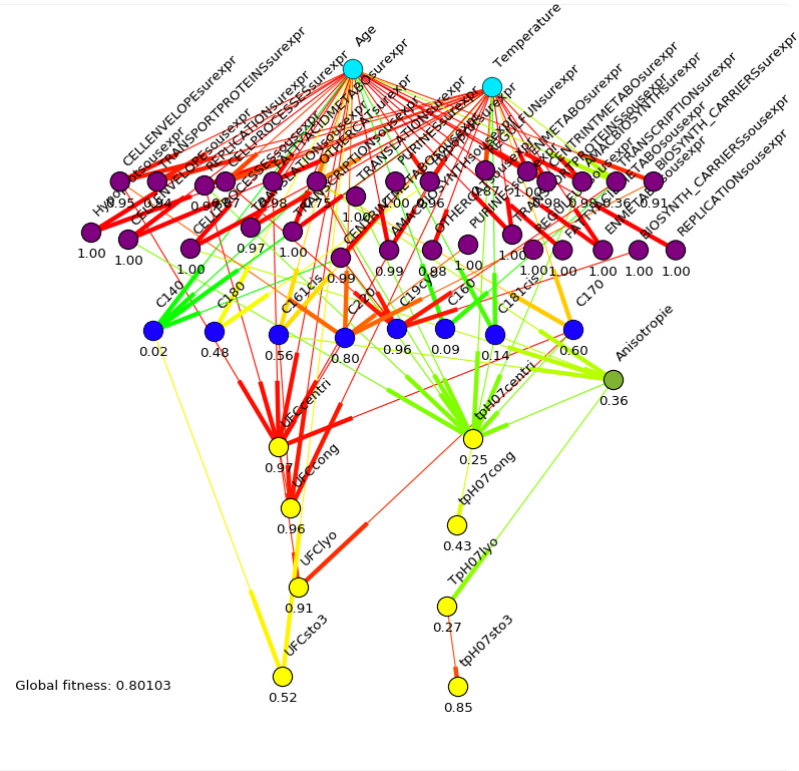


Fig. 12 (color online) Graphical model generated in LIDeOGraM representing a (optimized) global model. Nodes are organized in 4 categories: experimental conditions, genomic scale, cellular scale, and population scale. A Pearson correlation coefficient, calculated using the predictions from the global model compared to the experimental measurements, is printed below each node. An edge between two nodes means that the parent variable is used in the equation chosen to calculate the child variable. The color of an edge depends on the Pearson correlation coefficient, which represent the quality of the prediction. The color varies from red for a poor-quality prediction to green for a satisfying one.

will be collected, and experiments with several other experts on the freeze-drying process are scheduled.

5 Discussion and guidelines

Computational Modelling is an iterative process that comprizes three main activities: *designing* a model where the aim is to define a suitable representation for objects and their relationships; *exploring* the model to understand its behavior, and *tuning* it to find the best or optimal parameter values to obtain good predictions. Our approach in building interactive machine learning systems for food science and

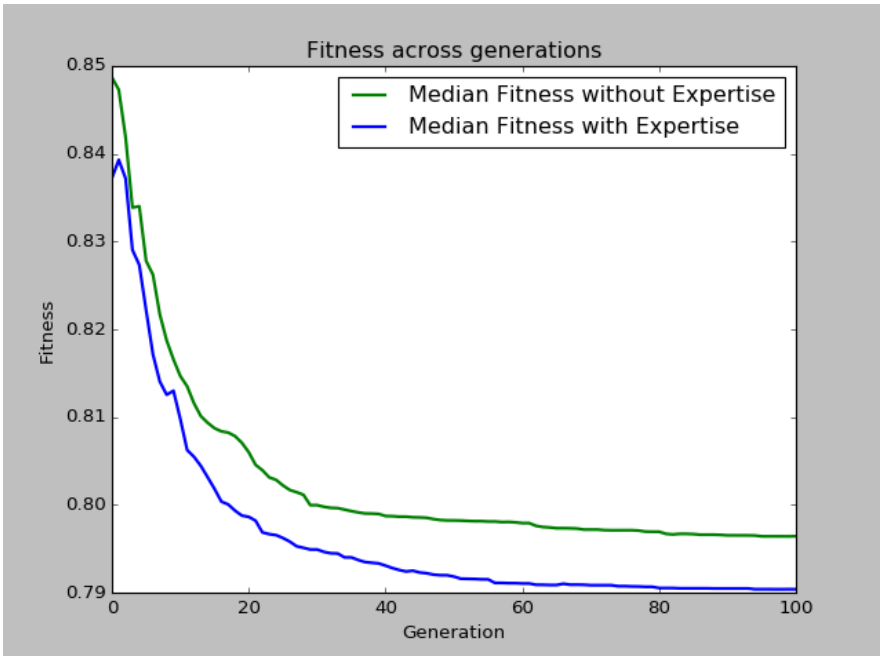


Fig. 13 Comparison of an experiment on the learning of the freeze-drying model, using LIDeOGraM with and without human interaction. The term *fitness* here refers to the average error, computed on all variables in the problem. The *generations* are the iterations of the evolutionary algorithm used for optimizing the global model.

technology focuses on involving experts of the process in one or more stages of this modelling pipeline, facilitating their interactions with the machine learning process through visual representations.

For the first two case studies, on modelling Camembert cheese ripening and grape maturity prediction, expert knowledge is integrated primarily at the design stage of model building. Using established methodologies from the knowledge elicitation domain (e.g. interviews, case studies, and observations), expert knowledge can be collected, coded and formalized into a probabilistic model. The goal, in these cases, is to create a knowledge representation of the process that matches the domain experts mental model.

For the last case study, on modelling bacterial production and stabilization, experts knowledge is integrated at each stage of the modelling process. At the design stage, to structure the relationship of variables and system constrains prior to launching the automatic machine learning and optimization algorithms; and post-model learning through various user interactions via the LIDeOGraM interface. For instance, domain experts can add or remove variables, classes and constraints. They could filter local models, or add new equations to explore how well they fit their data.

While interaction with experts is invaluable even with classical approaches, in the food science domain we argue for a more user-centered design approach to machine learning, whereby users can participate at each stage of modelling process, from design to exploration and tuning. This involvement not only helps domain experts understand computational models better, but it allows them to confront their domain knowledge and know-how with with the results of machine learning, ultimately making machine learning more transparent. Our informal evaluations and discussions with domain experts allowed us to observe the following:

- providing visual representations of machine learning models improves user engagement and encourages feedback, especially if domain experts are involved at the design stage and exploration stages.
- graph-based model representations are easy to understand, but multiple linked representations are more helpful when trying to understand the model.
- experts tend to take a multi-step approach to model validation, first to verify existing knowledge (most likely to build trust in the ML algorithm), then to assess new predictions. When doing so, they first look at the general high-level dependencies between variables, before looking at detailed information such as values of weights, or data in the conditional probability tables when DBNs are involved.

It remains to prove whether making machine learning more transparent helps domain experts better explore and validate computational models in food science. More research is needed to study whether user-centered design for modelling improves decision making and helps indeed build trust in constructed models. From our experience, we believe this to be the case, but a more formal assessment is required to properly evaluate our intuition.

6 Conclusions

In this chapter we illustrated through three case studies from the agri-food domain, how integrating experts knowledge into computational modelling can yield promising results. These real-world case studies portrayed different ways of combining machine learning with experts interaction to design, explore and tune machine learning models. In the first case study, domain experts helped design the structure of a dynamic Baysesian network to predict Camembert cheese ripening process. In the second case study, winemakers interacted with a dynamic Bayesian Model, to help choose the appropriate time for harvesting grapes. In the third case study, domain experts interacted with a symbolic regression model, to help create a grounded model of bacterial production and stabilization. Based on our experience in working closely with domain experts, we concluded this chapter with general observations and recommendations. We argue that more research in user-centered design methodologies for machine learning is needed, to enable domain experts to truly become model co-builders.

References

1. C. Baudrit, N. Perrot, J. M. Brousset, P. Abbal, H. Guillemin, B. Perret, E. Goulet, L. Guerin, G.Barbeau, and D. Picque. A probabilistic graphical model for describing the grape berry maturity. *Computers and Electronics in Agriculture*, 118:124–135, oct 2015.
2. C. Baudrit, M. Sicard, P.-H. Wuillemin, and N. Perrot. Towards a global modelling of the camembert-type cheese ripening process by coupling heterogeneous knowledge with dynamic bayesian networks. *Journal of Food Engineering*, 98(3):283–293, 2010.
3. T. Chabin, M. Barnabé, N. Boukhelifa, F. Fonseca, A. Tonda, H. Velly, N. Perrot, and E. Lutton. Interactive evolutionary modelling of living complex food systems: freeze-drying of lactic acid bacteria. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 267–268. ACM, 2017.
4. T. Chabin, M. Barnabé, N. Boukhelifa, A. Tonda, H. Velly, B. Lemaitre, N. Perrot, and E. Lutton. Lideogram : An interactive evolutionary modelling tool. In *Proceedings of the International Conference on Artificial Evolution (Evolution Artificielle)*, 2017.
5. C.P. Champagne, N. Gardner, E. Brochu, and Y. Beaulieu. freeze-drying of lactic acid bacteria. a review. *Canadian Institute of Food Science and Technology journal: Journal de l'Institut canadien de science et technologie alimentaire*, 1991.
6. J. Cheng, D. A. Bell, and W. Liu. An algorithm for bayesian belief network construction from data. In *proceedings of AI & STAT97*, pages 83–90, 1997.
7. C. Coulon-Leroy, B. Charnomordic, D. Rioux, M. Thiollet-Scholtus, and S. Guillaume. Prediction of vine vigor and precocity using data and knowledge-based fuzzy inference systems. *Journal International des Sciences de la Vigne et du Vin*, 46(3):185–205, 2012.
8. M.-J. Cros, M. Duru, F. Garcia, and R. Martin-Clouaire. A biophysical dairy farm model to evaluate rotational grazing management strategies. *Agronomie*, 23(2):105–122, 2003.
9. Z.-W. Dai, P. Vivin, and M. Génard. Modelling the effects of leaf-to-fruit ratio on dry and fresh mass accumulation in ripening grape berries. In *VIII International Symposium on Modelling in Fruit Research and Orchard Management 803*, pages 283–292, 2007.
10. Z.-W. Dai, P. Vivin, T. Robert, S. Milin, S. H. Li, and M. Génard. Model-based analysis of sugar accumulation in response to source–sink ratio and water supply in grape (vitis vinifera) berries. *Functional Plant Biology*, 36(6):527–540, 2009.
11. K. A. De Jong. *Evolutionary computation: a unified approach*. MIT press, 2006.
12. M. J. Druzdzel. Smile: Structural modeling, inference, and learning engine and genie: a development environment for graphical decision-theoretic models. In *Aaai/Iaai*, pages 902–903, 1999.
13. M. Fadock, R. B. Brown, and A. G. Reynolds. Visible-near infrared reflectance spectroscopy for nondestructive analysis of red winegrapes. *American Journal of Enology and Viticulture*, pages ajev–2015, 2015.
14. M. Ghoniem, J.-D. Fekete, and P. Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization*. IEEE.
15. J. R. Koza. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press, 1992.
16. J. Krause, A. Perer, and E. Bertini. Infuse: interactive feature selection for predictive modeling of high dimensional data. *IEEE transactions on visualization and computer graphics*, 20(12):1614–1623, 2014.
17. E. Lutton, A. Tonda, N. Boukhelifa, and N. Perrot. Complex systems in food science: Human factor issues. In Jan Van Impe, editor, *FoodSIM*. EUROSIS-ETI, 2016.
18. K. P. Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.
19. S. Passot, F. Fonseca, S. Cenard, I. Douania, and I. C. Trelea. Quality degradation of lactic acid bacteria during the freeze drying process: Experimental study and mathematical modelling. 2011.

20. J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.
21. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.
22. N. Perrot, C. Baudrit, J. Marie Brousset, P. Abbal, H. Guillemin, B. Perret, E. Goulet, L. Guerin, G. Barbeau, and D. Picque. 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, jul 2015.
23. M. Raynal, C. Debord, S. Guittard, and M. Vergnes. Epicure, a geographic information decision support system risk assessment of downy and powdery mildew epidemics in bordeaux vineyards. 2010.
24. M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.
25. M. Sicard, C. Baudrit, M.N. Leclerc-Perlat, P.H. Wuillemin, and N. Perrot. Expert knowledge integration to model complex food processes. application on the camembert cheese ripening process. *Expert Systems with Applications*, 38(9):11804–11812, sep 2011.
26. R. G. D. Steel and H. James. *Principles and procedures of statistics: with special reference to the biological sciences*. McGraw-Hill, 1960.
27. C. Turkay, A. Slingsby, K. Lahtinen, S. Butt, and J. Dykes. Supporting theoretically-grounded model building in the social sciences through interactive visualisation. *Neurocomputing*, 2017.
28. H. Velly, M. Bouix, S. Passot, C. Penicaud, H. Beinsteiner, S. Ghorbal, P. Lieben, and F. Fonseca. Cyclopropanation of unsaturated fatty acids and membrane rigidification improve the freeze-drying resistance of lactococcus lactis subsp. lactis tomsc161. *Applied microbiology and biotechnology*, 99(2):907–918, 2015.
29. H. Velly, F. Fonseca, S. Passot, A. Delacroix-Buchet, and M. Bouix. Cell growth and resistance of lactococcus lactis subsp. lactis tomsc161 following freezing, drying and freeze-dried storage are differentially affected by fermentation conditions. *Journal of applied microbiology*, 117(3):729–740, 2014.
30. L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338 – 353, 1965.