ORIGINAL ARTICLE

Artificial intelligence-based decision support system to manage quality of durum wheat products

Rallou Thomopoulos[1,2], Brigitte Charnomordic[3], Bernard Cuq[1] & Joël Abécassis[1]

1 IATE Joint Research Unit, INRA-Supagro-UM2-CIRAD, Montpellier, France
2 LIRMM, CNRS-UM2, Montpellier, France
3 ASB Joint Research Unit, INRA-Supagro, Montpellier, France

Key words
decision support; food processing; expert knowhow; durum wheat chain.

Correspondence
Joël Abécassis, IATE Joint Research Unit, INRA-Supagro-UM2-CIRAD, F-34060 Montpellier, France
Tel: +33 499 612 203
Fax: +33 499 613 076
Email: abecassi@supagro.inra.fr

Received 14 May 2009; Revised 23 June 2009; Accepted 15 July 2009

doi:10.1111/j.1757-837X.2009.00029.x

Abstract

Background The long term competitiveness of food companies as well as the general health and wellness of citizens depend on the availability of products meeting the demands of safe, healthy and tasty foods. Therefore there is a need to merge heterogeneous data in order to develop the necessary decision support systems. *Aims* The objective of this paper is to propose an approach for durum wheat chain analysis based on a knowledge management system in order to help prediction. *Material and Methods* The approach is based on an information system allowing for experimental data and expert knowledge representation as well as reasoning mechanisms, including the decision tree learning method. *Results* The results include the structure of the knowledge management system for durum wheat process data, statistics and prediction results using decision trees. The use of expert rules for decision support is introduced and a method for confronting expert knowledge with experimental data is proposed. Different case studies from the durum wheat process are given. *Discussion* Our specific original contributions are: the design of a hybrid system combining both data and knowledge, the advantage of not requiring an a priori model, and therefore, an increased genericity, and the potential use for both risk and benefit analysis. *Conclusion* The approach can be reused for other purposes within the chain, and can also be transferred to other domains. Such a project is a starting point to integrate new knowledge from multidisciplinary fields, and constitutes a tool for structuring the international cereal research community.

Introduction

Cereal food design until today still relies more on experience than on science, although the efforts during the last 20 years have considerably increased the number of research projects. This resulted in an explosion of scientific papers that – even if they are relevant on their scale of observation – can hardly be used in practice because they have not been completely integrated into a corpus of knowledge. At the same time, the cereal and pasta industry has developed from traditional companies relying on experience and having a low rate of innovation, to a dynamic industry geared to follow consumer trends: healthy, safe, easy to prepare, and pleasant to eat. Indeed, the long-term competitiveness of food companies as well as the general health and wellness of citizens depends on the availability of products meeting the demands of safe, healthy and tasty foods.

To meet such criteria, current industry requires knowledge from its own know-how as well as from different disciplines. A challenge for industry is to identify and extract only the key information from all the available data. Today it is not the scarcity of an information that makes its value, it is how to manage it, to integrate it into a system to make it available at the right time and the right place, which is becoming the major stakes. Therefore, there is a pressing need to merge different data into a comprehensive effort

dedicated to developing the necessary tools and decision support systems.

Knowledge management systems have been proposed in food science in order to help prediction

In particular, several works have dealt with the problem of food security, and therefore have been proposed for risk assessment, such as in predictive microbiology to prevent microbiological risk in food products. Such systems combine a database with mathematical models which, applied to the data, allow one to deduce complementary information (Zwietering *et al*., 1992; Peck *et al*., 1994; Baranyi & Tamplin, 2003; Haemmerlé *et al*., 2007).

For instance, among these models, the Sym'Previus model relies on three distinct databases: a classic database containing structured experimental data from the literature, and two semi-structured databases containing experimental data with unexpected structure, and experimental data from tables found on the web (Haemmerlé *et al*., 2007). In Sym'Previus the three databases are queried simultaneously by means of a unique interface, without any cross exchange.

In the field of cereal transformation, we can also note the 'virtual grain' system (Mueangdee *et al*., 2006), which gathers in a database heterogeneous information concerning the properties of cereal grains as well as the technical and agronomical itineraries for their production. The aim is to identify influential factors and potential relationships between properties that are usually studied separately: morphological, biochemical, histological, mechanical and technological properties. The database is connected to statistical and numerical computing tools, in particular a wheat grain cartography tool developed using Matlab. Based on wheat grain properties and information on the components distribution, it proposes a local representation of the properties in each tissue. However, the final objectives of the Virtual Grain project were related to the explanation of the grain behavior during fractionation, they do not include considerations on subsequent food products.

In the close domain of breadmaking technology, the 'Bread Advisor' tool has been a pioneer as a knowledge software for the baking industries (Young *et al*., 2001; Young, 2007). Although experimental data and dynamic prediction are not proposed, this tool, exclusively based on expert knowledge stored in a database, provides three kinds of information: text information about the processing methods, list of possible faults and their causes, and generic messages about the effects of process changes.

Proposed approach

The approach we propose to manage quality of durum wheat products has several original advantages:

- Firstly, the system allows evaluating both faults and qualities of food products: it is not only a risk assessment system, dedicated to a specific risk. While durum wheat is the subject domain, it can be generalized or adapted for other food application fields.
- Secondly, the system is a hybrid system, because it takes into account two categories of information: (i) experimental data (from the international literature of the domain) that describe in a quantified way experiments carried out and results obtained, to deepen the knowledge of the domain; (ii) expert statements that express generic knowledge rising from the experience of domain specialists and describing commonly admitted mechanisms, in a qualitative way. We can thus highlight a first important contribution of our work: *proposing a solution in case of lack of experimental data, by taking into account expert knowledge*.

Another specificity of our system is the possibility of confronting the different categories of information. We propose a mechanism to confront expert and experimental information when both are available.

- Finally, our approach for prediction is not based on predetermined models. In the food domain, such models are often unavailable, therefore learning techniques can prove useful. The second main contribution of our work: *providing prediction solutions in case of lack of models, by proposing dynamically learnt models* of two kinds. To exploit the experimental data, beside classic statistical data analysis, we use an existing classification and prediction model, namely decision trees. To exploit expert knowledge, we have introduced an original learning method that allows the refinement of existing knowledge through the comparison with the experimental data. Our current approach thus includes specific industry process analysis as well as expert know how reasoning, and for which final quality predictions are required.

Materials and methods

This section presents the construction of the representation system, the models used for data and knowledge representation, and the decision tree learning method.

Building a representation system

The building of the knowledge management system concerning the processing and qualities of food products mainly

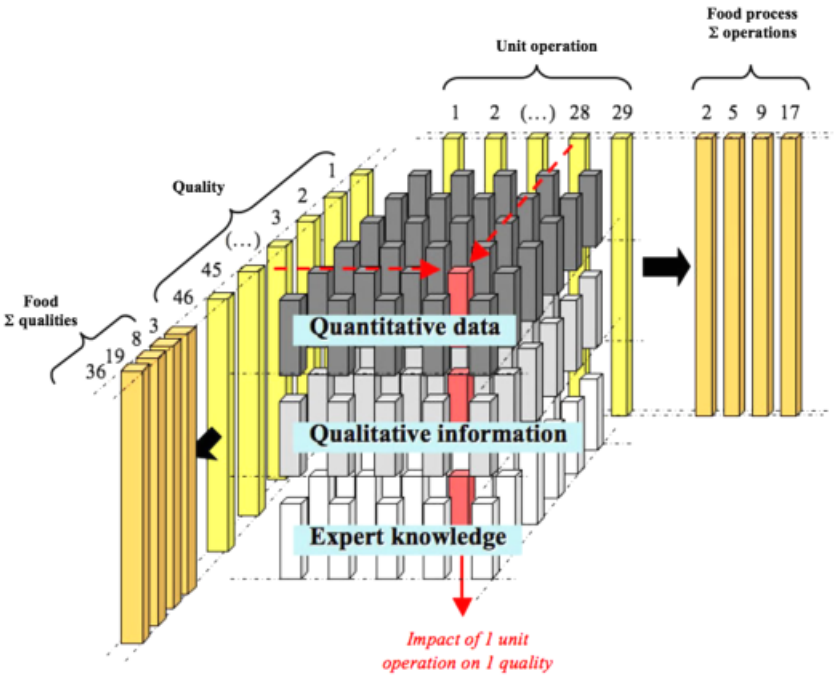

Figure 1 Building a representation system.

relies on the conception of an original structure, allowing organizing all information that is relevant of the studied domain. The diversity of information comes from:

- Its level: general information, process, qualities, impacts, etc.
- Its nature: quantitative scientific information, descriptive information, expert information, etc.

1. The originality of the structure that is proposed for the classification of the information comes from the development of a system, which is able to *integrate the information coming from different domains*. This integration is based on two points: the individualization of information and the creation of links between pieces of individualized information.

(i) The individualization of information is proposed as a principle for input classification within the knowledge management system. We thus define the inputs according to two dimensions:

- The 'technical' axis is defined by all the unit operations, which are involved in transformation from raw materials to end products (e.g., grinding, storage, drying, baking, etc.).
- The 'quality' axis is defined by all the criteria, which are used to represent the end-quality(ies) of food products, according to three aspects: organoleptic, nutritional and safety properties (e.g., colors, vitamins contents, pesticides contents, etc.).

(ii) The building of links between pieces of individualized information is designed to represent:

- The *impact of a unit operation on a given quality* (i.e., the crossing of a parameter of the 'technical' axis and a parameter of the 'quality' axis).
- The *multifactorial character of food quality parameters* (i.e., a 'sum' of several parameters from the 'quality' axis).
- The *technological ways*, which link a group of unit operations (i.e., the manufacture process of food) that transform the raw materials into finished products (i.e., a 'sum' of several unit operations).
- The *interactions between unit operations, to describe* the possible effect of a preliminary unit operation during a food processing, on the behavior during a subsequent unit operation.

2. The originality of the knowledge managing system also comes from the development of a system, which is able to *integrate information of different natures*. This integration is based on the use of knowledge management tools, adapted to the information structure and diversity.

- *Quantitative scientific information* (i.e., numerical data as variables or responses) are represented using a database system, which is structured in the form of a relational database using MySQL.

• *Descriptive scientific information* (i.e., analyses, hypotheses, graphs, illustrations, discussions, etc.) are text and pictures XML files.
• *Expert information* is relative to the knowledge of the different domains about food transformation and qualities. The formalism chosen to represent expert statements is the *conceptual graph model*, an artificial intelligence formalism of the semantic networks family.

A schema of the system is presented in Figure 1.

Data and knowledge representation models

The experimental data are represented in the relational model (Codd, 1970), which is the most commonly used database model, for the storage and querying of structured data. It is a well-studied, robust and efficient model based on the structuration of data into 'tables' called relations.

The model used to represent the expert knowledge is the conceptual graph model (Sowa, 1984). It is a knowledge formalism that comes from the artificial intelligence domain, initially introduced as well-suited for natural language representation. It was chosen for its graphical representation of both knowledge and reasoning, relatively intuitive for nonspecialists (Bos *et al*., 1997). The conceptual graph model presents interesting characteristics: it allows one to add or remove pieces of information easily; it is a graphic model by nature; it is defined on a 'support' that contains the vocabulary of the domain, partially ordered by the 'kind of' relation, useful to model and search information. To illustrate the 'kind of' relation, let us take the example of *Drying Processes*: *High temperature drying* is a 'kind of' *Quick Drying*, which is, with 'Slow drying,' a kind of *Drying*.

The conceptual graph model is based on two robust mathematical theories, the first-order logic and the graph theory. We use the formalization of the conceptual graph model presented by Mugnier (2000), and of conceptual graph rules (Salvat & Mugnier, 1996), which form an extension of the conceptual graph formalism, and in which rules of the form 'if A then B' are added to a knowledge base, where A and B are two simple conceptual graphs.

Conceptual graphs and conceptual graph rules are both illustrated in Figure 2. Figure 2a shows an example of rule. It represents the information 'If a drying has for input a pasta product in which peroxydase is active and produces an output pasta product, then this product is characterized by a brown color.' Figure 2b and c are two examples of simple conceptual graphs. The former reads easily as: 'the pasta product P2 is characterized by a brown color.' It is actually the result of applying the rule given above to the latter graph, which is called a *specialization* of the hypothesis of rule 2a, and therefore can be used as input to this rule.

Decision tree learning

Decision tree learning is a method commonly used in data mining and statistical multivariate analysis. Compared with other multivariate methods, the advantages of

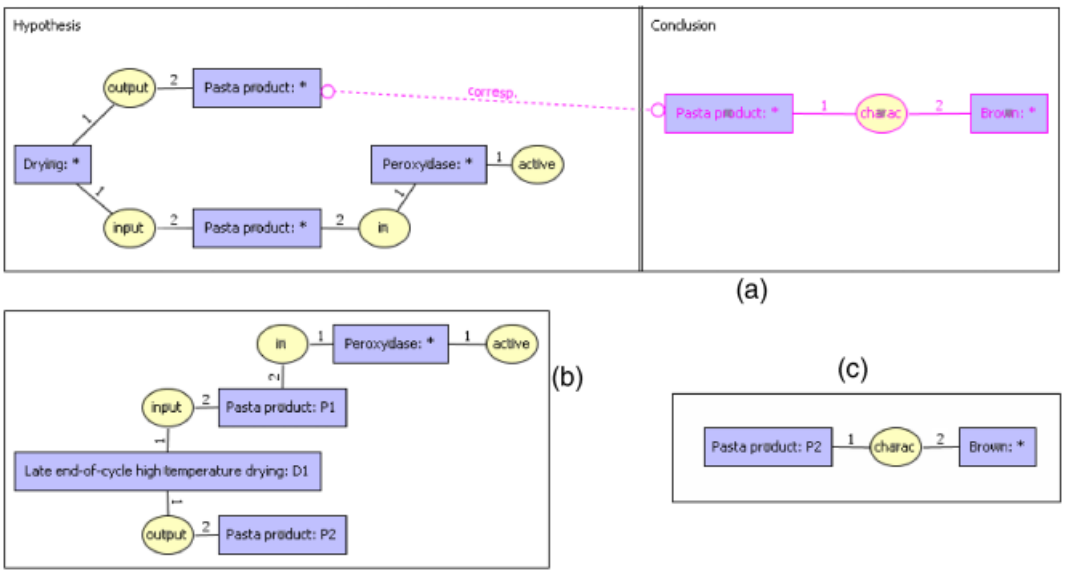

Figure 2 Example of use of expert knowledge. (a) Example of expert rule represented in the conceptual graph model. (b) Example of an input conceptual graph. (c) Output conceptual graph obtained by applying the rule (a) on the input conceptual graph (b).

decision trees are the descriptive/predictive capabilities, the simultaneous handling of numerical/symbolical features, the use of a similar formalism for regression and classification cases and finally the possibility of dealing with missing values. Decision trees are also simple to understand and interpret with a few explanations, and constitute a 'white box' model. As the most discriminant features are selected first, decision trees are often used in feature selection procedures previously to more precise (linear or non-linear) models.

Decision trees are both descriptive and predictive models. In the tree structure, leaves represent classifications (or average values) of a dependent variable. Branches represent conjunctions of input features that lead to those classifications (or average values). The tree can be seen as a collection of rules based on values of the more discriminant variables in the modeling data set.

Several implementations of decision trees are available in the literature, depending on the criteria used for building the tree. In this paper, we use CART recursive dichotomous partitioning (Breiman, 1984) in the R software (Ihaka & Gentleman, 1996) *rpart* implementation. Rules are selected based on how well splits determined using input features values can differentiate observations regarding the dependent variable. Once a rule is selected and splits a node into two, the same logic is applied to each 'child' node (i.e., it is a recursive procedure). Splitting stops when CART detects no further gain can be made, or some preset stopping rules are met. Each branch of the tree ends in a terminal node: each observation falls into one and exactly one terminal node; each terminal node is uniquely defined by a set of rules.

Decision trees are called classification trees when the dependent variable is a class, and regression trees when it takes continuous values. The splitting criterion is used to decide which variable gives the best split at each node.

Splitting criteria

Suppose data comes in n records of the form

$$(x, y) = (x_1, x_2, x_3 \ldots, x_p, y).$$

Let us first give the splitting criterion for classification trees. The dependent variable, Y, is the variable that we are trying to classify. The other variables, x_1, x_2, x_3, etc., are the input features. Suppose that y takes on values in labels (1, 2, . . . , C). The splitting criterion uses the Gini impurity and reaches its minimum (zero) when all cases in the node fall into a single target category (*pure node*). The Gini impurity at node T is defined as:

$$I(T) = 1 - \sum_{k=1}^{C} pk^2,$$

where pk is the proportion of items at node T that belong to class k.

The splitting criterion is chosen as the one with maximal impurity reduction:

$$\Delta I = p(T)I(T) - p(L)I(L) - p(R)I(R)$$

where $I(L)$ and $I(R)$ are the impurity indexes for the right and left son, respectively, while $p(T)$, $p(L)$ and $p(R)$ are the proportions of items assigned to nodes T, L and R, respectively.

Let us now consider the case of regression trees, when the dependent variable takes continuous numerical values. Then the splitting criterion is:

$$\mathrm{SS_T} = \mathrm{SS_L} + \mathrm{SS_R}, \text{ where } SST = \sum_i (y_i - \bar{y})^2$$

where $i = 1 \ldots n$, and y_i is the ith label value for the dependent variable, is the sum of squares for the node, and $\mathrm{SS_L}$, $\mathrm{SS_R}$ are the sum of squares for the right and left son, respectively. This is equivalent to choosing the split to maximize the between-group sum-of-squares in a simple analysis of variance.

In their descriptive form, decision trees summarize a set of data, while in their predictive form, once the trees have been learnt form a data set, they can be used to predict a classification (or value) from a set of input parameters.

Missing data

Missing values are one of the curses of statistical models and analysis. Most procedures deal with them by refusing to deal with them – incomplete observations are tossed out. This approach is prohibitive when data are costly, which is the case for many experiments we are dealing with.

Recursive partitioning in CART is more ambitious, as any observation with values for the dependent variable and at least one independent variable will participate in the modeling. The impurity indices are calculated over the observations which are not missing a particular predictor, and the probabilities $p(L)$ and $p(R)$ are adjusted so that they sum to $p(T)$.

Bagging

The main drawbacks of decision trees are the sensitivity to outliers and the risk of overfitting.

To overcome overfitting, tree pruning and cross-validation is a well-known solution. However, we prefer the use of 'bagging' as it also improves the quality of predictions. Bagging on decision trees has been introduced for this purpose (cf. Breiman, 1998).

It consists of aggregating multiple trees. Each tree is obtained by drawing a bootstrap sample from the multinomial distribution with parameters n and $p1 = 1/n, \ldots,$ $pn = 1/n$. The variables selected for splitting in the various nodes are not necessarily the same. The number of times each variable is selected for the split is informative on the tree stability, and therefore on the confidence in the ranking of the variables for the descriptive summary.

If *nbag* bootstrap samples are drawn, then *nbag* decision trees are built. When using these trees for prediction purposes, for each set of input parameters, we get *nbag* estimates. The estimates are then averaged to give the final prediction. This methodology has been shown to improve the prediction by about 10%, when drawing 25 samples.

Results and discussion

This section is organized as follows. The results concerning the structure of the knowledge management system for durum wheat process data are first presented. Statistics and tree learning results are given, as well as the prediction results using decision trees. The use of expert rules for decision support is then introduced. Finally a method for confronting expert knowledge with experimental data is exposed. Different case studies from the durum wheat process are taken as examples in each case, in order to illustrate various aspects of the proposed approach.

Structure of durum wheat process data

The knowledge management system concerning 'transformation and qualities of durum wheat based food' was specifically built according to the description of processing steps and quality parameters classically used to produce and evaluate these foods.

- The 'technical' axis was designed considering all the unit operations (*29 unit operations*) that are involved in the transformations of durum wheat grains into semolina or flour (five unit operations; e.g., conditioning, dehulling, milling, etc.), of semolina or flour into durum-based products (18 unit operations; e.g., extrusion, sheeting, drying, etc.), and of durum-based product into food (six unit operations; e.g., packaging, sterilization, storage, etc.).

Table 1 Unit operations involved during processing of foods based on durum wheat

Unit operations	Food quality parameters
→ Characteristics of grains	→ Nutritional parameters (types, contents, and availability of the components)
Storage of grains	Proteins
Cleaning of grains	Mineral
Conditioning or tempering	Vitamin
Dehulling	Dietary fiber
Milling/fractionation	Polyphenols
Storage of flours and semolinas	Lipids
→ Characteristics of flours and semolinas	Mono and oligo saccharides
Parboiling	Starch
Addition of ingredients (in flour or semolina)	→ Organoleptic parameters
Hydration	Aspect parameters
Kneading	Color parameters
Mixing	Texture parameters
Agglomeration	Sensory parameters (mouthfeel)
Fermentation	→ Safety parameters (types and concentration of different contaminents)
Dough sheeting	Microorganisms
Drying	Mycotoxins
Expansion	Pesticids
Extrusion (low temperature)	
Extrusion – cooking	
Baking	
Cooking in water	
Steaming	
→ Characteristics of wheat-based food	
Association with other foodstuffs	
Packaging	
Sterilization, pasteurization	
Storage of wheat-based foods	

- The 'quality' axis was built taking into account all the quality parameters (*56 parameters*) that could be used to describe the durum wheat-based foods, on a nutritional level (16 parameters; e.g., protein content, vitamin content, etc.), organoleptic level (20 parameters; e.g., color, elasticity, mouthfeel, etc.), and safety level (10 parameters; e.g., mycotoxin content, microbial stability, etc.).

The identified unit operations and food quality parameters are shown in Table 1.

The knowledge management system was defined to handle the processing steps involved in the production of different durum wheat-based food. Six families of food

Table 2 (a) *Cooking in water* parameters available in experiments; (b) Other process parameters available in experiments for studying *Cooking in water*

Cooking in water parameters	
Temperature	(90–100) °C
% of salt in water	0 %
Time	(10–30) min.
Kind of water	Deionized water, distilled deionized, distilled water, tap water, unknown
Other process parameters (interaction parameters)	
Vitamin B_6 addition	(0–1.23) mg 100 g^{-1}
Thiamin addition	(0–1.74) mg 100 g^{-1}
Riboflavin addition	(0–1.95) mg 100 g^{-1}
Niacin addition	(0–11.84) mg 100 g^{-1}
Vitamin A addition	(0 0.33) mg 100 g^{-1}
Grain variety	Capeiti, creso, unknown
Drying kind of cycle	Ht, Ht-A, HT-B, LT, unknown
Drying duration	(10–85) h
Drying maximum temperature	(39–86) °C
Flour storage temperature	(4–40) °C
Flour storage duration	(0–6) months
Method	AACC method, microbiological method

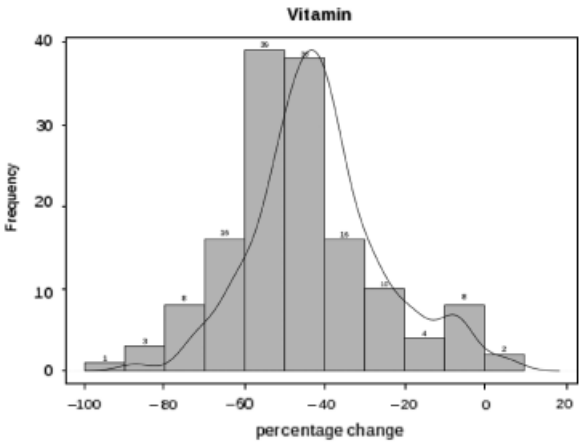

Figure 3 Distribution of vitamin decrease due to *cooking in water*.

products have been selected: pasta, couscous, flat bread, crackers, integer precooked grain (Ebly®) or precooked cracked grain (Bulgur). The processing comprises a specific series of unit operations that were defined for each product, and specific technical specification (i.e., settings for control parameters) for each unit operation involved in the processes.

Experimental data exploitation

To illustrate the decision tree approach, we present a case study about the impact of a unit operation on a product quality. The selected unit operation is *cooking in water for pasta products*, and we study its impact on one nutritional quality (vitamin content). One hundred and forty-five experimental results from 11 publications are available in the database. The experimental condition ranges that were investigated for each parameters of the cooking unit operation are given in Table 2. The distribution of 'vitamin decrease' is plotted on Figure 3, and the integration of results mainly demonstrates that cooking in water induces a decrease of vitamin content by 46% for pasta products. A detailed summary of results by subcomponent (i.e., the different types of vitamins) is given in Table 3.

Table 3 Summary of vitamin content decrease by subcomponent

Subcomponent	# results	Mean vitamin decrease	Vitamin decrease standard deviation	Vitamin decrease (min, max)
Folic acid (vitamin B_9)	3	−21.6	1.1	(−23, −21)
Niacin (vitamin B_3 or PP)	34	−51.0	15.9	(−99.5, −29)
Panthothenic acid (vitamin B_5)	9	−18.8	14.5	(−44, −3.7)
Riboflavin (vitamin B_2)	37	−51.4	13.6	(−73, −18.3)
Thiamin (vitamin B_1)	48	−52.8	12.4	(−83.7, −31)
Vitamin A	8	−6.6	9.2	(−19,8)
Vitamin B_6	6	−42.0	8.8	(−53, −28)

Table 4 Features most often selected by the bagging procedure

	First split	Second split	Third split
Component	100		
Times		52	22
Kind of water		26	16
Thiamin addition		17	17
Miscellaneous		5	45

Tree learning is based on these parameters, but also takes into consideration parameters of previous unit operations involved in the pasta process (before cooking in water), such as vitamin addition, drying parameters or storage conditions. Those so called *interaction* parameters are summarized at the bottom of Table 2.

As explained in section "Materials and methods," a bagging procedure (with 100 samples, as the trees are not very stable) is used to generate a family of trees. The number of times that each factor is selected for the first two splits is indicated in Table 4. When analyzing the results presented in Table 4, the most discriminating important factor appears as perfectly stable and corresponds to the type of vitamins. The remaining factors are less stable, the most important ones

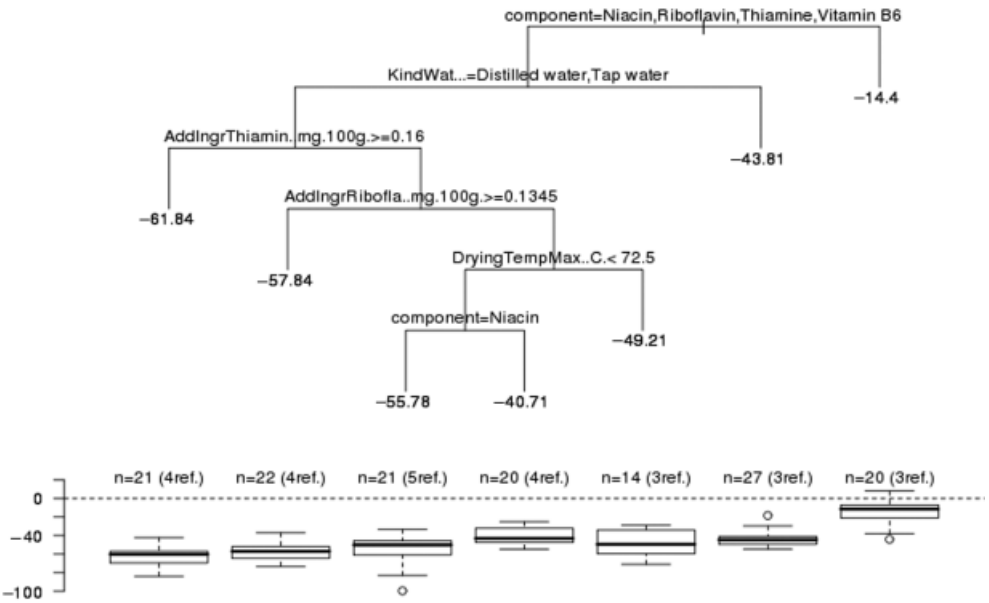

Figure 4 The descriptive decision tree learnt from the whole *cooking in water* data set.

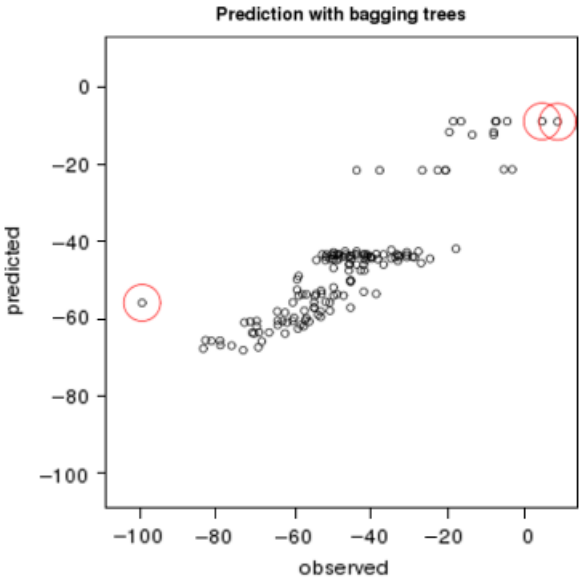

Figure 5 Prediction results versus observed values for the *cooking in water* data set obtained with tree bagging.

being cooking time, type of water and preliminary thiamin addition.

As a typical example, Figure 4 shows one tree among the family of trees obtained through the bagging procedure. At each node, the splitting variable is shown with the set of values leading to the left branch. The terminal leaves are labeled with the average value of the dependent variable (*decrease in vitamin content*) for the examples assigned to each of them. At the bottom of the figure, a boxplot is drawn below each leaf to show the distribution of values of the dependent variable within the leaf. The number of related experiments and references is indicated above each boxplot.

The family of trees can also be used to predict the impact of unit operation parameters on product qualities (e.g., vitamin content). The quality of prediction is shown in Figure 5, where predicted values are plotted versus observed values. The general trends are rendered within a wide band, the outliers (extreme observed values, circled in Figure 5) are more difficult to predict.

We investigated further for one specific component among the vitamin (thiamin) for which enough data are available to generate a tree. Figure 6 displays the descriptive tree obtained from the data set reduced to the impact of cooking in water on the *thiamin* content (48 values), which shows the importance of the analytical method and of *thiamin* addition on the response (i.e., the residual *thiamin* content).

The bagging procedure can also be used for predicting the impact on vitamin content of the *cooking in water* process, using a set of parameters different from the ones used in the experimental data sets, but within the same range.

Table 5 shows the predicted values for four sets of parameters. As expected, the model indicates a decrease in vitamin content, and highlights the effects of parameters.

Expert knowledge exploitation

The case study that we present to illustrate the use of expert rules concerns the impact of drying on the color and texture

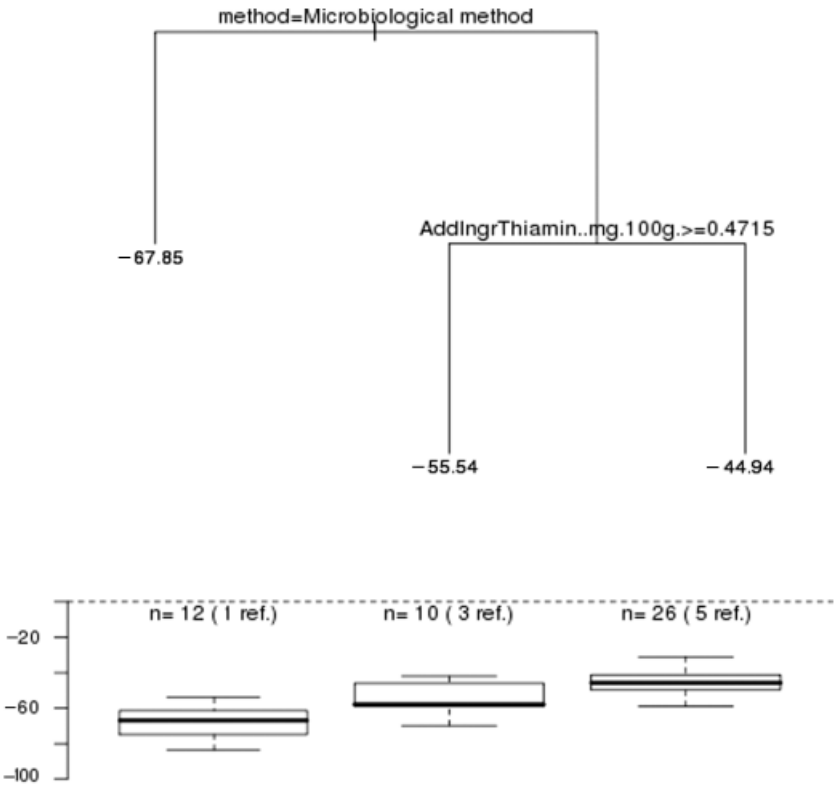

Figure 6 The descriptive decision tree learnt from the *cooking in water* data set reduced to the *thiamin* component.

Table 5 Predicted values obtained by the bagging procedure for various parameters

Component	Temperature (°C)	Time (min)	Kind of water	Vitamin A addition	Thiamin addition	Vitamin content decrease
Riboflavin	90	20	Distilled water	0.33	0	− 54.3
Riboflavin	90	10	Distilled water	0.33	0	− 47.6
Thiamin	100	12	Tap water	0	0.9	− 59.2
Thiamin	100	12	Tap water	0	0	− 53.6

of pasta products. The impact of drying, beyond a given temperature (about 60–65 °C), and depending on relative humidity conditions, is known to rely on the combination of several molecular phenomena, namely: (i) the inhibition of enzymatic activity; (ii) starch gelatinization; (iii) protein insolubilization.

Without giving a complete description of the case study, we will focus on the following rules expressed by domain experts. The representation of rule 6 in the conceptual graph model is shown as an example in Figure 2a.

Yellow color

Rule 1 High-temperature drying inactivates lipoxygenase

Rule 2 Lipoxygenase oxydates carotenoids

Rule 3 Carotenoids are responsible for the yellow color of pasta

Rule 4 Oxydating carotenoids reduces the yellow color of pasta

Brown color

Rule 5 High-temperature drying, if applied at the beginning of cycle, inactivates peroxydase

Rule 6 Peroxydase is responsible for the brown color of pasta

Rule 7 The brown color hides the yellow color of pasta

Reddish color

Rule 8 High-temperature drying, if applied at the end of cycle, induces a reddish color

Rule 9 Reddish color is due to Maillard reaction

Rule 10 The reddish color hides the yellow color of pasta

Texture

Rule 11 High-temperature drying causes protein insolubilization

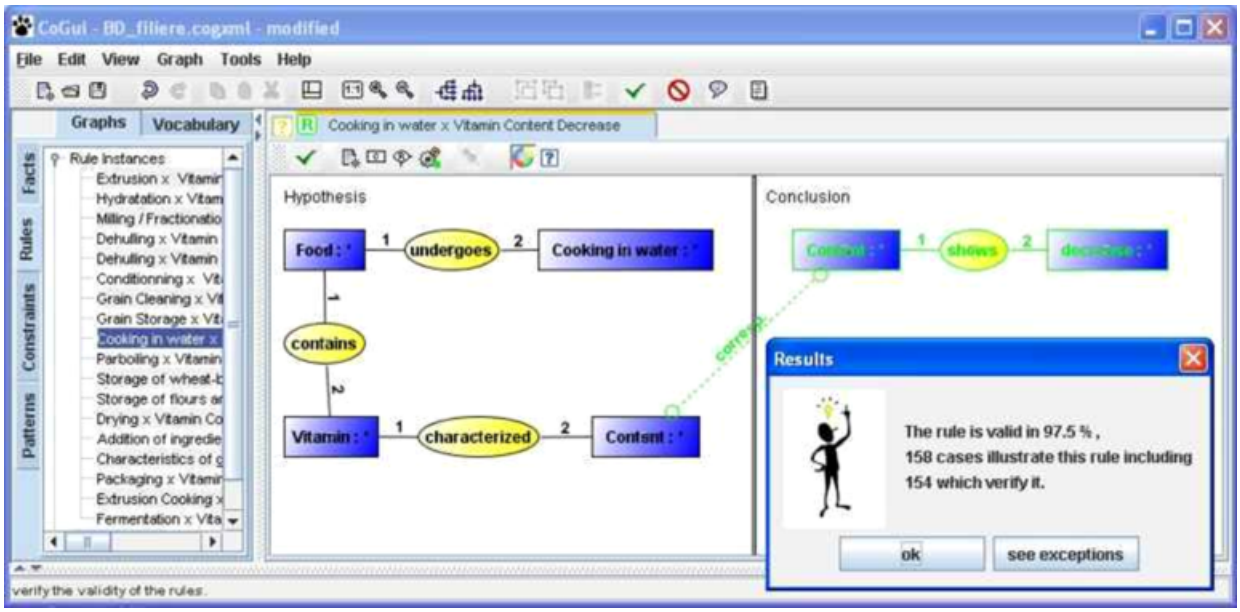

Figure 7 Evaluation of the validity of an expert rule.

Rule 12 Protein insolubilization is responsible for pasta firmness

Stickiness

Rule 13 High-temperature drying, if applied when humidity is relatively high (> 15%), modifies the state of starch by causing its partial gelatinization

Rule 14 Below 15% humidity, starch remains stable even under high temperature

Rule 15 The beginning of drying cycle is characterized by high humidity

Rule 16 The end of drying cycle is characterized by low humidity

Rule 17 Starch partial gelatinization leads to sticky pasta

The rules expressed in the conceptual graph model can be exploited for two uses:

• Use for *prediction*. Starting from given process conditions, described by a conceptual graph, all the rules whose hypotheses are more general can be applied in 'forward' chaining, producing a final conceptual graph that represents the expected result. As explained in section "Materials and methods," the rule represented in Figure 2a can be applied to the conceptual graph of Figure 2b, which represents the following information: 'the late end-of-cycle high temperature drying D1 is applied on the pasta product P1, in which peroxydase is active, and produces as an output the pasta product P2.' Therefore we can infer the conclusion of the rule, which leads to the conceptual graph of Figure 2c, indicating that the output pasta product P2 is characterized by a brown color.

• Use for *reverse engineering*. Starting from wanted end-product properties, described by a conceptual graph, all the rules whose conclusions are compatible with this graph can be applied in 'backward' chaining, suggesting possible process conditions to reach the expected properties. For instance, if the objective for the end product is to avoid a sticky texture, applying the rule 17 in backward chaining allows concluding that starch grain partial gelatinization must be avoided. Then the application of the rule 13 in backward chaining indicates that high-temperature drying must not be associated with > 15% humidity.

Confrontation of expert knowledge with experimental data

We have proposed a method that allows one to test the validity of expert rules within the experimental data (Thomopoulos, 2008). When an expert rule appears not to be valid, a second-step objective is to refine it in order to take into account the exceptions.

Searching for exceptions to expert rules within experimental data

The objective is to automatically test whether the expert knowledge expressed as conceptual graph rules is valid within the experimental data of the relational database. A

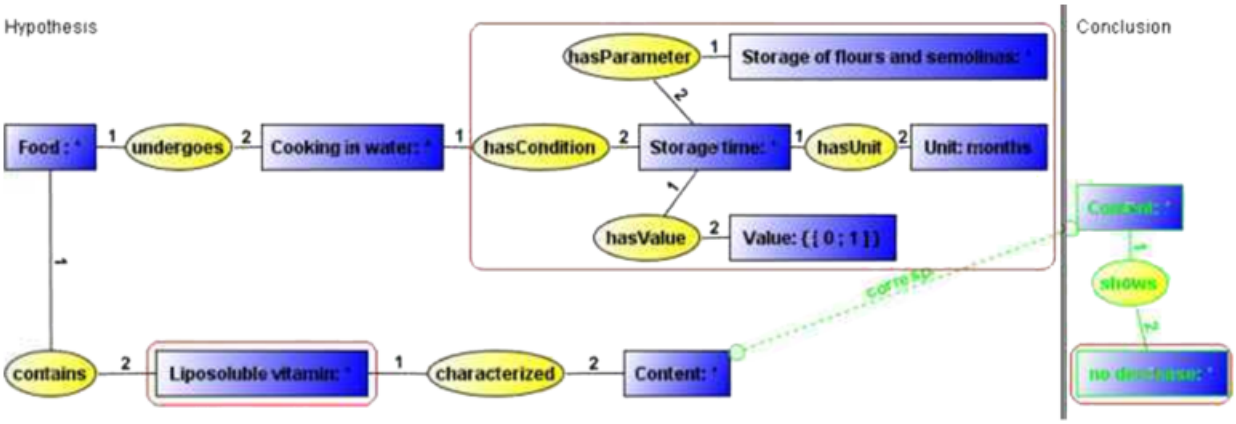

Figure 8 The rule of Figure 7 after the negation, specialization and completion steps.

confidence rate is computed for the tested rule and the data that constitute exceptions to the rule are identified and can be visualized by the user. For example the confidence τ of the rule of Figure 7 is equal to 97.5%. Confidence spread out from 73% to 100% in the application.

Using exceptions to build new rules

The aim of this part is to learn new rules corresponding to the exceptions identified in the previous stage. Three steps are proposed to achieve this objective, respectively, *negation*, *specialization* and *completion*.

After the *negation* step, the rule instance of Figure 7 is changed into: 'if a food product undergoes cooking in water, then its vitamin content does *not* decrease' (it may increase or stagnate in the exceptions to the rule). After the *specialization* step, the rule of Figure 7 has its 'vitamin' vertex specialized into 'liposoluble vitamin.' After the three steps (*negation*, *specialization* and *completion*), the rule of Figure 7 produces the rule of Figure 8. It represents the information: 'If a food product undergoes cooking in water, with a storage time of flours and semolina lower than 1 month, then its liposoluble vitamin content does not decrease.'

Conclusion and perspectives

To help food companies meet the demands of safe, healthy and tasty foods, there is a need to merge different research disciplines into a comprehensive effort dedicated to developing the new tools and decision support systems.

The approach presented above has a high potential to succeed in integrating various data and knowledge. Our specific original contributions are

- the design of a hybrid system combining both data and knowledge;
- the advantage of not requiring an *a priori* model, and therefore, an increased genericity;
- the potential use for both risk and benefit analysis.

Furthermore, the building of the system led to an in-depth formalization of the durum wheat chain, which can be reused for other purposes within the chain, and can also be transferred to other domains.

The work presented in this paper could be seen as a starting point to integrate new knowledge from multidisciplinary fields, in order to facilitate knowledge transfer to the food industry, with a general aim of improving existing products and developing new ones. In this way, the project constitutes a tool for structuring and promoting the co-operation between science, practice and application in the area of cereal processing.

References

Baranyi J., Tamplin M.L. (2003) A new international effort for the advance of predictive microbiology, ComBase: Combined database of microbial responses to food environments. In: *Proceedings of the International Conference on Predictive Modelling in Food (ICPMF'03)*, Quimper, France, pp. 50–51.

Bos C., Botella B., Vanheeghe P. (1997) Modeling and simulating human behaviors with conceptual graphs. In: *LNAI: Vol. 1257. Conceptual Structures: Fulfilling Peirce's Dream* eds Lukose D. pp. 275–289, Springer, Heidelberg, Germany.

Breiman L. (1984) *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton.

Breiman L. (1998) Arcing classifiers. *The Annals of Statistics*, **26**, 801–824.

Codd E.F. (1970) A relational model of data for large shared data banks. *Communications of the ACM*, **13**, 377–387.

Haemmerlé O., Buche P., Thomopoulos R. (2007) The MIEL system: uniform interrogation of structured and weakly-structured imprecise data. *Journal of Intelligent Information Systems*, **29**, 279–304, Springer.

Ihaka R., Gentleman R. (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5** (3), 299–314.

Mueangdee N., Mabille F., Thomopoulos R., Abécassis J. (2006) Virtual Grain: a data warehouse for mesh grid representation of cereal grain properties. In: *Proceedings of the 9th European Conference on Food Industry and Statistics, Agrostat'2006,* Montpellier, pp. 291–299.

Mugnier M.L. (2000) Knowledge representation and reasoning based on graph homomorphism. In: *LNAI: Vol 1867. Conceptual Structures: Logical, Linguistic, and Computational Issues* eds. Ganter B., Mineau G.W. pp. 172–192, Springer, Heidelberg, Germany.

Peck M.W., Roberts T.A., Sutherland J.P., Walker S.J. (1994) Modelling the growth, survival and death of microorganisms in foods: the UK Food Micromodel approach. *International Journal of Food Microbiology*, **23**, 265–275.

Salvat E., Mugnier M.L. (1996) Sound and complete forward and backward chaining of graph rules. In: *LNAI: Vol. 1115. Conceptual Structures: Knowledge Representations as Interlingua* ed Eklund P.W. pp. 248–262, Springer, Heidelberg, Germany.

Sowa J.F. (1984) *Conceptual Structures – Information Processing in Mind and Machine*. Addison-Welsey, Reading, MA.

Thomopoulos R. (2008) Learning exceptions to refine a domain expertise. In: *Encyclopedia of Data Warehousing and Mining – 2 nd Edition, August 2008* ed Wang J. pp. 1129–1136, Information Science Reference, Hershey, PA, USA.

Young L.S. (2007) *Application of Baking Knowledge in Software Systems. In Technology of Breadmaking – 2nd edition*. Springer, Berlin, New York, US: pp. 207–222.

Young L.S., Cauvain S.P., Davies P.R. (2001) Rise again, fair knowledge. *In Applications and Innovations in Intelligent Systems IX: Proceedings of ES2001, the Twenty-first SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence,* Cambridge, Springer, pp. 89–99.

Zwietering M.H., Wijtzes T., de Wit J.C., Van't Riet K. (1992) A decision support system for prediction of the microbial spoilage in foods. *Journal of Food Protection*, **55**, 973–979.

