

**CONTRIBUTION OF XTRACTIS<sup>®</sup> METHODOLOGY TO THE  
AUTOMATIC EXTRACTION OF ROBUST FUZZY MODELS.**

**Application to the prediction of consumer liking and sensory  
evaluation, and to the optimization of product formulation.**

Zyed ZALILA, Julien CUQUEMELLE, Arezki CHIKH, Cédric PENET,  
Benjamin LORENTZ, Dimitri DESCHAMPS

*Proceedings AgroStat 2008, 10<sup>th</sup> European Symposium on Statistical Methods for the Food  
Industry, Louvain-la-Neuve, Belgium, January 22-25, 187-199.*

# **Contribution de la méthodologie xtractis<sup>®</sup> à l'extraction automatique de modèles flous robustes. Application à la prédiction de préférence de consommateurs et d'évaluation sensorielle et à l'optimisation de formulation de produit.**

## **Contribution of xtractis<sup>®</sup> methodology to the automatic extraction of robust fuzzy models. Application to the prediction of consumer liking and sensory evaluation, and to the optimization of product formulation.**

Z. Zalila<sup>1,2</sup>, J. Cuquemelle<sup>1</sup>, A. Chikh<sup>1</sup>, C. Penet<sup>1</sup>, B. Lorentz<sup>1</sup>, D. Deschamps<sup>1</sup>

<sup>1</sup> *intellitech [intelligent technologies] – 14 rue du Fonds Pernant – 60200 Compiègne – France*  
E-mail: [zyed.zalila@intellitech.fr](mailto:zyed.zalila@intellitech.fr) – tel: +33 3 44 23 48 90 – fax: + 33 3 44 23 48 99

<sup>2</sup> *University of Technology of Compiègne (Costech) – 60203 Compiègne – France*

### **Résumé**

Les systèmes d'inférence floue permettent de modéliser facilement et intuitivement tout processus de prise de décision sous la forme d'une fonction multidimensionnelle déterministe, implicitement définie par des règles linguistiques. Ces règles floues sont habituellement construites à partir d'une expression de la connaissance préexistante sur le processus étudié.

Cependant, dans de nombreuses situations telle l'analyse sensorielle, aucune expertise explicite à propos du processus n'est disponible. Dans ces cas, des données entrées / sorties mesurées sur le processus représentent la seule information disponible. A partir de ces données, nous proposons d'induire automatiquement les règles floues de décision adéquates, pour représenter le processus d'évaluation subjective le plus fidèlement possible, ce qui impose d'évaluer systématiquement la capacité de généralisation des modèles générés. Ceci permet d'éviter les situations classiques d'overfitting (apprentissage par cœur).

Pour cette étude, l'analyse de tomates fraîches illustre l'efficacité de l'application de la modélisation floue automatique xtractis<sup>®</sup> à l'ingénierie sensorielle.

**Mots-clés :** induction de règles – apprentissage automatique – modélisation floue non linéaire – extraction de connaissances – optimisation non linéaire – ingénierie sensorielle – capacité de généralisation – validation de modèles – stratégie d'apprentissage – overfitting – validation croisée

### **Abstract**

Fuzzy inference systems allow to easily and intuitively model any decision making process as a deterministic multidimensional function implicitly defined by linguistic rules. These fuzzy rules are classically built thanks to a prior expression of knowledge about the process under study.

However, in a number of situations and particularly in sensory analysis, no explicit expertise on the process is available. In these cases input / outputs databases collected

from the process are the only available information. From this data, we propose to automatically induce suitable fuzzy rules to represent the subjective evaluation process as faithfully as possible, which imposes to systematically evaluate the generalization capacity of the generated models. This avoids classical situations of overfitting (learning by heart).

In this study, fresh tomatoes analysis illustrates the effective application of **xtractis**<sup>®</sup> automatic fuzzy modelling to sensory engineering.

**Keywords:** rule induction – machine learning – fuzzy non linear modelling – knowledge extraction – fuzzy non linear optimization – sensory engineering – generalization capacity – model validity – learning strategy – overfitting – cross validation

## 1. Introduction

This paper proposes to show how techniques not widespread in the sensory engineering community can be effectively used to solve practical and common sensory issues.

We will first present the main specificities of the fuzzy theory (derived from the symbolic approach of artificial intelligence) that explain the great benefits of the use of fuzzy models to solve real life problems, especially when human (subjective) evaluation is involved.

Then, according to recent advances in the machine learning theory, we will expose how learning algorithms (that can be considered as a generalization of analytical regression) can be evaluated in terms of generalization capacity to ensure that a model is actually representative of a process, and is not only valid on a specific learning sample.

Thanks to a real world case concerning the sensory evaluation of fresh tomatoes, we will expose how the **xtractis**<sup>®</sup> model extraction method we have developed, based on both above-mentioned theories, allows to extract efficient and robust models from the available databases.

These models being both interpretable (thanks to a linguistic structure) and predictive (thanks to the implicit definition of a numerical function), they can solve a large variety of sensory tasks.

## 2. Main concepts and capabilities of fuzzy modelling

### 2.1 Imprecision and uncertainty

Fuzzy mathematics, also known generically as fuzzy theory, includes a number of theories which are generalizations or extensions of their classic equivalents: thus the theory of fuzzy subsets is an extension of set theory, fuzzy logic is an extension of binary logic, the theory of fuzzy quantities is an extension of number and interval theories, possibility theory extends probability theory. All these theories offer formally rigorous concepts, techniques and methods for collecting, representing and analysing fuzzy data. The specificity of fuzzy data is that it is imprecise, uncertain and subjective and these three main characteristics often co-exist. A whole range of other synonyms fall under this notion of fuzziness, such as poorly specified, imperfect, vague, qualitative, linguistic, partial or approximate.

Because fuzzy logic is nuanced and gradual, it more closely approaches human logic allowing the measure of possibility to become an accurate replacement for the measure of probability, when the available information is sparse and/or of poor quality [Dubois & Prade, 1994].

From a formal point of view, the fuzzy theory can be considered as an interface between qualitative data or symbolic concepts, and quantitative values. Thus, this natural capacity of the fuzzy theory to handle heterogeneous data (quantitative / qualitative, precise / vague, continuous / discrete) makes it particularly suitable to handle real life problems and especially sensory engineering tasks.

## 2.2 Knowledge modelling, fuzzy model

The objective of modelling is to obtain a formal model which describes a natural, human or industrial process to understand it better, and/or with a view to predicting the effects of the process. Unlike other approaches that can be difficult to interpret in a qualitative point of view, a fuzzy system is a collection of linguistic rules<sup>1</sup> that can be intuitively interpreted. For example:

Rule 1: « if [CLime] is *high* and [Csugar] is *low* then [Acidity] is *strong* ».

Rule 2: « if [CLime] is *low* and [Csugar] is *high* then [Acidity] is *mild* ».

*low* and *high* are fuzzy subsets on a quantitative scale characterizing the fact that a concentration is considered as low or high

*strong* and *mild* are characterizations of the intensity of the acidity as if evaluated by an assessor.

Thanks to adequate schemas of approximate reasoning [Zadeh, 1975] [Zalila, 1993], a collection of fuzzy rules implicitly defines a multidimensional non linear function, which actually allows a fuzzy model to merge the benefits of qualitative and quantitative approaches. As fuzzy systems are universal approximators of non-linear functions [Kosko, 1992] connecting output variables to observed variables, they can in theory model any complex non-linear process.

However, the classical fuzzy modelling approach may not be used “as is” in sensory engineering, as it needs explicit *a priori* knowledge of the process behaviour to be translated into fuzzy rules [Zadeh, 1973]. This explains why despite all the benefits of this theory, it is not yet widely used in areas where only input-output databases are available. This has led us to develop a comprehensive method<sup>2</sup> for automatic extraction of robust fuzzy models from data.

## 3. Automatic extraction of models

### 3.1 Data types

Databases used for model extraction are divided into three different types: Objective data (**O**) - analysis results or physico-chemical measurements, demographic or financial data ... -, Subjective data (**S**) - consumer liking - and Subjective Objectivised data (**SO**) given by human experts - sensory panels - that represent the most objective evaluation of subjective attributes.

When possible, we recommend using non aggregated repetitions of measures (individual assessors' estimations and repetitions, noisy sensors repetition) for the learning process, which will constraint the model to take this variability into account. This improves the robustness of the generated models, as they must be accurate and robust on all individual learning points and not only on their aggregation.

---

<sup>1</sup> According to the model “if *Premise* then *Conclusion*” derived from cognitive psychology.

<sup>2</sup> This method and other techniques to exploit fuzzy models are implemented in the **xtractis**<sup>®</sup> software suite.

An interesting property of fuzzy systems is their ability to perform a prediction even with missing inputs. This property allows us to get rid of absurd points or to take account of missing information, without the need to replace these missing values by estimations<sup>3</sup>.

## 3.2 Learning from data

### 3.2.1 Learning Strategy

A learning strategy defines a method to process a learning sample (provided by a series of experiments on the process to be modelled) to induce a model that is accurate enough on this learning sample. An example of a classical learning strategy might be: perform a PCA on the input space, keep the 2 main axes, and perform a quadratic regression to identify a relation between the 2 main axes and the output.

More generally speaking, we define a learning strategy as a choice of:

- A model structure, including the definition of its input space (list of variables, number of inertia axes), and of its form (parametric function, number of fuzzy rules, complexity of a neural net ...);
- An error measurement (quadratic error, maximal absolute error, correlation ...);
- An algorithm that identifies optimal parameters for the model structure according to the error measure on the learning sample.

### 3.2.2 Generalization capacity of models

The learning problem consists in automatically identifying a model, or a transfer function, relating inputs and outputs to represent the process as faithfully as possible. To achieve this, one defines a learning strategy that is applied to a (usually) small learning sample, and hopes that the generated model that is accurate on the learning sample actually models the underlying process.

However, as model identification is an ill-posed problem [Hadamart, 1902], an infinity of significantly different models may accurately approximate a finite learning sample. Moreover, given a dataset without any prior knowledge about the process to model, it is impossible to *a priori* define the learning strategy yielding the best model, as there is no single algorithm which is the best for any given field and any given problem [Jain & Mao, 1997].

The study of generalization capacity (also referred to as “robustness” hereafter) aims at selecting a learning strategy that will induce a model that actually represents the process. Conversely, it must discard all the strategies that only learn the sample by heart and which lead to models qualified as “overfitted”, as they usually lose their generalization capacity because of fitting the learning sample too closely.

### 3.2.3 Regulation methods

Regulation methods are means of constraining a learning strategy to make it less prone to over fit a data set. To achieve this goal, three main ways can be used:

- reduce the complexity of the model structure (zeroing some parameters of a function, merging similar fuzzy classes, deleting links in a neural network);
- constrain the learning algorithm to generate models following specified criteria (boundaries on the model output variation range, on first or second derivative values, neural network weight decay);
- perturb the learning algorithm so that it does not reach a minimal error on the learning set (early stopping, noise injection).

---

<sup>3</sup> Missing values are usually replaced by imputations, computed from information contained in the database. We consider that this approach may introduce false information in the database.

The use of regulation methods might surprise as they lead to a sub-optimal accuracy of the predictions on the learning sample, but the goal of these methods is precisely to avoid the model too closely fitting the learning points in order to improve its robustness.

However, these methods give no assessment about the actual performance of the generated models and hence do not allow to efficiently choose the best learning strategy for a given data sample. In fact, as these methods aim at constraining the learning algorithm, they must be considered as part of the generation strategy which has to be evaluated in terms of generalization capacity.

### 3.3 Model selection

#### 3.3.1 Insufficiency of a unique validation sample

In order to assess the robustness of a model, one usually relies on computing predictions on unknown points to check if the model is able to correctly predict points not involved during model generation. However, defining a fixed set of validation points as robustness estimator has two major drawbacks when the amount of data is low regarding the complexity of the process:

- Firstly, the initial choice of the validation sample has a strong influence on the outcome of the robustness evaluation. This is illustrated by the plot of Figure 1 which shows the evaluation of the same learning strategy on 200 different data splits<sup>4</sup>. Thus, using a fixed learning sample to validate a model is akin to draw conclusions based on one randomly chosen experiment among the 200 represented here (with a correlation ranging from 0.2 to 0.92).

- Secondly, the whole available data is neither used for learning (which may actually prevent to model the process), nor for validation (which adds more uncertainty to the validity of this robustness estimation).

Cross Validation methods avoid these two major drawbacks, yet at the price of a substantial computational cost.

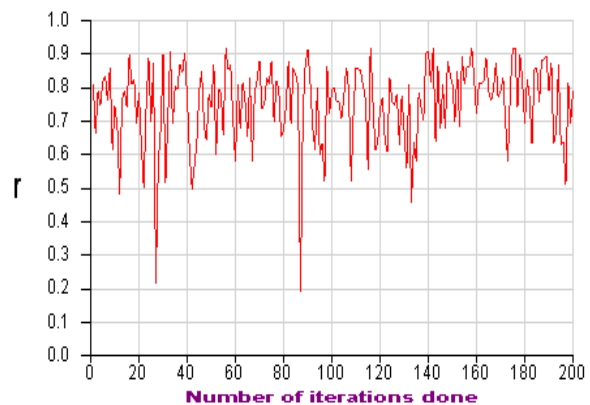


Figure 1: Variability due to the choice of a learning sample

#### 3.3.2 Cross validation methods

Among the several available cross validation methods (referred to as “CV” hereafter) [Plutowski, 1996], the Monte Carlo CV gives the best compromise between low variance and computational load in the estimation of the model robustness [Zalila, Cuquemelle et al., 2007b].

This method basically consists in aggregating the outcome of the learning strategy applied on several splits instead of considering only one split to assess the validity of the model. We propose to gather all the predictions done with the different validation splits then to calculate the correlation actual values / predicted values only during the last step on the whole scatter of points [Zalila, Cuquemelle et al.,

---

<sup>4</sup> >From the complete data sample (17 points), 200 different splits into a learning set (70% of data, 11 points) and a validation set (30% of data, 6 points) are randomly generated. For each split, the same learning strategy is applied to the learning set and the resulting model is evaluated with the validation set. The plot shows the correlation between actual and predicted values on the validation set for each of the 200 experiments.

2007a]. This delayed aggregation assures the convergence of the estimated robustness of the model after a sufficient number of iterations and a better accuracy of this robustness as more information is kept for its evaluation.

Figure 2 shows the convergence of the CV robustness analysis (for the same learning strategy and data sample used in Figure 1), as the aggregation of the outcome of each split is iteratively computed. The estimate that assesses the robustness of the model is reached at the convergence of the process.

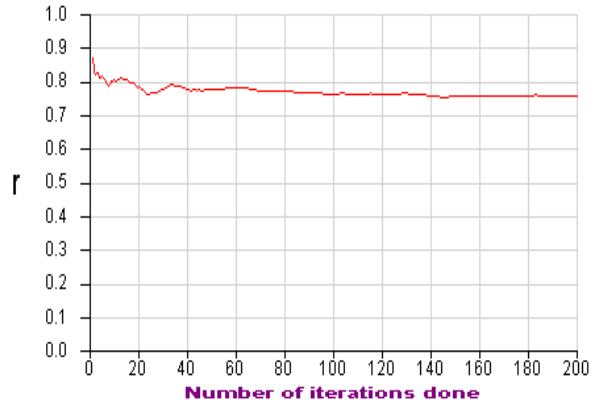


Figure 2: Convergence of the Monte Carlo CV

The whole available data is thus used to validate the learning strategy, and as this estimate is computed on a large number of subsamples, its variance is greatly reduced over a computation on a single validation sample.

For a whole available i.i.d sample of size  $N$ , performing the CV with validation samples of size  $P$  gives an unbiased estimation of the robustness of the learning strategy applied to learning samples of size  $N-P$  [Grandvalet, 2001]. According to [Vapnik, 2000], for a given learning strategy, the robustness of models generated with a sample of size  $N$  is statistically better than the robustness of models generated with samples of size  $N-P$ . From these both statements, it can be derived that the bigger amount of points is used for the validation samples, the more pessimistic the Monte Carlo CV is to evaluate the strategy applied to the full sample of size  $N$ . Hence, once a learning strategy has been evaluated as robust according to CV, it can then be applied to learn on the whole available data without risk of overfitting (in a statistical sense).

### 3.3.3 Generalized heuristic for model extraction

As there is no objective way to relate the characteristics of a given learning strategy to the generalization capacity of the model it generates, it is impossible to choose *a priori* what learning strategy will give the best model. Given a dataset, in order to get the best possible model, there is no other option than evaluating a large number of different learning strategies and select the best one.

We propose a multi-criteria selection process based on data-driven performance indexes as given by CV, as well as the complexity of the models that are compared. In fact, between several models of similar robustness, one should choose the simplest according to regulation principles (*cf.* 3.2.3).

Moreover, thanks to the linguistic structure of fuzzy models, the modeller may use his knowledge about the process to select models. Actually, to maximize the benefit of this property, **xtractis**<sup>®</sup> fuzzy models only use original variables available in the database. The input space simplification is only done by variable selection, without any transformation (as done by a PCA or a PLS)<sup>5</sup>.

<sup>5</sup> A common pitfall of relying on a transformed input space for model extraction is to forget that as the computation of the new axes is data-driven, it suffers the same risk of overfitting as model extraction. As such, the stability of such transformations should be studied [Daudin, Duby et al., 1989], or they should be considered as part of a learning strategy evaluated by CV, so that the estimation of robustness also includes the input space transformation.

## 4 Application to fresh tomatoes evaluation

In this paper, fresh tomatoes are used to illustrate the effective application of fuzzy modelling to sensory engineering, starting from the Objective (**O**) data set (15 physical and chemical variables), the Subjective Objectivised (**SO**) data set (11 sensory descriptors evaluated by 14 trained assessors with 2 sessions for each measure) and the Subjective (**S**) data set (379 consumers, each evaluating his liking of 10 tomatoes among the 17 types available in the set), provided by the Centre Technique des Fruits et Légumes (CFIFL) and the Institut National de la Recherche Agronomique (INRA) in France. This benchmark reference base was used by all the contributors to the Consumer Segmentation and Key Drivers Analysis Workshop, held at Sensometrics 2004 (Davis, CA, USA).

4 consumer segments have been identified by performing segmentation on the liking information only. As for any **xtractis**<sup>®</sup> study, each of these segments is modelled independently as each segment usually have specific liking behaviours (and might consequently respond to specific sets of drivers).

### 4.2 Modelling results

In the following presentation of modelling results, when talking about “best” (or “top-”) models, we are referring to models with the best generalization capacity among all evaluated models (usually several thousands for each segment). This must be taken into account when comparing these results with studies for which no robustness evaluation has been performed, as it is basically simple, on this dataset, to generate models with a close to perfect accuracy on the learning sample, but with a very poor generalization capacity.

For each model, the error is evaluated in three ways (Table 1). On the first line of each model performance report lies the accuracy, *i.e.* the performance of the model in predicting the learning sample. The second line gives the results of a Monte Carlo CV using 15% of the database for the validation samples (3 points), and the third line represents a Monte Carlo CV with 30% of data (6 points). Each Monte Carlo computation has been performed 200 times, which is sufficient to reach convergence. Hence, for each analysis, the generation strategy is applied 200 times on different samples, each of the 200 generated models giving predictions on 3 (respectively 6) unknown points representing 15% (resp. 30%) of the database, thus giving a final scatter of 600 (resp. 1200) validation points.

The Hamming distance is the mean of absolute prediction errors. The percentages are related to the definition range of the output.

#### 4.2.1 Best results achieved by **O/SO → S** models over **O→S** or **SO→S** models

The classical approach to model the liking of consumer is to identify a relationship between sensory descriptors and consumer liking (**SO-S** model). This has been performed here, with good results reported in Table 1.

We have also tried to make a direct prediction of the liking using the instrumental measures only, but the performance of the **O-S** generated models is significantly lower than the **SO-S** models and thus is not represented in this paper.

However, only relying on the panel evaluations to predict the liking of a product means following common prior assumptions stating that only sensory descriptors can explain the liking, and that the sensory profile is complete, which is not always the case and anyway difficult to prove. As **xtractis**<sup>®</sup>

proposes to automatically identify relevant variables in the input space, we have performed a model extraction to predict the liking from both sensory and instrumental measures.

Output	Nb. of		Variables	Error				Remarque
	var.	rules		Correlation	Mean	Hamming	Max.	
Seg 1	4	2	Firm_inside	0.984	0.102	3.57 %	10.90 %	Gen. 1568
			Juicy	0.954	0.085	5.55 %	23.65 %	200 x MC 15 %, all points
			Melty Tomato_flavor	0.943	0.109	6.12 %	32.75 %	200 x MC 30 %, all points
Seg 2	5	3	Firm_inside	0.909	-0.04	5.03 %	14.39 %	Gen. 2144
			Tomato_odor	0.810	-0.10	7.42 %	30.10 %	200 x MC 15 %, all points
			Firm Mealy Tomato_flavor	0.760	-0.11	8.44 %	38.60 %	200 x MC 30 %, all points
Seg 3	2	2	Juicy	0.897	-0.06	7.2%	31.73%	Gen. 1325
			Skin width	0.874	-0.12	8.77%	32.73%	200 x MC 15 %, all points
				0.808	-0.19	10.42%	42.31%	200 x MC 30 %, all points
Seg 4	3	3	Firm_inside	0.964	0.034	5.33 %	11.70 %	Gen. 1591
			Firm	0.867	-0.05	10.16 %	40.00 %	200 x MC 15 %, all points
			Skin_width	0.747	0.100	13.58 %	57.23 %	200 x MC 30 %, all points

Table 1: SO-S models performance

Output	Nb. of		Variables	Error				Remarque
	var.	rules		Correlation	Mean	Hamming	Max.	
Seg 1	6	2	Average_weight	0.999	0.001	0.81 %	2.94 %	Gen. 697
			G_Total_acidity	0.971	0.005	4.40 %	19.02 %	200 x MC 15 %, all points
			Firm_inside Juicy Sweet Tomato_flavor	0.938	0.012	5.88 %	37.97 %	200 x MC 30 %, all points
Seg 2	4	2	G_Total_acidity	0.946	0.052	4.37 %	10.46 %	Gen. 4445
			Ext_color	0.802	0.107	8.14 %	23.65 %	200 x MC 15 %, all points
			Firm_inside Mealy	0.786	0.051	8.44 %	28.08 %	200 x MC 30 %, all points
Seg 4	3	2	M_Total_acidity	0.958	-0.16	6.41 %	14.03 %	Gen. 9050
			G_Sum_of_sugars	0.830	-0.11	11.07 %	43.44 %	200 x MC 15 %, all points
			Mealy	0.784	0.034	11.81 %	62.54 %	200 x MC 30 %, all points

Table 2: O/SO-S models performance

The results of this extended study are summarized in the Table 2. For segment 1, a more robust model but using 6 variables instead of 4 has been generated. For segment 2, a model of comparable robustness has been extracted, but with the benefit of a simpler structure (4 variables instead of 5<sup>6</sup>, and only 2 rules instead of 3). For segment 3, no particular improvement was made over the SO-S model, so no result is shown. The benefit of the O/SO-S approach is also clear for the modelling of segment 4:

<sup>6</sup> This is all the more interesting as 1 of these 4 variables is an instrumental measurement. Hence, the measurement of 3 sensory attributes is sufficient for this O/SO-S model instead of 5 for the SO-S version.

the extracted model provides similar robustness indexes (MC15% slightly lower and higher MC30%) but a simpler structure (2 rules instead of 3).

This ability to extract models with no *a priori* explicit knowledge and the possibility to process large databases with several hundreds of inputs allows to efficiently discover decision-making strategies that were not obvious even for experts of the considered field [Zalila, Davodeau et al., 2005].

#### 4.2.2 O→SO modelling results

As *xtractis*<sup>®</sup> is able to extract models from any input-output database, it is also possible to try extracting models that would predict the sensory profile of the tomatoes from the instrumental measurement.

On this study, the results strongly depend on the sensory descriptors we try to model. Some of the descriptors (“Firm\_inside”, “Skin\_width”, “Sweet”, “Acidity”) are represented by very robust models. For other descriptors (“Ext\_color”, “Tomato\_odor”, “Juicy”, “Mealy”, “Tomato\_flavor”), the best models have a lower robustness that questions their relevance<sup>7</sup>. Finally, two descriptors (“Firm”, “Melly”) are impossible to model from this database: all the extracted models, even if very accurate, are not robust at all.

This ability to prove the impossibility to model a descriptor from a database is a very interesting feature of the use of CV techniques. As stated above, even a regulated learning strategy can generate an overfitted model. In such situations, one would easily be mistaken in using a model that seems to have good performance but has a poor generalization capacity. This particular situation is illustrated by Figure 3 which exposes the performance of a very accurate but definitely non robust model, which means that predictions on unknown situations will be absurd.

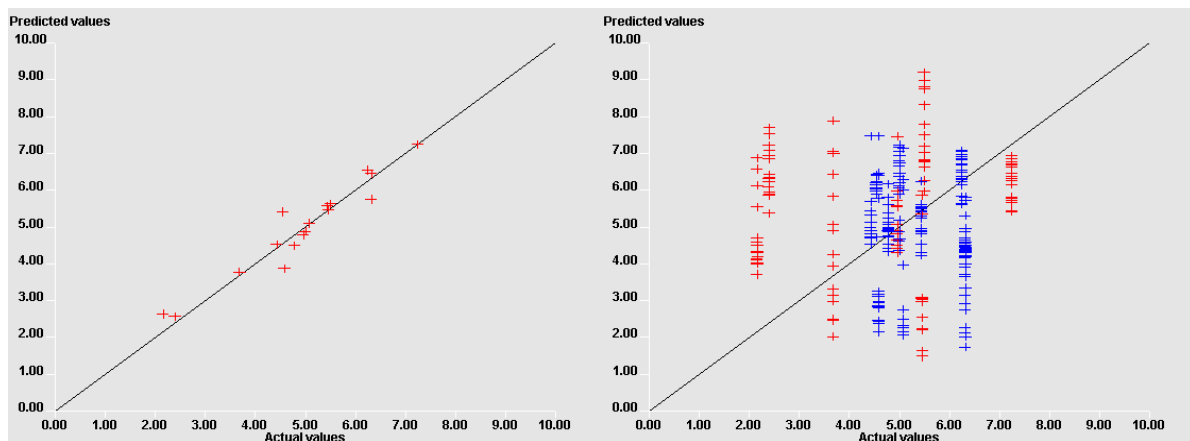


Figure 3: Good accuracy and poor robustness (Monte Carlo 15%) of a model for descriptor “Firm” (5 variables, 3 rules)

<sup>7</sup> A more thorough analysis would certainly allow a strong improvement of the performance of these models, as the low robustness is in most cases due to one or two types of tomatoes only. As such, it is reasonable to think that these tomatoes could be very specific and that a few more products in the database would help modelling these descriptors.

## 4.3 Model exploitation

### 4.3.1 Decision strategy interpretation

The generated models are, as stated before, a collection of linguistic rules that have been induced from the examples gathered in the database. This allows an expert of the domain to compare the knowledge extracted from the data with his own knowledge on the processes to model, to validate or discuss the relevance of these models. Furthermore, this technique is able to linguistically synthesize the behaviour of a completely unknown process.

The variable selection performed during the extraction process allows to identify the sets<sup>8</sup> of variables of major influence on the output of the process. In the case of preference modelling, the optimal subset of variables constitutes the drivers of preference. As the extracted models may be strongly non linear and are usually non monotonous, it is irrelevant to qualify the drivers as being “positive” or “negative” across the whole input space<sup>9</sup>. This qualification may only be performed locally, and is represented by the decision rules.

### 4.3.2 Direct prediction

As the linguistic structure of each model implicitly defines a non linear function, it is possible to provide a model with new inputs to instantly get the prediction of the output for the modelled process. Therefore, by using the 4 models for consumer liking, it is possible to simultaneously predict the preference of each consumer segment for new tomatoes. This materializes a “virtual consumer panel”.

With the same approach, if robust models can be extracted for sensory descriptors, they can be used as a “virtual panel of assessors”.

### 4.3.3 Model inversion

The model inversion feature proposed by **xtractis**<sup>®</sup> OPTIMIZE allows to identify sets of input values for which the output of several models will reach a specified constraint. This constraint is defined as a fuzzy multi-objective request<sup>10</sup>, which in the case of non existence of an optimal solution will nonetheless propose satisfactory solutions. Thus, in the example presented hereafter, we have performed an optimal solution research to maximize the liking of the 4 segments of consumers. It has not been possible to identify a single tomato simultaneously maximizing the preference of the 4 segments, but thanks to the fuzzy definition of the request, the sensory profiles and the instrumental characterisations of several “close to optimal” different tomatoes with high liking scores for all segments have been discovered in less than 30 seconds. The best one is proposed in Figure 4.

---

<sup>8</sup> It frequently happens that some variables are irrelevant if used alone, whereas when used together they give useful information on the process. **xtractis**<sup>®</sup> variable selection preserves those synergies between variables.

<sup>9</sup> For example, when relating preference to salt concentration, this concentration would be a positive driver in areas of low concentration, and a negative driver in areas of high concentration, depicting the fact that consumers usually dislike both non salty and too salty products. Hence there is no sense in trying to qualify positiveness or negativeness of salt concentration on the whole span of possible values.

<sup>10</sup> The fuzzy request is a conjunction of elementary requests defined by the application of fuzzy / crisp operators on the different outputs. It is also possible to add fuzzy / crisp analytical constraints on the inputs. Example: “Which formulation will maximize [Acidity] and reach a value of about 4 for [Sweet], given the fact that [CSugar] must equal 2 x [Clime]?”

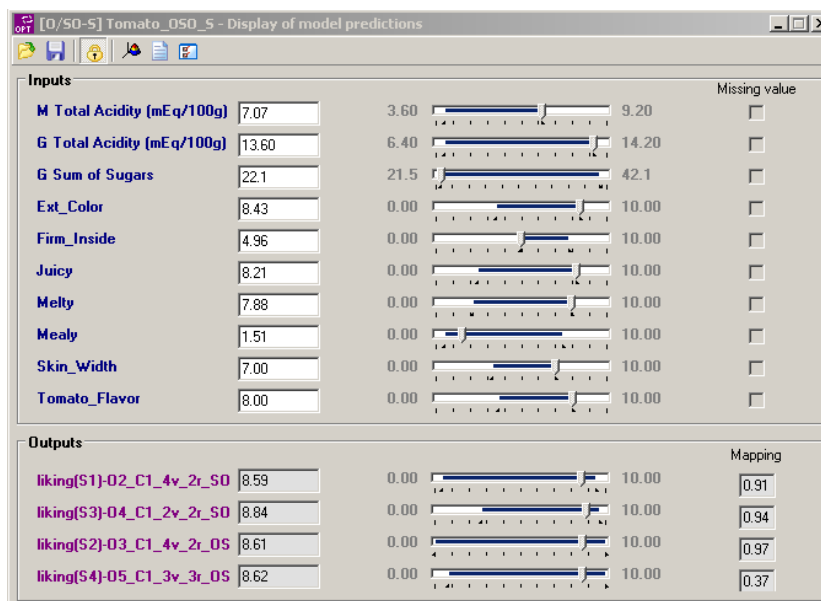


Figure 4: Result of optimization (maximize preference for all segments)

## 5. Conclusion and prospects

We have emphasized that studying the capacity of generalization of the generated models is mandatory, and shown that evaluating the performance of a model by relying on the learning sample only or on a single validation sample is insufficient. This is especially the case in most sensory engineering studies because of the small size of the samples. Therefore we recommend systematically relying on robustness analyses such as cross validation to assess the quality of models generated by any kind of extraction technique.

Using efficient learning algorithms and cross validation analyses for robustness assessment, the **xtractis**<sup>®</sup> approach allows to take benefits from the use of fuzzy models, without the need to master the fuzzy theory, even when no *a priori* knowledge about the process is available.

The duality of fuzzy model merges qualitative and quantitative modelling:

- the linguistic definition of models qualitatively explains the decision making strategy (identification of drivers, influence of drivers on consumer liking, sensory effects of specific components);
- the implicit definition of a multidimensional non linear function gives these models the capacity to make direct predictions (virtual trained panel, virtual consumer group), but also to perform model inversion that give a practical and efficient help for product optimization.

Hence, the use of **xtractis**<sup>®</sup> modelling approach allows to reduce design time by helping to identify optimal products meeting a request defined on outputs and thus reduce the number of prototypes to be made. This also reduces the number of sensory evaluation sessions and the number of variables to measure, as well as consumer tests thanks to the availability of deterministic predictive models. Moreover, it may help favouring the models using variables that are cheap to measure.

This modelling approach being independent of the type of data, it can also be successfully applied to other fields than sensory engineering, such as risk analysis, finance, Customer Relationship Management or natural phenomena.

## Bibliography

- Daudin, J. J., DUBY, C., & Trécourt, P. (1989). PCA stability studied by the bootstrap and the infinitesimal Jackknife method. *Statistics*, 20, 255-270.
- Dubois, D., & Prade, H. (1994). *Possibility theory and data fusion in poorly informed environments*. Institut de Recherche en Informatique de Toulouse (I.R.I.T.) - C.N.R.S. (23 p).  
<http://citeseer.ist.psu.edu/47324.html>
- Grandvalet, Y. (2001). *Sélection de modèles pour l'apprentissage statistique*, Coursebook. University of Technology of Compiègne, France (44 p).  
<http://www.hds.utc.fr/~grandval/apprentissage.pdf>
- Hadamard (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 49-52.
- Jain, A. K., & Mao, J. (1997). Guest editorial: Special issue on artificial neural networks and statistical pattern recognition. *IEEE Transactions on Neural Networks*, 8(1), 1-3.
- Kosko, B. (1992). Fuzzy systems as universal approximators. *Proceedings of the 1<sup>st</sup> IEEE International Conference on Fuzzy Systems - FUZZ-IEEE FUZZ92* (pp. 1153-1162). San Diego, CA, USA.
- Plutowski, M. E. P. (1996). *Survey: Cross-validation in theory and in practice*. David Sarnoff Research Center, Princeton, New Jersey, USA (33 p).
- Vapnik, V. N. (2000). *The nature of statistical learning theory*, 2<sup>nd</sup> Ed. Springer Verlag.
- Zadeh, L. A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. Systems, Man and Cybernetics*, 3, 28-44.
- Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning. *Information Science*, Part I: 8, 199-249, Part II: 8, 301-357, Part III: 9, 43-80.
- Zalila, Z. (1993). *Contribution à une théorie des relations floues d'ordre n*. PhD thesis, D 573, University of Technology of Compiègne, France (474 p).
- Zalila, Z., Cuquemelle, J., Penet, C., Chikh, A., Lorentz, B., Deschamps, D., & Assemat, C. (2007a). *xtractis<sup>®</sup> non linear modelling approach*, White Paper, **intellitech**, Compiègne, France (10 p).
- Zalila, Z., Cuquemelle, J., Penet, C., & Lorenz, B. (2007b). Why Cross-Validation methods are particularly relevant to robust modeling? Interests in sensory science. *Proceedings of the 7<sup>th</sup> Pangborn Sensory Science Symposium*. Hyatt Regency, Minneapolis, MN, USA, August 12-16.
- Zalila, Z., Davodeau, S., Assemat, C., Cuquemelle, J., Chikh, A., Deschamps, D., Marbach, S., Lorentz, B., & Penet, C. (2005). *xtractis<sup>®</sup> fuzzy predictive models for virtual testing and prototyping. Sensory evaluation of cheese spreadability*. *Proceedings of the 6<sup>th</sup> Pangborn Sensory Science Symposium*. Harrogate International Centre, North Yorkshire, UK, August 7-11.