Imperfect knowledge and data-based approach to model a complex agronomic feature: application to vine vigor

C. Coulon Leroy, Brigitte Charnomordic, M. Thiollet Sholtus, Serge Guillaume

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Imperfect knowledge and data-based approach to model a complex agronomic feature – application to vine vigor.

Cécile COULON-LEROY [(1)(2)], Brigitte CHARNOMORDIC [(3)], Marie THIOLLET-SCHOLTUS [(1)], Serge GUILLAUME [(4)]

[(1)] INRA UE1117, UMT Vinitera, 49071 Beaucouzé, France

Tel: (33) 241235555; mail: c.coulon@groupe-esa.com

[(2)] Present address: LUNAM Université, Groupe ESA, UPSP GRAPPE, 55 rue Rabelais, BP30748, 49007 Angers, France

[(3)] INRA Supagro, UMR MISTEA, 34060 Montpellier, France

[(4)] Irstea, UMR ITAP, 34196 Montpellier, France

Abstract

Vine vigor, a key agronomic parameter, depends on environmental factors but also on agricultural practices. The goal of this paper is to model vine vigor according to both kinds of influential variables. The perspective is to design a decision support tool to adapt the agricultural practices to the environment in order to get a given vigor target.

The approach was based upon a collected dataset and the available expert knowledge. It included a data selection step, which was needed because of data imperfection and incompleteness. Usually implicit in the literature, data selection was carried out with explicit criteria. Then a fuzzy model was designed from the selected data. Owing to the fuzzy model interpretability, its structure and behavior were analyzed to identify input-output relationships and interactions between variables.

The case study was located in a French vineyard in the middle Loire valley. The input features were related to soil, rootstock and inter-crop management, the output was an expert assessment of vine plot vigor. Results showed that, despite the data imperfection, the approach was able to select data that yielded an informative model.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Well-known relationships were identified, and some elements of new or controversial knowledge were discussed.

Keywords: Fuzzy logic, knowledge imprecision, hidden variables, automatic learning, data selection, data inconsistency, interaction

Highlights

- Supervised learning is done using imperfect data, knowledge and databases
- A selection procedure based on the k-means algorithm is used to select consistent data
- Fuzzy inference systems built using automatic learning allow to identify relationships and interaction between variables

1. Introduction

In modern agriculture, an important issue is to optimize the agricultural practices according to environmental factors, in order to reach a given yield level and product quality. Models can be used as support for decision making.

In general, agricultural systems are complex systems; this is the case of vine growing. Vegetative vine development, called ‘vigor’, takes into account the rhythm and the intensity of the vine shoot growth (Carbonneau et al., 2007). Empirically, vine vigor level is well known as being stable over the years (Johnson, 2003; Kazmierski et al., 2011). It is highly influenced by environmental factors, such as soil or climate, but can also be modified by agricultural practices (choice of rootstock, inter-row management, pruning type, among others). Vine vigor is a key parameter to control the balance between vegetative growth and productivity that influences berry composition and then wine characteristics (Bramley et al., 2011; Kliewer and Dokoozlian, 2005).

Some complex mathematical models are available for vine development. These models work at a very large scale and for contrasting environmental conditions

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

(Garcia de Cortazar Atauri, 2007; Valdes-Gomez et al., 2009). Some of them were designed for decision support with respect to very specific problems as the salinity in Australia (Walker et al., 2005). Some other models were not validated under various field conditions (Nendel and Kersebaum, 2004). For complex systems, it is difficult to design formal mathematical models. An alternative approach consists in deriving empirical statistical models from experiments.

However, for perennial crops such as vine, full experimental designs to test a large number of factors in interaction are very difficult to implement. On-going research consists, in most cases, in experimentally quantifying the impact of one variable on vine development while the other variables are being fixed *e.g.* (Bavaresco et al., 2008). Even if, at vineyard scale, interactions between variables involved in the agricultural system are empirically observed by winegrowers, these observations are not sufficient to analyze the functioning of the agronomical system. A special case of interesting interactions is the simultaneous impact of some environmental factors and agricultural practices. Some interactions between variables have been highlighted for vine vigor *e.g.* interactions between cover cropping and water supply (Celette et al., 2005), or between cover cropping and rootstock (Barbeau et al., 2006; Hatch et al., 2011). To identify these interactions is an important step toward a decision support system to adapt agricultural practices to the environment. However, vine vigor is difficult to model from experiments, essentially for two reasons. Firstly, the collected data are tainted with uncertainty; the features can suffer from imprecision, especially when they are assessed by human beings. Secondly, the data set is likely to be incomplete, because the agronomical system has some hidden features that are unknown or hard to assess. Due to these hidden features, the data base will probably include conflicting data: similar recorded combinations of input features may have contradictory output assessment.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Therefore it is important to include a data selection step in the modeling approach. In the literature, that step is often implicit and not described. In this paper, a selection method with explicit criteria is proposed.

Once the data are selected, various learning methods can be used to produce a model to study interactions between variables. They include artificial intelligence or statistical techniques. Both can deal with some kinds of data imperfection and both have been used in environmental modeling (Chen et al., 2008).

Common choices include classical linear models (LM) and decision trees (DT), or for more recent developments, Bayesian networks (BN). These statistical models are efficient in a wide range of situations, and often yield a confidence interval, since they are based on probability theory. However, they may be difficult to interpret for a human being. For instance it is problematic to give a meaning to a LM coefficient. DT are easy to interpret, and have proven very useful for discriminant feature selection but this is not the main objective here. BN can incorporate expert knowledge and yield a graphical model easy to read, provided the number of nodes is not too high. They have been used for diagnosis purposes (Sicard et al., 2011). There are also some clear limitations to BN with respect to the proposed application. It may be difficult for experts to express their knowledge in terms of probability distributions. BN also have a limited ability to deal with continuous data, and discretization assumptions can significantly impact the results. Structure learning of a BN is still an open challenge, and the learning methods have a high complexity. Furthermore, as all statistical methods, they require a large amount of data to produce significant results, which is not always possible to get.

Fuzzy logic and inference systems (FIS) are part of artificial intelligence techniques. In FIS, fuzzy logic is used as an interface between the linguistic space, the one of human reasoning, and the space of numerical computation. FIS handle linguistic concepts, e.g. High or Low, implemented using fuzzy sets. Data imprecision is taken into account thanks to a progressive transition between the qualitative labels

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

used for input or output variables. Fuzzy models are able to represent imprecise or approximate relationships that are difficult to describe in precise mathematical models. Historically, FIS were designed from expert knowledge (Mamdani and Assilian, 1975). This approach is limited to small systems and may give poor accurate results. Specific learning algorithms for FIS have then been proposed by Guillaume and Charnomordic (2012a) and by Guillaume and Magdalena (2006). Fuzzy logic based models are interpretable, under a few restrictions (Guillaume and Charnomordic, 2011), this being particularly important for decision support (Alonso and Magdalena, 2011).

Fuzzy modeling was used in a previous work to predict the vine vigor imparted by the environment (Coulon-Leroy et al., 2012). The objective of the present paper is to propose a more ambitious work using fuzzy modeling to study the interactions between environmental factors, agricultural practices and vine vigor. The approach pays a particular attention to data selection, which is a critical step in supervised learning; even it is usually not explicitly dealt with in the literature. It attempts to make the best of domain expertise and of available field data, though they are incomplete, in order to design an interpretable model. The interpretability makes it possible to analyze the system behavior and to evaluate interactions between variables.

2. Material and methods

In this section, we propose to follow five steps:

- to describe the case study with its input and output variables of (Section 2.1).
- to select data used prior to the automatic learning (Section 2.2) by clustering (Section 2.1.1), generating sub-clusters (Section 2.1.2) and selecting consistent sub-clusters (Section 2.1.3).
- to build the fuzzy model (Section 2.3) by partitioning input variables according to data and expertise (Section 2.3.1) and generating ‘if-then’ rules from data (Section 2.3.2).

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

- to optimize the fuzzy model and to evaluate the system performance (Section 2.4).
- to analyze the optimized system and the interaction between variables (Section 2.5).

The overall procedure is summarized in Figure 1. The multidimensional data are denoted by $(x_1,x_2…x_p, y)$ where x_i (i=1, …,p) are the input variables, and y is the output variable. In the following, the output variable is a categorical variable with a given number of ordered levels.

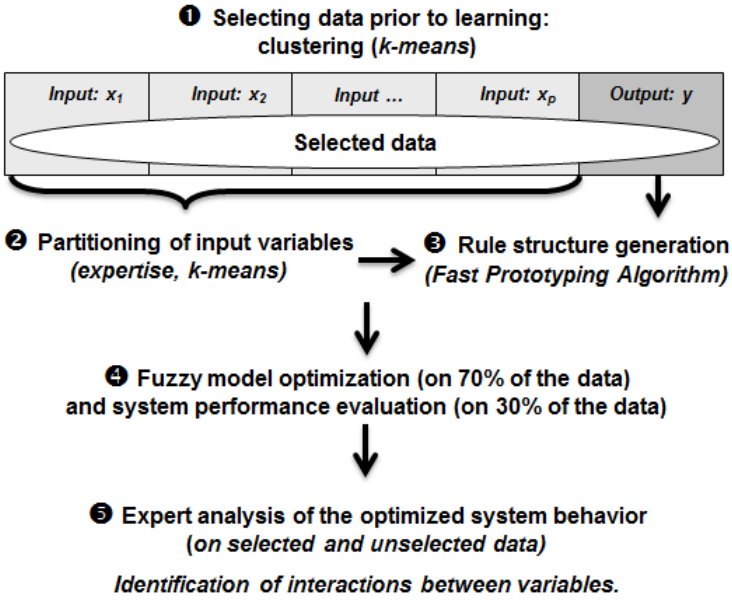

Figure 1: Overall procedure.

All of the developments described in the present works are accessible using the R software (R Development Core Team, 2008) and the FisPro toolbox (Guillaume and Charnomordic, 2012a). R[1] is a free software environment for statistical computing and graphics. FisPro[2] is an open source software that corresponds to ten years of research and software development on the theme of learning interpretable fuzzy inference systems from data. It has been used in the fields of agriculture and

1 http://www.r-project.org
2 http://www7.inra.fr/mia/M/fispro/

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

149 environment (Colin et al., 2011; Coulon-Leroy et al., 2012; Rajaram and Das, 2010;
150 Tremblay et al., 2010).

151 **2.1 Case description**

152 The case study is located in the middle Loire Valley, on the Saumur vineyard,
153 in France. It includes 152 vine plots of a cooperative of winegrowers. Their
154 localization and the soil and sub-soil characteristics are known. The winegrower'
155 practices were surveyed.

156 Some practices are controlled by Protected Designation of Origin (PDO, the
157 French "Appellation d'Origine Contrôlée") Saumur. Thus, some of the practices that
158 influence vine vigor *e.g.* planting density (Carbonneau et al., 2007; Morlat et al., 1984;
159 Murisier, 2007) are not taken into consideration in this study because they are
160 homogeneous over the whole studied area according to the 'Saumur PDO'.

161 The main grape variety is: *Vitis vinifera* cultivar 'Cabernet franc', planted in
162 all studied vine plots.

163 In the studied area, the vine vigor is influenced by soil factors and by two
164 main agricultural practices: rootstock choice and inter-row management. These
165 influential factors are the input of the system.

166 **2.1.1 Input variables**

167 There are three input variables corresponding to the three influential factors:

168 i. Vine vigor imparted by soil (VIG_S). An indicator of the vigor imparted by
169 soil factors to the vine was previously built (Coulon-Leroy et al., 2012). VIG_S is a
170 continuous variable varying between 1 (low imparted vigor) and 3 (high imparted
171 vigor).

172 ii. Vigor conferred by rootstock (VIG_R). Vine is grafted on a rootstock *e.g.*
173 the 3309C to fight against the attack of an insect called *Phylloxera vastatrix*. The
174 rootstock, at the interface between soil and vine variety, interacts with the variety to

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

modify the development of the whole plant (Ollat et al., 2003). For each rootstock, vigor level was determined from the literature (Galet, 1979; Institut Français de la Vigne et du Vin et al., 2007).VIG_R is a discrete variable with five values (1 - very low; 1.5 - low; 2 - medium; 2.5 - high and 3 - very high vigor, as mentioned Table 1).

Table 1: Level of vine vigor imparted to the vine variety by the rootstock (Galet, 1979, Institut Français de la Vigne et du Vin, 2007).

VIG_R value	Vine vigor conferred	Rootstocks
1	Very Low	Riparia, 420A-MG, 44-53M
1.5	Low	101-14, 333EM
2	Medium	3309C, Gravesac, Fercal, 420A, 8BB, 161-49, RSB1
2.5	High	SO4, 5BB, 41B
3	Very High	Rupestris, 1103P, 110R, 99R, 140Ru, 196-17

iii. The inter-row management constraint on vine vigor (VIG_C). A grass cover is introduced in the inter-rows of vineyards to limit runoff and soil erosion however it also limits vine vegetative development of the vine on account of competitions for soil water and nitrogen (Celette et al., 2009). VIG_C is a discrete variable with 10 values (between 0 - no constraint and 3 - high constraint). Values of constraints were obtained by crossing the constraint imparted by the cover crop variety and the cover crop area (Table 2). The constraint imparted by the cover crop variety was determined thanks to technical reports of advisory services. The cover crop area was measured for each vine plot of the studied area. Under a cover of 10%, the surface was considered by the technicians as Low, and over 30% as High.

Table 2: Level of inter-row crop constraint on vine vigor (VIG_C).

		Cover crop area		
		Low	Intermediate	High
Constraint imparted by the cover crop variety	Very low	1	1.5	2
	Low	1.25	1.75	2.25
	Intermediate	1.5	2	2.5
	High	1.75	2.25	2.75
	Very high	2	2.5	3

2.1.2 Output variable

The vigor evaluation (named VIG_OBS) linked to the shoot growth and leaf areas observed on vine plots was used as reference output data to evaluate the

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

196 interactions between environmental factors, agricultural practices and vine growth. A
197 wide range of direct or indirect, destructive or undestructive methods to assess vine
198 vigor exists (Tregoat et al., 2001). Among them, some are based on measurements
199 such as pruning wood weights or leaf areas. However, remote sensing is the most
200 widely used technique to evaluate vine vigor in precision viticulture. Various
201 indicators, *e.g.* the Normalized Difference Vegetation Index (NDVI), are based on leaf
202 reflectance. High-resolution images and specific algorithms are necessary to
203 discriminate pixels result from a mix of vine leaf area, inter-row soil, grass and even
204 shadows (Homayouni et al., 2008; Santesteban et al., 2013). Expert evaluation can
205 also be used (Carey et al., 2007; Morlat and Lebon, 1992). In that case, the assessment
206 is performed in 'three-dimensions'.

207 It appears that expert evaluation is often the only way to make complex
208 assessments, and it is currently used to characterize the sensory properties of an
209 agricultural product. Sensory data are likely to show inconsistency when judges are
210 untrained (Lesschaeve, 2003). This is the case in vine vigor evaluation. However, we
211 chose to use expert evaluation. The main reason was that distinct inter-row crop
212 management strategies made NDVI value not comparable over the study area
213 (Homayouni et al., 2008). The vine vigor was assessed in 2011 by a skilled technician
214 employed by the Saumur wine cooperative (VIG_OBS). Vine vigor is a discrete
215 variable labeled using four ordered levels (1 - very low; 2 – low; 3 - high and 4 - very
216 high). This expert evaluation was used as output training data to build the model.

217 **2.2 Selecting data prior to learning**

218 Classical data cleaning includes feature selection, which was done as
219 described above, by keeping the main influential input variables, according to
220 expertise, literature and field availability. When dealing with complex systems in
221 agronomy, another step may be required. For instance, the soil plant interaction cannot
222 be reduced to a few scalars; other hidden influential variables, that are not usually

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

recorded yet, contribute to explain output variations. That is likely to generate inconsistencies in the data base. So data items need also to be selected, as pointed out by Taskin (2009) about classification image, even if, in many applications, the quality of the learning data is not questioned and the dataset is directly employed in the learning stage. The R software (R Development Core Team, 2008) was used for these developments.

2.2.1 Data clustering

Many clustering techniques could be considered: k-means, fuzzy c-means or hierarchical clustering. We opted for the *k-means* (Hartigan and Wong, 1979; MacQueen, 1967) clustering method for the following reasons. It is a simple and efficient method, with only one parameter: the number of clusters. By contrast, hierarchical clustering requires the choice of the agglomeration method and the dendrogram analysis to determine the suitable number of clusters. Fuzzy c-means, which is the fuzzy generalization of the *k-means* algorithm, was considered, but rejected. In fuzzy c-means, each item is assigned a membership degree to each cluster elements. The membership degree would be responsible for a higher complexity, and difficult to take into account in the next steps. So the *k-means* clustering was carried out on all of the features: input and output variables.

It is well known that the *k-means* algorithm is highly sensitive to the initial clustering centers, which are randomly chosen, so the *k-means* algorithm was run 10 times. Then 10 different partitions of 10 clusters were obtained.

2.2.2 Sub-cluster generation

Because of the random choice of the initial cluster centers, the cluster composition was likely to change from one run to another. The aim of the second step was to select sub-clusters with the same composition over a given number of runs. To ensure group robustness, we focused on items which had been assigned together in a

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

common cluster at least 7 times over the 10 runs. This way, different sets of stable sub-clusters were obtained, denoted by S_{10} (10 times over the 10 runs), S_9 (9 times over the 10 runs), S_8 (8 times over the 10 runs), and S_7 (7 times over the 10 runs).

2.2.3 Consistent sub-cluster selection

For each of the S_i sets, a final selection step was applied in order to use consistent and representative data at the learning step. Only clusters for which the output variance was less than a given threshold, set according to expertise, were chosen. In our case study, as the number of output levels was small (4 levels), the output variance threshold was set to zero. Therefore, each cluster included items with "similar" input values and the same output label. Then, the clusters were ordered according to their output level. To get a learning set that best represented the whole data; the most populated clusters of each output level were selected.

The result is a data set, $D_{i,}$ for each set S_i. Among all these data sets, the one with the highest cardinality was selected for learning the fuzzy model.

2.3 Fuzzy modeling

Fuzzy inference systems were chosen as they provide a modeling framework, able to combine expertise and data. The inference engine is a set of rules whose premises use linguistic terms. Each of these linguistic terms was implemented as a fuzzy set in the numerical space. The fuzzy system design involved two different steps: first the input variable partitioning and then the rule generation. Next the fuzzy inference system was optimized. Variable partitioning only involved the feature data distribution, without considering any further use. This way the same fuzzy partition can be used with several rule induction algorithms. Fuzzy partitions and fuzzy rules define the FIS structure. Model optimization, introduced in Section 2.4, aimed to tune FIS parameters, membership function location and rule conclusion, while preserving both system structure and semantics.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Let us now detail steps 2: partitioning of input variables, and 3: rule structure generation, of the approach summarized in Figure 1.

2.3.1 Partitioning of input variables

A fuzzy set is defined by its membership function (MF). A point in the universe, x, belongs to a fuzzy set with a membership degree, $0 \leq \mu(x) \leq 1$. If H is a fuzzy set representing High vigor levels, the membership degree of a given vigor value, x, $\mu_H(x)$, can be interpreted as the level up to which the x vigor level should be considered as High. Several fuzzy sets, e.g. Low, Medium and High, can be defined in the same universe, as illustrated in Figure 2.

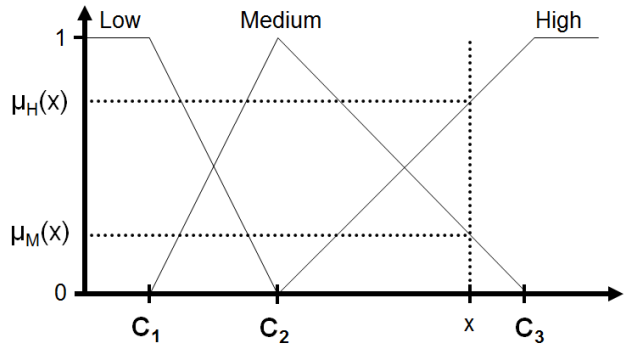

Figure 2: Example of three fuzzy sets defined in the same universe. They define a fuzzy partition of the variable. 'x': a point of the universe, $\mu_M(x)$: the membership degree in the 'Medium' membership function, $\mu_H(x)$: the membership degree in the 'High' membership function.

As fuzzy sets usually overlap, a data point is likely to belong to more than one fuzzy set. In the partition shown in Figure 2, the value x belongs to the fuzzy sets Medium and High with the corresponding membership degrees $\mu_M(x)$ and $\mu_H(x)$. Moreover, for each point in the universe, the sum of the membership degrees to all the fuzzy sets of this kind of partition is equal to one. These so called "strong fuzzy partitions" have good properties regarding semantics. They allow managing the progressiveness of the phenomenon as well as a smooth transition between categories. Working with the membership degrees in the different linguistic concepts, instead of

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

the raw data values, reduces the system sensitivity to raw data variation. This is a convenient and meaningful way to tackle biological variability.

Discrete variables can also be considered under the condition that their values are ordered and have a progressive semantic meaning.

The process of partitioning comes to choose the number of fuzzy sets and the corresponding characteristic points (C_1, C_2 and C_3 in Figure 2).

The number of fuzzy sets was determined by expertise, in order to have a number of concepts corresponding to the usual expert vocabulary. VIG_S and VIG_C were partitioned into three fuzzy sets corresponding to the usual terms 'Low', 'Medium' and 'High', used by domain experts and technicians. The discrete variable, VIG_R, was described by five ordered values (Very Low, Low, Medium, High and Very High), corresponding to the rootstock imparted potential vigor as indicated in Table 1 (the 'Very Low' label is not represented in the dataset).

The characteristic points of continuous inputs were not so easy to determine only by expertise so mathematical algorithms were be used. We run the monodimensional *k-means* algorithm on the input data, independently for each variable, and the cluster centers were chosen as characteristic points. More sophisticated methods, such as hierarchical fuzzy partitioning, available in FisPro, could be used. Once again, we decided in favor of the *k-means*, for its simplicity and efficiency.

Since VIG_S was partitioned into 3 fuzzy sets, VIG_R into 5 and VIG_C into 3, the number of possible rules was 3.5.3=45. However, the rule learning methods did not generate all of them, as described below.

2.3.2 Rule structure generation

Fuzzy sets are used in a Fuzzy Inference System (FIS) to build linguistic rules. A fuzzy rule is written as follows:

$$If\ X^1\ is\ A^1_r\ and\ X^2\ is\ A^2_r \ldots and\ X^p\ is\ A^p_r\ then\ Y\ is\ C^r$$

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

where A^k_r is the fuzzy set of the *k*th input variable used within the *r*th rule, and C^r is the rule conclusion.

The truth degree of the fuzzy proposition X^1 is A^1_r is given, for a sample x_i whose value for the X^1 variable is x^1,by the membership degree of x^1 in A^1_r , $\mu A_r^1(x^1)$. All the partial degrees in the conditional part of the rule are combined using an operator, called a t-norm, which generalizes the logical AND operator:

$$W^r(x_i) = \mu A_r^1(x^1) \wedge \mu A_r^2(x^2) \wedge \ldots \wedge \mu A_r^p(x_i^p)$$

where $\wedge$ is the t-norm. The most common t-norms are the minimum and the product.

$W^r(x_i)$ is called the matching degree of rule r for the *i*th sample.

The rule conclusion can be either fuzzy, Mamdani type FIS (Mamdani and Assilian, 1975), or crisp. When the output is crisp, and the rule conclusion is reduced to scalar, the type of system is referred to as a zero-order Sugeno FIS (Takagi and Sugeno, 1985) which is equivalent to a Mamdani FIS (Glorennec, 1999).

In the following, the system is a zero-order Sugeno FIS and the t-norm is the minimum.

Thanks to the fuzzy set overlap, a given input is likely to fire several rules simultaneously. Consequently, all these rules will be involved in the system inference and the rule conclusions will be aggregated to give the final output. The Sugeno rule aggregation is performed using a weighted sum of the rule conclusions, the weights being the respective rule matching degrees (Equation 1).

$$\hat{y}_i = \frac{\sum_{r=1}^{n} W^r(x_i) C^r}{\sum_{r=1}^{n} W^r(x_i)} \quad (1)$$

Where $\hat{y}_i$ is the final output value, *n* the number of rules, $W^r(x_i)$ the *r*th rule matching degree and C^r the *r*th rule conclusion. That way, the output is continuous.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Many rule generation methods are available in the literature. Four of them, tuned to yield interpretable results, are implemented in FisPro: Fuzzy Decision Trees (FDT), a procedure proposed by Wang and Mendel (1992) (WM), Fuzzy Orthogonal Last Squares (F-OLS) and the Fast Prototyping Algorithm (FPA). Let us give a quick summary of them.

FDT are an extension of classical decision trees, starting from a root node including all data set items, FDT use a recursive procedure to split each node into Mj child nodes, where Mj is the number of fuzzy sets in the j*th* input variable partition selected for the split. For each node, the algorithm selects the variable according to a discriminant criterion, based on entropy or variance. The FDT implementation in FisPro is based on Weber (1992).

In its FisPro implementation, WM is not very different from FPA, The main difference stems from the way the rules are initialized. With FPA, they are calculated using a subset of examples, whereas WM only takes into account a single item.

F-OLS is inspired from linear regression model fitting. The algorithm maps the input variables into a transformed linear space, and ranks the induced rules by decreasing order of explained output variance.

The Fast Prototyping Algorithm (Glorennec, 1999) consists of generating the rules that, among all possible combinations of antecedents, satisfy the two following criteria: (i) the rule matching degree is higher than a given threshold for (ii) at least a given number of data items.

FPA has the advantage of providing a summarized but fair view of the dataset. It is less sensitive to outliers than FDT and F-OLS. WM has a rough management of conflicts, which is not adequate here. For those reasons, we decided to use FPA as a rule generation method. Let us give some more details about it.

Using FPA, in a first step, the rules corresponding to the input combinations are generated, only if there are corresponding data in the data set. In a second step their conclusions are initialized according to the data values as given by the Equation 2.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

$$C^r = \frac{\sum_{i \in E^r} W^r(x_i) \times y_i}{\sum_{i \in E^r} W^r(x_i)} \quad (2)$$

Where $W^r(x_i)$ is the matching degree of the *i*th example for the *r*th rule, and E_r is a subset of examples chosen according to their matching degree to the rule. C^r is the *r*th rule conclusion.

If there are not enough items that fire the r^{th} rule with a degree higher than the user defined threshold, the rule is not kept. Thus, FPA yields a subset of all the possible rules. We set the threshold to a membership degree of 0.2, and the minimum cardinality of E^r to 1. In order to carry a complete analysis, we did not exclude rules that only correspond to a few examples, as the sample has been carefully selected.

2.4 Fuzzy model optimization and system performance evaluation

The FIS accuracy can be improved using an optimization sequence without losing the system interpretability (Casillas et al., 2003; Evsukoff et al., 2009). As partition parameters and rules have been generated separately, it is interesting to run an optimization procedure of the model as a whole. The optimization algorithm used in this work has been proposed in Guillaume and Charnomordic (2012a). It is adapted from Glorennec (1999) and based upon the work of Solis and Wets (1981). It allows optimizing all of the FIS parameters: input or output partitions and rule conclusions.

The input variables were optimized each in turn, the order depending on the variable importance. To assess that importance, the variables were ranked according to a fuzzy decision tree.

The selected data set was split into a learning set (70% of the vine plots) and a test set (30% of the vine plots). Ten pairs of learning and test sets were randomly created, taking into account the output distribution levels. The optimization procedure yielded as many FIS as training test pairs. Then a *median* FIS was computed, resulting of the combination of the ten optimized FIS; the various optimized parameters were

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

replaced by their median value, which is statistically more robust than the mean (Guillaume and Charnomordic, 2012b).

The optimization procedure was guided by the root mean square error (RMSE) index, given in Equation 3 , and the R-squared (R2), given in Equation 4.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \quad (3)$$

where $\hat{y}_i$ is the inferred value for the *i*th item, y_i its observed value and N the number of items.

The R squared (R^2), defined in Equation 4, was used to characterize the system accuracy.

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \quad (4)$$

where $\overline{y}$ is the average of observed values.

The optimization process does not change the system structure; the number of MFs remains the same for all the variables as well as the rule premise structure. Only the MF parameters and the rule conclusions are modified. This allows the semantic properties of the initial model to be preserved while the model accuracy is improved.

The fuzzy characteristics points and the rule conclusions were compared before and after optimization.

2.5 Optimized system analysis

Due to the linguistic reasoning, the system behavior can be analyzed through the study of input-output relationships. Ideally, some well-known relationships should be found as they have been identified in the literature, and some others should appear and raise questions about the empirical or scientific knowledge. Their analysis may yield interesting information about variable interactions (Delgado et al., 2009).

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Finally, the optimized system was tested on the items that were part of the initial data set, but did not belong to the learning data set. A classical validation procedure was not reasonable, due to the presence of conflicting data in the initial data set. Let us note that these conflicts arise from the complexity of the phenomena, and that we only have a partial view of the studied system. The main features were recorded, but some auxiliary ones were not.

Therefore the test procedure had for main objective to check which cases were consistent with the system and to focus on the reasons behind the inconsistent cases. The expected outcome was some complementary knowledge on the agricultural system behavior.

3. Results and discussion

We now present the results of the approach, applied to the case study described in Section 2.1. The various steps detailed in Sections 2.2, 2.3 and 2.4 are illustrated, each in turn.

3.1 Selection of learning data

Our objective was to select consistent data in order to learn coherent input output relationships, using the procedure described in Section 2.2, with a three-step selection scheme based on the *k-means* clustering.

3.1.1 *k-means* clustering

The cluster cardinalities ranged from 4 to 32 and the cluster composition varied from one run to another. This experimentally confirmed the necessity to repeat the *k-means* clustering. As an example, let us analyze the results shown in Table 3, where the 19 reported vine plots have the same values of VIG_C and VIG_OBS. We focus on the 8 plots that were in cluster #1 or #6. The vine plots 320-24, 372-6, 436-40 and 45-19 were together in the same cluster over the ten runs (cluster #1). The same phenomenon occurred for another set of vine plots: 339-27 and 406-8 (cluster #1 or

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

#6). But for run 7, 339-27 and 406-8 were in the cluster #6 with plots 339-22 and 339-23 that were in cluster #2 over the other runs. Therefore all these 8 plots are in S_9 (the selection of sub-clusters with the same composition over 9 runs), but only the 4 plots 320-24; 372-6, 436-40 and 45-19 are in S_{10}.

Table 3: Some clustering results. VIG_S: vine vigor imparted by soil. VIG_R: vine vigor conferred by the rootstock, the level of inter-row management constraint on vine vigor (VIG_C) is here equal to 2.25 and the observed vine vigor (VIG_OBS) equal to 4 for all of the 19 vine plots; *k-means* were run 10 times, the Run i column gives the cluster assignment for each row and the i*th* run.

Vine plot	VIG_S	VIG_R	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
320-24	2.75	2	1	1	1	1	1	1	1	1	1	1
323-14	1.471	2	2	2	2	2	2	2	5	2	2	2
339-16	2.228	2.5	3	3	3	3	3	3	3	3	3	3
339-22	1.699	2	2	2	2	2	2	2	6	2	2	2
339-23	1.753	2	2	2	2	2	2	2	6	2	2	2
339-27	2.147	2	1	1	1	1	1	1	6	1	1	1
372-6	2.809	2	1	1	1	1	1	1	1	1	1	1
403-18	1.734	2.5	4	3	4	4	4	4	3	4	3	4
406-52	1.5	2.5	4	4	4	4	4	4	3	4	4	4
406-8	2.216	2	1	1	1	1	1	1	6	1	1	1
426-25	1.457	2	2	2	2	2	2	2	5	2	2	2
433-7	1.214	2.5	4	4	4	4	4	2	5	4	4	4
436-40	2.787	2	1	1	1	1	1	1	1	1	1	1
45-19	2.543	2	1	1	1	1	1	1	1	1	1	1
476-8	1	2	2	2	2	2	2	2	5	2	2	2
485-9	1.932	3	3	7	3	3	3	3	3	3	3	3
510-16	1.229	2	2	2	2	2	2	2	5	2	2	2
516-42	2	2.5	3	3	3	3	3	3	3	3	3	3
516-52	1.5	2.5	4	4	4	4	4	3	3	4	4	4

3.1.2 Sub-cluster generation

Sub-clusters of S_7, S_8, S_9 and S_{10} were generated according to the results of the 10 *k-means* runs. The characteristics of the sub-clusters that belong to the S_8 (items which have been assigned a common cluster at least 8 times over 10 runs) set are given in Table 4, sorted by increasing values of VIG_OBS. S_8 included 19 sub-clusters, totaling 148 vine plots, out of 152. Three sub-clusters #3, #8 and #14 are composed of vine plots with different VIG_OBS levels, as indicated by a non-null variance. Sub-cluster cardinality ranges from 2 to 16.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Table 4: Characteristics of sub-clusters obtained by the selection of vine plots that were together in the same cluster eight times out of the ten k-means runs. VIG_OBS: the observed vine vigor, VIG_S: vine vigor imparted by soil. VIG_R: vine vigor conferred by the rootstock and VIG_C: inter-row management constraint on vine vigor.

Sub-Clusters	Number of vine plots	Mean VIG_OBS	Variance (n) VIG_OBS	Mean VIG_S	Variance (n) VIG_S	Mean VIG_R	Variance (n) VIG_R	Mean VIG_C	Variance (n) VIG_C
# 1	10	1	0	1.4	0.1	2.1	0	2.1	0.1
# 2	3	1	0	2.8	0	2.5	0	1.92	0.06
# 3	9	1.8	0.2	2.9	0	1.9	0	2.03	0.12
# 4	16	2	0	1.8	0	2	0	1.97	0.09
# 5	8	2	0	1.5	0	2.5	0	1.72	0.1
# 6	6	2	0	2.4	0.1	2.5	0	1.75	0.08
# 7	3	2	0	2.3	0	2	0	1.58	0.06
# 8	7	2.7	0.2	2.5	0.1	2.1	0.1	0	0
# 9	10	3	0	2.5	0.1	2.5	0	1.9	0.07
# 10	7	3	0	1.5	0.1	2.5	0	1.71	0.15
# 11	6	3	0	1.4	0.1	2	0	1.92	0.06
# 12	3	3	0	2.1	0	2	0	1.75	0
# 13	2	3	0	2.8	0	2	0	2.25	0.25
# 14	7	3.1	0.4	1.7	0.1	2.4	0.1	0	0
# 15	14	4	0	2.6	0.1	2	0	1.95	0.18
# 16	13	4	0	1.4	0	2.6	0	1.83	0.1
# 17	12	4	0	2.1	0	2.7	0.1	1.88	0.06
# 18	10	4	0	1.5	0.1	2	0	2.18	0.08
# 19	2	4	0	2.2	0.2	3	0	1.75	0

3.1.3 Consistent sub-cluster selection

Data sets were generated from the S_7, S_8, S_9 and S_{10} clusters.

To illustrate the procedure for the generation of D_8 from S_8, let us examine Table 4. Sub-clusters #3, #8 and #14 were discarded because of their non-null VIG_OBS variance. In order for the selected data to be representative of the initial data set, the two most populated sub-clusters for each VIG_OBS level were kept. There was only one remaining sub-cluster for VIG_OBS level 1, this reflecting the unbalanced VIG_OBS levels.

Some situations *i.e.* combinations with a rootstock that conferred a very low vigor, are not represented in the D_8 dataset. Only four vine plots had this type of rootstock in the whole dataset, and all four of them were assigned to sub-cluster #3 (Table 4), which was discarded from the selection, for the reason given above; they are associated with high values of VIG_S, that corresponding to the choice of the

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

rootstock by the winegrowers to comply with the environmental factors. This complex phenomenon is not integrated in the fuzzy model.

We chose the D_8 dataset for fuzzy model learning, because it had the highest number of vine plots (78), while D_{10} has 55 plots, D_9 has 76 plots and D_7 has 75 plots. The vigor level distributions of the dataset D_8 are quite similar to those of the initial dataset as shown in Figure 3.

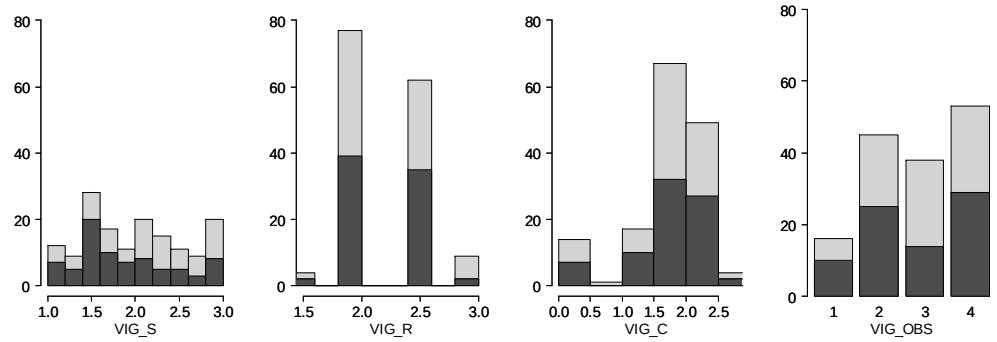

Figure 3: Vigor level (VIG_S: vine vigor imparted by soil, VIG_R: vigor conferred by the rootstock and VIG_C: inter-row management constraint on vine vigor) distribution of the initial dataset (in light grey) and of the selected dataset D_8 (in dark grey).

3.2 Initial system design

The initial system was built considering the D_8 data set (78 vine plots chosen to be consistent and representative of the initial data set). The fuzzy set characteristics points are indicated in Section 3.3 (see in particular Tables 5 and 6). The rule base is shown in Table 7, and we now give some comments on the rules.

First of all, only 19 rules were generated because some combinations were absent from the learning data set. No vine plots were planted with a rootstock that confers either a Very Low or a Very High vine vigor level. Some incoherent combinations from an agricultural point of view were absent, winegrowers choosing the agricultural practices according to the environmental factors. These 19 rules summarize the data using approximate concepts defined by experts.

Rule analysis (Table 7) shows the adaptation or not of the agricultural practices according to the environmental factors. Each rule is matched by different vine plots/examples (see Table 7). Some rules are fired by an important number of

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

examples *i.e.* rules 1, 3 and 4. Other rules are only matched by a single example or a few, *i.e.* rules 16, 17 and 19.

In rules 11, 13 and 19, environmental factors imparting a high vigor are associated to a rootstock that confers a high vigor level. Goulet and Morlat (2010) already noticed that the practices in the vineyard are sometimes unsuitable because they have not been well adapted to environmental factors. For example, the authors indicate that in the vineyard of the Sarthe in Loire Valley (France), 72% of the vine plots have a too vigorous rootstock since the environmental factors induce already a very strong vigor. Combinations existing in a vineyard reflect various levels of practice adaptation according to environmental factors. In the Saumur area, regarding the number of vine plots that activate rules 11, 13 and 19 (Table 7); the adaptation of practices seems to be better.

The performance of the initial system is as follows: RMSE and R^2 are respectively equal to 0.67 and 0.62.

3.3 System optimization

The initial FIS built using the dataset D_8 was optimized, according to the learning and test procedure described in Section 2.4.

After optimization, the fuzzy set parameters C_2 and C_3 of VIG_S were identical (2.1), so that there was no smooth transition between a Medium level of VIG_S and a High level (Table 5). The scale of VIG_S varies between 1 and 3, meaning that the half-scale of VIG_S (values >2.1) is considered as a High vigor level.

Table 5: Fuzzy parameters of VIG_S before and after optimization.

VIG_S fuzzy parameters	Initial FIS	Optimized FIS
C_1	1.4	1.4
C_2	2.0	2.1
C_3	2.8	2.1

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Fuzzy characteristic points of VIG_R correspond to the discrete values of VIG_R: 1.5, 2 and 2.5. VIG_R can take only five values so the optimization is not relevant.

Even if VIG_C is a discrete variable, with 16 possible values, as for VIS_S, it was difficult to determine fuzzy parameters only by expertise. Optimization procedures led to adjust the fuzzy parameter values of VIG_C (Table 6).

Table 6: Fuzzy parameters of VIG_C before and after optimization.

VIG_C fuzzy parameters	Initial FIS	Optimized FIS
C_1	1.00	1.02
C_2	1.65	1.50
C_3	2.25	2.18

Rule conclusions are shown in Table 7. Consequents of rules 8 and 9 strongly decreased after optimization (-1.3 and -1.6 on a [1-4] scale) in contrast with the consequent of rule 2 that did not much change. For the rules corresponding to a Medium VIG_S, the rule conclusions systematically decreased after the optimization.

Table 7: Rule conclusions of the three input variables combinations VIG_S, VIG_R and VIG_C.

	Inputs			Values of rule conclusions		Number of vine plots that activate each rule	
Rules	VIG_S	VIG_R	VIG_C	Initial FIS	Optimized FIS	Initial FIS	Optimized FIS
1	Medium	Low	High	2.6	2.1	26	28
2	Low	Medium	Medium	3.7	4.0	20	14
3	Low	Low	High	1.3	1.2	20	38
4	Low	Low	Medium	1.2	1.3	20	26
5	Medium	Low	Medium	2.5	2.4	26	19
6	Low	Medium	High	3.5	3.8	15	15
7	High	Low	High	4.0	4.0	13	19
8	Medium	Medium	Medium	2.7	1.4	23	10
9	Medium	Medium	High	2.7	1.1	17	10
10	Low	Medium	Low	2.5	2.2	8	7
11	High	Medium	Medium	3.0	2.9	10	7
12	Medium	Medium	Low	3.3	3.2	10	6
13	High	Medium	High	3.0	2.9	9	9
14	High	Low	Medium	4.0	4.0	15	13
15	Low	Low	Low	3.2	3.8	3	10
16	Low	High	Medium	4.0	3.9	2	2
17	High	Low	Low	4.0	3.9	3	3
18	Medium	Low	Low	2.7	2.5	3	8
19	High	Medium	Low	3.5	4.0	3	1

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

The optimization procedure managed to improve the system accuracy. Table 8 summarizes the results of optimization runs, comparing the average results of the initial and the median FIS over the learning and test samples. The median FIS significantly improved the accuracy over the test samples, with a relative gain of 19% for the RMSE and 22% for the R2. It will be used in the following.

Table 8: Performance of the system before and after optimization over the test set.

	FIS	RMSE	R2
Learning set	Initial	0.67	0.64
	Optimized	0.52	0.77
	Relative gain	22%	20%
Test set	Initial	0.67	0.60
	Optimized	0.54	0.73
	Relative gain	19%	22%

3.4 Identification of relationships by expert analysis of the optimized system behavior

The model based on the fuzzy inference system, built using the selected data, has a relatively good accuracy, as discussed in Section 3.3. so its behavior can be interpreted and validated by the agronomists, according to the objectives stated in Section 2.5.

Let us discuss the effect of the VIG_C variable. When vine plots have no intercrop, *i.e.* no constraint on vine vigor, VIG_C=Low (rules 10, 12, 15, 17, 18 and 19), the estimated vigor is always higher than '2', unlike vine plots with an intercrop (Table 7). The impact of a grass cover as intercrop on vine is well known in the literature due to competition for water and nitrogen (Celette et al., 2009). The same authors indicated that intercrop reduces vine growth, *i.e.* the vigor, of the present year but also of the next years by decreasing grapevine nitrogen reserves. These already known relationships, interpreted by expertise, confirm the ability of the method to extract knowledge from a database.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

The study of the impact of rootstock in combination with the other variables required a detailed analysis. Expert analysis of the system behavior disclosed unexpected or new relations.

Let us consider non intercropped vine plots (Figure 4(a)), i.e. without constraint on vine vigor.

When the soil imparts a high vigor level (VIG_S>2.1), the effect of the rootstock is reduced or even erased. The soil effect is predominant. We can visualize these relations in Figure 4.

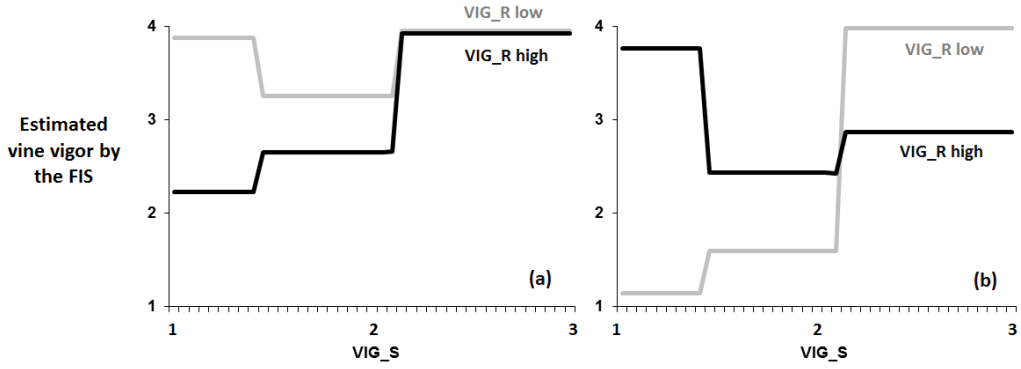

Figure 4: Fuzzy inference system output, estimated vine vigor, according to VIG_S (vine vigor imparted by the soil), VIG_R (vine vigor imparted by the rootstock); (a): Intercropped vine plots – high constraint of the inter-row management on vine vigor, (b): Non-intercropped – low constraint.

The new element brought out by our procedure is to study the combinations of features, while the expertise is often related to the effect of one feature, independently from the other ones.

When the soil imparts a Low or Medium vigor level, the rootstock impact is not as expected. Vine plots with a rootstock that imparts a High vigor (rules 10, 12 and 19) have a lower predicted vigor than vine plots with a rootstock that imparts a low vigor (rules 15, 17 and 18, Table 7). This is at first sight puzzling. After investigation together with technicians of the wine cooperative, the following plausible reason for that contradiction came out. Winegrowers, knowing the potential vigor of their vine plots, fertilized their plots to compensate for that low potential. In the case of non-intercropped plots, a great quantity of fertilizer became available for the top-soil roots

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

of the vine, and that may have increased the vegetative development. This reveals the potential impact of a variable - the level of soil fertility - not yet taken into account. Presently this variable is not systematically measured by winegrowers, except at the time of planting, so it is only available for a few number of vine plots.

Let us now consider plots intercropped with a crop that involves a High constraint (Figure 4(b)). When the soil imparts a Medium or a Low vigor, the estimated vigor is coherent with the empirical knowledge: a Low vigor rootstock leads to a lower vigor; the more the soil imparts a Low vigor, the greater the difference between rootstocks. As can be seen in Figure 4(b), when the soil imparts a High vigor level, and for Low vigor rootstock, the system estimates a higher vigor level than expected. Let us discuss that effect.

Several rootstock varieties impart the same vigor level, nevertheless, in the studied area, most of the time; a low vigor rootstock corresponds to the 101-14 kind and a high vigor rootstock to the SO4 kind. Recent works have shown that some rootstocks are more efficient to extract the soil water content, independently of the conferred vigor (Marguerit et al., 2011). The adaptation of the 101-14 rootstock to the humidity is better than the adaptation of the SO4 rootstock. That way, in the case of soils imparting a high vine vigor level due to high water content, the 101-14 rootstock could be better adapted and so could lead to higher vine vigor. The rootstock ability to adapt to soil humidity should also be taken into account in the model. However it has to be considered in relation with the type of soil and with the climate.

Modeling with linguistic rules allowed the experts to analyze the agricultural system behavior. Induced rules can be considered as pieces of extracted knowledge, and well-known relationships were identified that support the validity of the approach, while unexpected ones were found that led to interesting hypotheses.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

3.5 Running the optimized system on unselected data

We run the optimized system on the 74 (152 – 78) vine plots that were removed from the learning data set. This is not a test procedure in the classical way, when the available data set is split into two parts: learning and test. It would have been interesting to run such a classical validation procedure, had more data been available. In our case, where data have a lot of inconsistencies, that would have drastically reduced the representativeness of the model.

So the unselected data used in this section are not new data. They were not removed randomly from the initial data set, but after a careful and explicit analysis, the objective being to learn the model on consistent data. The model generalization ability cannot be assessed in this way. Nevertheless, some useful information can be found from these experiments on the unselected data.

First of all, 11 vine plots out of 74 agree with the system, the inferred value being equal to the observed value.

Then let us analyze some inconsistencies, by focusing on the plots with the highest differences between inferred and observed vine vigor, whose characteristics are given in Table 9.

For instance, two vine plots have an inferred vine vigor value equal to 4 and an observed value equal to 1. The inferred value is explained by the high VIG_S values (3 and 2.8). We can formulate the hypothesis that hidden variables not taken into account have an impact on the observed vine vigor. The same remark can be done for inferred values equal to 1 instead of 4, mainly due to low vigor imparted by soil factors. Our hypothesis is that soil fertility may explain such results. Winegrowers can compensate for a low vigor imparted by soil factors, by fertilizing their vine plots. The algorithms of Coulon-Leroy (2012) do not take into account the soil fertility. According to the variables taken into account, a high vigor level imparted by soil factors should be predicted but a mineral deficiency can explain such apparent inconsistencies.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Table 9: Values of input variables of 9 vine plots with the highest differences between inferred and observed vine vigor. VIG_S: vine vigor imparted by soil. VIG_R: vine vigor conferred by the rootstock and VIG_C: the level of inter-row crop constraint on vine vigor and VIG_OBS: the observed vine vigor.

Vine plots	VIG_S	VIG_R	VIG_C	VIG_OBS	Vine vigor inferred by the FIS
284-14	3	2	1.75	1	4
372-24	2.8	2	2.25	1	4
161-20	1.5	2	2	4	1
323-14	1.5	2	2.25	4	1
426-25	1.5	2	2.25	4	1
433-6	1.2	2	1.75	4	1
444-24	1.5	2	1.75	4	1
476-8	1	2	2.25	4	1
510-16	1.2	2	2.25	4	1

Finally, the interpretation of some other prediction errors may be partly due to uncertainties in expert evaluation, in particular vigor assessment.

4. Conclusion

The modeling approach developed in this work proposed a methodology to analyze a complex agricultural system, by using available data and knowledge. The key points in the approach are the use of a selection procedure, to select consistent and representative data from the data set, and the choice of a Fuzzy Inference System-based model, built using automatic learning and expertise.

In the fields of Agriculture and Environment, it is very difficult, not to mention naïve and perhaps delusional, to build a full experimental design to study a complex system, such as vine, because of the many features to test out. Therefore the observed data are incomplete, and cannot be used as such for learning a model, while of course the characteristics of the learning dataset have a deep influence on the model design. Data inconsistency would be likely to result in incoherence in the model, so we proposed a method to select consistent agricultural data, with the aim to study the interactions between variables. Both input and output variables were considered in the selection process.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

An interesting asset of the model built using a fuzzy inference system is its interpretability, due to the use of linguistic terms. These terms are implemented by fuzzy sets that avoid the systematic use of crisp thresholds and allow for data uncertainty management. Results could be interpreted, and their analysis showed deep interactions between variables, which comforted the hypothesis that a simplistic expert system based on direct relationships cannot be sufficient.

We considered the main influential input variables for the studied area; other variables, such as soil fertility, could be added in future work because soil fertility impact can explain some of the results obtained by the fuzzy inference system that was built. The future directions could also integrate the impact of fertilization practices.

This work raised some questions about new methodological developments to deal with the uncertainty of input and output measurements or assessments. Undergoing work includes the definition of a new index taking into account a fuzzy target *i.e.* a fuzzy value of the expert evaluation of the vine vigor.

From the agronomical perspective, the interest of this kind of work is to lay down the foundations of a decision support tool aiming to adapt the agricultural practices to the environment in order to get a given vigor target. The methodology used in this paper is generic, and has been applied to a French vineyard, in the Saumur area. A next step consists in testing the method in other vineyards, including rule analysis and system behavior assessment.

Acknowledgements

This work is part of a PhD project that aims to study the combined effects of environmental factors and agricultural practices and link them with wine quality. The first author received a fellowship from the 'INRA-SAD' department and Pays de la Loire Region (France). We are grateful to 'La Cave des Vignerons de Saumur' to make data available to us for this study.

References

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

694
695 Alonso, J.M., Magdalena, L., 2011. Special issue on interpretable fuzzy systems. Inf. Sci. 181,
696 4331-4339.
697
698 Barbeau, G., Goulet, E., Ramillon, D., Rioux, D., Blin, A., Marsault, J., Panneau, J.P.P., 2006.
699 Effets de l'interaction Porte-Greffe / Enherbement sur la réponse agronomique de la vigne *Vitis*
700 *vinifera* L cvs. Cabernet Franc et Chenin. Prog. Agric. Vitic. 123, 80-86.
701
702 Bavaresco, L., Gatti, M., Pezzutto, S., Fregoni, M., Mattivi, F., 2008. Effect of leaf removal on
703 grape yield, berry composition, and stilbene concentration. Am. J. Enol. Vitic. 59, 292-298.
704
705 Bramley, R.G.V., Ouzman, J., Boss, P.K., 2011. Variation in vine vigour, grape yield and
706 vineyard soils and topography as indicators of variation in the chemical composition of grapes,
707 wine and wine sensory attributes. Aust. J. Grape Wine R. 17, 217-229.
708
709 Carbonneau, A., Deloire, A., Jaillard, B., 2007. La vigne. Physiologie, terroir, culture. Dunod,
710 Paris (France).
711
712 Carey, V., Archer, E., Barbeau, G., Saayman, D., 2007. The use of local knowledge relating to
713 vineyard performance to identify viticultural terroirs in stellenbosch and surrounds, In: Nuzzo,
714 V., Giorio, P., Giulivo, C. (Eds.), Proceedings of the International Workshop on Advances in
715 Grapevine and Wine Research. International Society Horticultural Science, Leuven 1, pp. 385-
716 391.
717
718 Casillas, J., Cordón, O., Herrera, F., Magdalena, L., 2003. Interpretability Improvements to
719 Find the Balance Interpretability-Accuracy in Fuzzy Modeling: An Overview, In: Springer
720 (Ed.), Interpretability Issues in Fuzzy Modeling., Heidelberg (Germany), pp. 3-22.
721
722 Celette, F., Findeling, A., Gary, C., 2009. Competition for nitrogen in an unfertilized
723 intercropping system: the case of an association of grapevine and grass cover in a
724 Mediterranean climate. Eur. J Agron. 30, 41-51.
725
726 Celette, F., Wery, J., Chantelot, E., Celette, J., Gary, C., 2005. Belowground Interactions in a
727 Vine (Vitis vinifera L.) - Fescue (Festuca arundinacea Shreb.) Intercropping System: Water
728 Relations and Growth. Plant Soil 276, 205-217.
729
730 Chen, S.H., Jakeman, A.J., Norton, J.P., 2008. Artificial Intelligence techniques: An
731 introduction to their use for modelling environmental systems. Math. Compu. Simulat. 78, 379-
732 400.
733
734 Colin, F., Guillaume, S., Tisseyre, B., 2011. Small Catchment Agricultural Management Using
735 Decision Variables Defined at Catchment Scale and a Fuzzy Rule-Based System: A
736 Mediterranean Vineyard Case Study. Water Resour. Manage. 25, 2649-2668.
737
738 Coulon-Leroy, C., Charnomordic, B., Rioux, D., Thiollet-Scholtus, M., Guillaume, S., 2012.
739 Prediction of vine vigor and precocity using data and knowledge-based fuzzy inference
740 systems. J. Int. Sci. Vigne Vin. 46, 185-205.
741
742 Delgado, G., Aranda, V., Calero, J., Sanchez-Maranon, M., Serrano, J.M., Sanchez, D., Vila,
743 M.A., 2009. Using fuzzy data mining to evaluate survey data from olive grove cultivation.
744 Comput. Electron. Agr. 65, 99-113.
745
746 Evsukoff, A.G., Galichet, S., de Lima, B.S.L.P., Ebecken, N.F.F., 2009. Design of interpretable
747 fuzzy rule-based classifiers using spectral analysis with structure and parameters optimization.
748 Fuzzy Set Syst. 160, 857-881.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Galet, P., 1979. A practical ampelography: grapevine identification. Cornell University Press, New-York (USA) and London (England).

Garcia de Cortazar Atauri, I., 2007. Adaptation du modèle STICS à la vigne (*Vitis vinifera L.)* utilisation dans le cadre d'une étude d'impact du changement climatique à l'échelle de la France. PhD of the "Ecole Nationale Supérieure Agronomique de Montpellier", p. 292.

Glorennec, P.-Y., 1999. Algorithmes d'apprentissage pour systèmes d'inférence floue. Hermès Sciences Publicat., Paris (France).

Goulet, E., Morlat, R., 2010. The use of surveys among wine growers in vineyards of the middle-Loire Valley (France), in relation to terroir studies. Land Use Policy 28, 770-782.

Guillaume, S., Charnomordic, B., 2011. Learning interpretable fuzzy inference systems with FisPro. Inf. Sci. 181, 4409-4427.

Guillaume, S., Charnomordic, B., 2012a. Fuzzy inference systems: An integrated modeling environment for collaboration between expert knowledge and data using FisPro. 39, 8744-8755.

Guillaume, S., Charnomordic, B., 2012b. Parameter optimization of a fuzzy inference system using the FisPro open source software, IEEE International Conference on Fuzzy Systems, Brisbane (Australia), pp. 1-8.

Guillaume, S., Magdalena, L., 2006. Expert guided integration of induced knowledge into a fuzzy knowledge base. Soft Comput. 10, 773-784.

Hartigan, J.A., Wong, M.A., 1979. A k-means clustering algorithm. Appl. Stat. 28, 100-108.

Hatch, T.A., Hickey, C.C., Wolf, T.K., 2011. Cover Crop, Rootstock, and Root Restriction Regulate Vegetative Growth of Cabernet Sauvignon in a Humid Environment. Am. J. Enol. Vitic. 62, 298-311.

Homayouni, S., Germain, C., Lavialle, O., Grenier, G., Goutouly, J.P., Van Leeuwen, C., Da Costa, J.P., 2008. Abundance weighting for improved vegetation mapping in row crops: application to vineyard vigour monitoring. Can. J. Remote Sens. 34, 228-239.

Institut Français de la Vigne et du Vin, INRA, Montpellier SupaAgro, Viniflhor, 2007. Catalogue of grapevine's varieties and clones cultivated in France.

Johnson, L.F., 2003. Temporal stability of an NDVI-LAI relationship in a Napa Valley vineyard. Aust. J. Grape Wine Res. 9, 96-101.

Kazmierski, M., Glemas, P., Rousseau, J., Tisseyre, B., 2011. Temporal stability of within-field patterns of NDVI in non irrigated mediterranean vineyards. J. Int. Sci. Vigne Vin 45, 61-73.

Kliewer, W.M., Dokoozlian, N.K., 2005. Leaf area/crop weight ratios of grapevines: influence on fruit composition and wine quality. Am. J. Enol. Vitic. 56, 170-181.

Lesschaeve, I., 2003. Evaluating wine ‘‘typicité” using descriptive analysis, 5th Pangborn sensory science symposium, Boston (USA).

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, Berkeley Symposium on Mathematical Statistics and Probability, Berkeley (USA), pp. 281-297.

Mamdani, E.H., Assilian, S., 1975. An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man-Mach. Stud. 7, 1-13.

Marguerit, E., Brendel, O., Van Leeuwen, C., Delrot, S., Ollat, N., 2011. Grapevine rootstock genetically determine scion transpiration and its response to water deficit: an integrated approach using ecophysiology and quantitative genetics, 'Systems approaches to crop improvement' conference, Harpenden (United Kingdom).

Morlat, R., Lebon, E., 1992. Experience of multisite trials for the study of vineyards. Prog. Agric. Vitic. 109, 55-58.

Morlat, R., Remoue, M., Pinet, P., 1984. The influence of the planting density and the method of soil management on root growth in a vineyard planted on good soil. 4, 485-491.

Murisier, F., 2007. The influence of plant density and hedgerow height on grape and wine quality. Trial on Gamay vines in Leytron (Wallis, CH). Revue suisse Vitic. Arboric. Hortic. 39, 251-255.

Nendel, C., Kersebaum, K.C., 2004. A simple model approach to simulate nitrogen dynamics in vineyard soils. Ecol. Model. 177, 1-15.

Ollat, N., Tandonnet, J.P., Bordenave, L., Decroocq, S., Geny, L., Gaudillere, J.P., Fouquet, R., Barrieu, F., Hamdi, S., 2003. Vigour conferred by rootstock: hypotheses and direction for research. Bull. OIV 76, 581-595.

R Development Core Team, 2008. R: A language and environment for statistical computing, Vienna (Austria).

Rajaram, T., Das, A., 2010. Modeling of interactions among sustainability components of an agro-ecosystem using local knowledge through cognitive mapping and fuzzy inference system. Expert Sys. Appl. 37, 1734-1744.

Santesteban, L.G., Guillaume, S., Royo, J.B., Tisseyre, B., 2013. Are precision agriculture tools and methods relevant at the whole-vineyard scale? Precis. Agric. 14, 2-17.

Sicard, M., Baudrit, C., Leclerc-Perlat, M.N., Wuillemin, P.H., Perrot, N., 2011. Expert knowledge integration to model complex food processes. Application on the camembert cheese ripening process. Expert Syst. Appl. 38, 11804-11812.

Solis, F.J., Wets, R.J.B., 1981. Minimization by random search techniques. Math. Oper. Res. 6, 19-30.

Takagi, T., Sugeno, M., 1985. Fuzzy Identification of Systems and Its Applications to Modeling and Control. 15, 116-132.

Taskin, K., 2009. Increasing the accuracy of neural network classification using refined training data. Environ. Model. Softw. 24, 850-858.

Tregoat, O., Ollat, N., Grenier, G., Leeuwen, C.V., 2001. Comparative study of the accuracy and speed of various methods for estimating vine leaf area. J. Int. Sci. Vigne Vin 35, 31-39.

Author-produced version of the article published in Computers and Electronics in Agriculture, 2013, 99, 135-145.
The original publication is available at http://www.sciencedirect.com/science/article/pii/S0168169913002275
DOI : 10.1016/j.compag.2013.09.010

Tremblay, N., Bouroubi, M.Y., Panneton, B., Guillaume, S., Vigneault, P., Belec, C., 2010. Development and validation of fuzzy logic inference to determine optimum rates of N for corn on the basis of field and crop features. Precis. Agric. 11, 621-635.

Valdes-Gomez, H., Celette, F., Garcia de Cortazar Atauri, I., Jara-Rojas, F., Ortega-Farias, S., Gary, C., 2009. Modelling soil water content and grapevine growth and development with the STICS crop-soil model under two different water management strategies. J. Int. Sci. Vigne Vin 43, 13-28.

Walker, R.R., Zhang, X., Godwin, D.C., White, R., Clingeleffer, P.R., 2005. Vinelogic growth and development simulation model - rootstock and salinity effects on vine performance, XIV International GESCO Viticulture Congress Geisenheim (Germany), pp. 443-448.

Wang, L.-X., Mendel, J.M., 1992. Generating fuzzy rules by learning from examples., IEEE Transactions on Systems, Man and Cybernetics, pp. 1414-1427.

Weber, R., 1992. Fuzzy-ID3: a class of methods for automatic knowledge acquisition, 2nd Internat. Conf. on Fuzzy Logic and Neural Networks, Iizuka (Japan), pp. 265-268.