PRECISION VITICULTURE DATA ANALYSIS USING FUZZY INFERENCE SYSTEMS

Mathieu GRELIER[1], Serge GUILLAUME[2],
Bruno TISSEYRE[3] and Thibaut SCHOLASCH[4]

1: Comité Interprofessionnel du Vin de Champagne, 5 rue Henri Martin, 51200 Epernay, France
2: Cemagref, Umr Itap, B.P. 5095, 34196 Montpellier cedex 5, France
3: ENSA.M, 2 place Viala, 34060 Montpellier cedex, France
4: University of California, Berkeley, ESPM Department 151 Hilgard Hall, Berkeley, CA 94720-3110, United States

Abstract

Aims: Various types of data are likely to be used in a precision viticulture framework, to adjust management actions according to within field variations. This paper proposes an alternative way of analysis to classical methods.

Methods and Results: Data are analysed using fuzzy logic techniques. The result is a set of linguistic fuzzy rules induced from data. In this paper, the rules are build in order to explain the relationship between vintage quality, reduced to sugar content, and other available variables. The resulting system is proved to be accurate, moreover thanks to fuzzy logic interpretability, the induced rules are analyzed and compared to expert knowledge.

Conclusion: This example highlights the potential of fuzzy logic to deal with precision viticulture datasets.

Significance and impact of study: This is a preliminary work, it has been carried out using a free software available in the internet.

Keywords: Precision viticulture, machine-learning, fuzzy logic, non-linearity, interpretability, linguistic models, spatialized data, Fispro

Résumé

But : Des jeux de données à l'échelle intraparcellaire sont aujourd'hui disponibles et doivent permettre d'orienter le management intraparcellaire en viticulture de précision. Cet article propose une voie d'analyse de ces jeux de données alternative aux méthodes classiques.

Méthodes et résultats : Les données sont analysées à l'aide de techniques basées sur la logique floue. Le résultat est une base de règles linguistiques. Dans cet article, les règles visent à expliquer les relations entre la qualité de la vendange, réduite ici au taux de sucre, et les autres variables disponibles. Le système construit est précis et, grâce à l'interprétabilité de la logique floue, les règles sont analysées et comparées à la connaissance experte.

Conclusion :Cet exemple souligne le potentiel de la logique floue pour traiter les données de viticulture de précision.

Signification et impact de l'étude : Il s'agit d'une étude préliminaire dont l'avantage est d'avoir été réalisée à partir d'un logiciel libre disponible sur internet.

Mots clés : viticulture de précision, apprentissage, logique floue, non-linéarités, interprétabilité, modèles linguistiques, données spatialisées, Fispro

manuscript received : 12 May 2006 - revised manuscript received: 27 November 2007

*Corresponding author: serge.guillaume@montpellier.cemagref.fr

J. Int. Sci. Vigne Vin, 2007, **41**, n°1, 19-31

INTRODUCTION

Over the last years, many new technologies have been developed or adopted in viticulture: positioning systems, such as the Global Positioning Systems, yield and quality monitoring systems embedded on grape harvesters, canopy monitoring systems (airborne imagery or on machine sensors), soil monitoring systems such as soil resistivity, low-cost and reliable devices to store and share the information such as Personal Digital Assistant (PDA)…

When combined, these new technologies produce a large amount of affordable high resolution information, managed by Geographical Information Systems (GIS), and have lead to the development of finer-scale or site-specific management that is often termed Precision Viticulture (PV). Many research and developments projects already exist in most of the significant wine production areas in the world (Tisseyre *et al.*, 2006; Bramley and Hamilton, 2004). These projects show a significant within vineyard variability in yield, vigour and quality in most of the studied areas. They also highlight the opportunity to manage within field variability in order to optimize the production protocol. Thus, many site specific management practices can be considered as control parameters such as differential harvest, variable fertilization rates or water applications.

To improve production efficiency, these data must be related to vine physiology. Unless they are clearly understood, these new pieces of information will not help the grower to make a better decision. Common practice is to use multi-dimensional data analysis techniques such as Principal Component Analysis. But result interpretation requires expertise in the technique itself, and, unfortunately most viticulturists lack the knowledge to properly analyze their own data (Taylor *et al.*, 2005).

Another approach to data analysis is to use automatic learning procedures. This starts with laying down an objective, which can be the prediction of one specific variable of agronomic interest. For that purpose, we give a different status to the variables of the database: we differentiate one variable which becomes the output from those used to achieve the prediction, the inputs. Taken together, inputs and output define a system.

Based on the input and output values, the learning phase consists in finding the model parameters that better reproduce the input/output relationships. Several techniques are available to design such systems, artificial neural networks being the most popular one (Shatar and Mcbratney, 1999; Drummond *et al.*, 2000). These types of models are called black box behaviour models as it is difficult to the user to understand their inner structure and neural network weights do not have any valuable meaning. They cannot be used, for instance, to assess the influence of a given input variable within the model.

In this work we use Fuzzy Inference Systems to address the problem of precision viticulture, as fuzzy logic is well-known for its natural language modeling ability. The inference system is made up of a set of rules of the form If *Situation then Conclusion*, where both parts of the rule are described by linguistic concepts, such as "a high temperature". These rules can be either written by a domain expert or induced from data. Induction is a generalization process which aims at finding what general statements can be made from a set of particular observations. In this case, the statements are expressed as interpretable linguistic rules. This kind of systems can be seen as a fuzzy extension of classical expert systems. The main difference from expert system is that fuzzy predicates allow gradual transitions and robustness. Classical rule induction has already been applied to ecological problems (Dzeroski *et al.*, 1997).

This work investigates the within field variation of one vintage quality characteristic, the sugar content at harvest. It aims to verify the feasibility of (i) assessing the spatial variability of this variable, (ii) extracting the main parameters related to spatial variations, (iii) providing the user with linguistic rules which describe in a simple way how sugar is related with these parameters, and (iv)

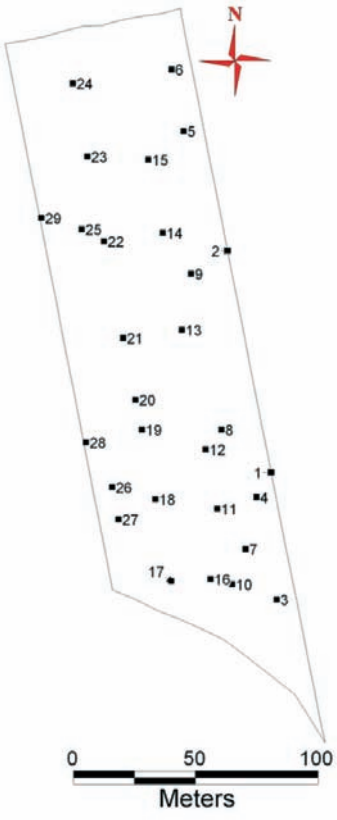


Figure 1. Location in the field of the 29 sample sites.

Tableau 1. The variables used in this study.

Name	Abbreviation	Unit	min	max	mean	var
yield	yld	Tons/ha	1.891	5.882	4.01	0.9571
1/09/03 berry weight	91bw	grams	80.4	203.8	154.79	1142.1
8/09/03 berry weight	98bw	grams	98.4	217.8	167.76	1143.27
6/10/03 berry weight	106bw	grams	103.8	219.7	164.67	852.15
1/09/03 pH	91pH		3.39	3.88	3.565	0.01357
8/09/03 pH	98pH		3.39	3.91	3.647	0.01733
6/10/03 pH	106pH		3.41	3.86	3.666	0.01658
1/09/03 titrable acidity	91ttac	grams of	3.6	5.4	4.428	0.3594
8/09/03 titrable acidity	98ttac	grams of	4.2	5.85	4.866	0.2006
6/10/03 titrable acidity	106ttac	grams of	3	5.1	3.838	0.2805
1/09/03 sugar content	91deg	eq. alcohol	11.67	13.17	12.39	0.162
8/09/03 sugar content	98deg	eq.alcohol	9.92	13.24	12.52	0.424
6/10/03 sugar content	106deg	eq. alcohol	11.96	14.9	13.88	0.4138
1/09/03 malic acid	91malac	g/L	0.4	2.5	0.731	0.1644
8/09/03 malic acid	98malac	g/L	0.3	0.8	0.5207	0.01813
6/10/03 malic acid	106malac	g/L	0.2	0.7	0.3483	0.0183
exposed leaf area	ela	m^2/m of row	1.71	2.64	2.265	0.05935
weight of wood	wood	Kg	1.05	4.15	2.74	0.6777
trunk circumference	circ	cm	152.95	177.1	158.03	121.43
trunk diameter	diam	cm	30.3	52.19	45.56	19.31
0.5 m soil resistivity	res05	ohm*m	26.4	152.6	66.51	1355.89
1 m soil resistivity	res1	ohm*m	26.4	152.6	66.51	1355.89

verifying the spatial relevancy of the rules and their possible use within a decision support system.

MATERIALS AND METHODS

This section includes the data presentation as well as an introduction to fuzzy rule induction.

1. The dataset

The data used in this study have been previously collected, in 2003, in a vine field located in Navarra (Spain). The field is planted of Merlot variety has an area of 2 ha and coordinates are 42°38 N, 1°59 W (WGS84).Vines are planted in non-irrigated conditions. Data were collected for other investigations in the framework of an European Eureka project (VITIS project). Both the sampling locations and the measured variables were out of our control during this experiment.

The data set consists of 12 different variables, 5 of them available at 3 different dates, assessed on 29 sample locations as shown in figure 1.

Details of the 12 variables are described in table 1.

The yield data come from a monitoring system mounted on a grape harvester: 11,807 measurements are available, covering the whole area of the field. To assign a yield value to each of the 29 sample sites, the mean of all the values falling into a six meter radius circle around the site has been computed. This radius value ensures at least 30 measurements are taken into account for each sample site.

Some variables, related to grape quality, are measured at three different dates: 1/09/03, 8/09/03 and 6/10/03 (harvest). These variables are berry weight, pH, titrable acidity, sugar content (given in equivalent of alcohol) and malic acid.

Other variables are related to the vine plant: trunk circumference, trunk diameter, weight of wood, exposed leaf area. Trunk circumference and diameter were measured in the beginning of 2003. The spatial variability of these parameters is assumed to be time stable in non irrigated conditions since vigour (and all the related parameters) are strongly related with time stable parameters such as soil depth, water availability and nutrient availability. The weight of wood is measured during pruning and is usually interpreted as a vigour indicator. Exposed leaf area was measured at the end of growth, that generally occurs in July under these latitudes, three months before harvest.

Finally soil resistivity is measured at two different depths, 0.5 and 1 m. Although the absolute values are likely to change according to climate variations, the exhibited soil electrical spatial variability remains time stable since it is related to soil characteristics.

The aim of our study is to model the relationships between a vintage quality variable and other relevant parameters. Since sugar content at harvest is one of the

most important parameter for the growers, we chose the 6/10/03 vintage sugar content. Thus, sugar content at harvest is the output parameter of our system.

2. Fuzzy rule induction method

Fuzzy Inference Systems are one of the most famous applications of fuzzy logic and fuzzy sets theory (Zadeh, 1965).

The strength of Fuzzy Inference Systems relies on their twofold identity: on one hand they are able to handle linguistic concepts; on the other hand they are universal approximators able to perform non linear mappings between inputs and outputs, through automatic learning procedures.

But, applying that type of procedures with only numerical performance improvement in mind conflicts with fuzzy logic originality: its interpretability.

The goal of this section is not to propose an extensive introduction to fuzzy logic (see Zadeh, 1965; Dubois and Prade, 2000; Bouchon-Meunier and Marsala, 2003), but only to provide the reader with the basic elements of fuzzy linguistic modeling.

First, we recall how fuzzy sets are used to model linguistic concepts and then the two main steps of rule generation are detailed: variable fuzzy partitioning design and rule induction.

a- Fuzzy sets and linguistic terms

A fuzzy set is defined by its membership function. A point in the universe, x , belongs to a fuzzy set, A , with a membership degree, $0 \le \mu_A(x) \le 1$. Figure 2 shows a triangle membership function.

Fuzzy sets can be used to model linguistic concepts. If A is the set of high temperatures, the membership degree of a given temperature, x, $\mu_A(x)$, can be interpreted as the level to which the x temperature should be considered as high. The rule "If Temperature is High then ..." is implemented as "If X is A then ...". For the x temperature the matching degree of the rule is given by its membership degree, $\mu_A(x)$. Usually several variables are involved in the rule description. In this case the membership degrees are combined using a *AND* operator, the minimum and the product being the most common ones.

Several fuzzy sets, corresponding to linguistic concepts, can be defined on the same universe, e. g. low, average, high. The set of the fuzzy sets defined on the same universe form a fuzzy partition of the variable.

b- Fuzzy partitioning

The readability of fuzzy partitioning is a pre-requisite condition to build an interpretable rule base. The necessary conditions for interpretable fuzzy partitions have been studied by several authors (Ruspini, 1969; Valente de Oliveira, 1999; Glorennec, 1999; Espinosa, 2000). Let us recall the main points:

- Distinguishability: Semantic integrity requires that the membership functions represent a linguistic concept and different from each other.

- A justifiable number of fuzzy sets.

- Coverage: Each data point, x, should belong significantly, $\mu(x) > \varepsilon$, at least to one fuzzy set.

- Overlapping: All the fuzzy sets should significantly overlap.

We implement these constraints as follows:

$$\begin{cases} \forall x \sum_{f=1,2,\ldots,m} \mu^f(x) = 1 \\ \forall f \exists x \mu^f(x) = 1 \end{cases} \quad (1)$$

where m is the number of fuzzy sets in the partition and $\mu^f(x)$ is the membership degree of x to the *fth* fuzzy set. Equation 1 means that any point belongs at most to two fuzzy sets when the fuzzy sets are convex.

We choose fuzzy sets of triangular shape, except at the domain edges, where they are semi trapezoidal. A triangle fuzzy set f is defined by its breakpoints $left^f$, c^f, $right^f$. Conditions from equation 1 are implemented by choosing fuzzy set breakpoints as shown in figure 3, defining a standardized fuzzy partition.

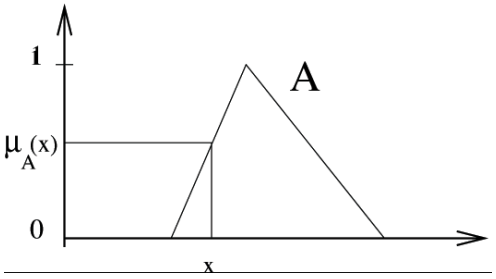


Figure 2. A triangle membership function.

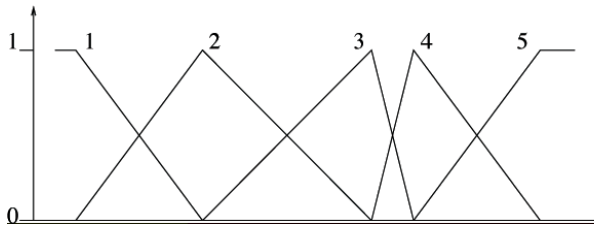


Figure 3. A standardized fuzzy partition.

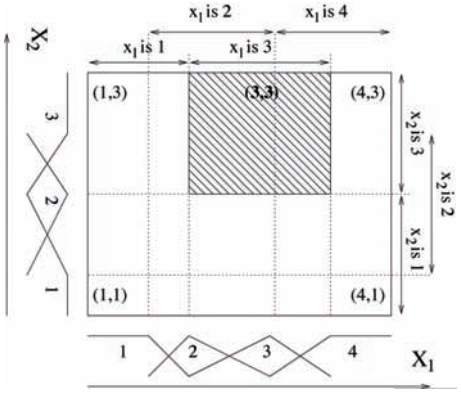


Figure 4. From partitions to fuzzy rules.

These kind of partitions can be either designed by a domain expert or learnt from data.

c- Fuzzy rule generation

The next phase of FIS design consists in rule generation from the previous unidimensional partitions. The goal is to produce a small number of general rules.

The automatic rule generation process involves three fundamental steps: building all the possible rules according to partition structures, select only the rules which are supported by data, and, finally, initialize rule conclusions.

1. Building all the possible rules

Let us examine figure 4, which illustrates a 2-dimension system. The input variables are X_1 *and* X_2. The corresponding fuzzy partitions are shown below X_1 axis, and on the left side of the X_2 axis. The label (3,3) is given to the following rule "If X_1 is 3 and if X_2 is 3 then Output is C".

This rule is, more or less, true for X_1 and X_2 values falling in the hatched area of figure 4. As, in this example, the X_1 (X_2) fuzzy partition is made up of 4 (3) membership functions (or linguistic terms) the number of possible rules, which depends only on the partition sizes, is 12. They are: (1,1), (1,2) …(4,3).

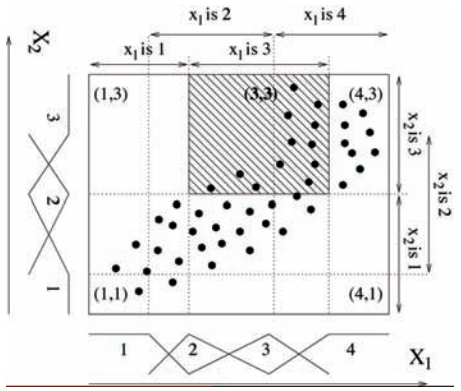


Figure 5. Possible rules and learning data

2. Confrontation with data

Then the possible rules are confronted with data, as illustrated by the scatter plot in figure 5.

Since several points of the sample correspond to the (3,3) rule, this rule will be generated. But the (1,3) rule corresponds to an empty area: there is no data point firing this rule. In this case, rule generation using an automatic algorithm does not make sense.

3. Conclusion initialization

Rule conclusion is generally computed from the observed output of the items corresponding to the rule. The computation also takes into account the matching degree of the different items.

The Fuzzy Inference System design and management have been carried out using an open source software called FisPro (Guillaume *et al.*, 2002). Among the available methods for fuzzy rule induction (Guillaume, 2001), FisPro only implements those which yield interpretable fuzzy rules. We use a refinement algorithm (Guillaume and Charnomordic, 2004) to determine the number of linguistic terms for each of the variables. The refinement procedure starts by considering the simplest system, which has only one rule including a single fuzzy set for each variable. The selection procedure builds new systems by refining the fuzzy partitions. The key idea is to introduce as many variables, described by a sufficient number of fuzzy sets, as necessary to get a good rule base. A good system represents a reasonable trade-off between complexity, in relationship with the number of rules, and accuracy, measured by the performance index.

The performance index (PI) is computed as follows:

$$PI = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left\|\hat{y}_i - y_i\right\|^2} \quad (2)$$

n being the number of items in the sample, y_i the observed output for the *ith* example and y_i the one inferred by the system.

RESULTS

We started our study with a first Principal Component Analysis followed by an expert analysis of maturity, knowing the PCA results. This led to expert assumptions concerning the causes of within field sugar content variability that occurred at harvest. In order to achieve the same goal but without the expert, interpretable fuzzy rules induction methods were applied on the same dataset.

1. Classical Analysis

The data set is now analyzed using classical means. First data are processed using a Principal Component

J. Int. Sci. Vigne Vin, 2007, **41**, n°1, 19-31

a b

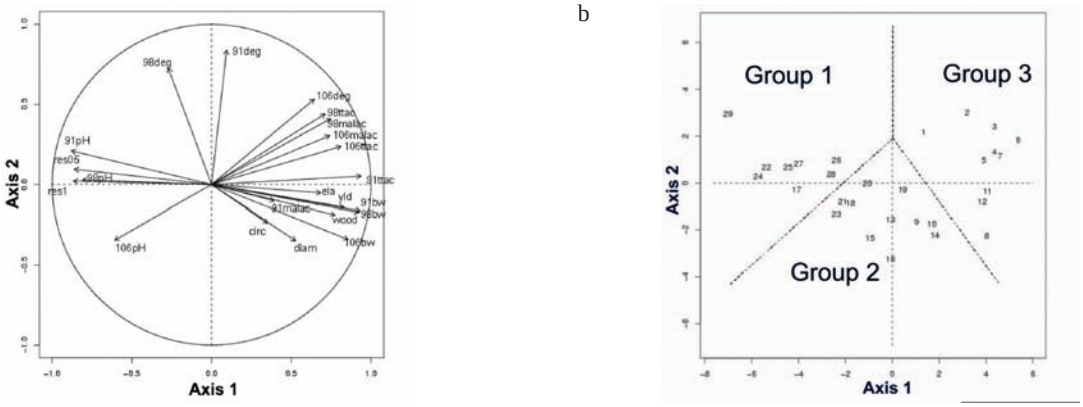


Figure 6. PCA plots: Main plane variable and example projection.

Analysis, then ripening spatial variability is analyzed by an expert.

a- Principal Component Analysis

A PCA is carried out on the whole dataset. Since the samples are georeferenced, the PCA plots correspond to geographical zones. Figure 6 shows the variable and sample projections on the main plane. The explained variance is 65%, 53% by the first axis, 12% by axis 2 resulting in 3 groups of data as seen in figure 6b.

Two groups of variables are clearly identified according to their first axis coordinates: soil resistivity (res05 and res1) and pH (91pH, 106pH and 98pH) on the negative side, yield (yld), berry weight (91bw, 98bw and 106bw) and all of the parameters related to vigour (weight of wood -wood-, exposed leaf area -ela-) and acidity (91ttac, 106ttac, 106malac, 98malac) on the positive side. Axis 1 can be interpreted as a soil-plant-vigour gradient.

Axis 2 is mainly explained only by two sugar content variables (98deg,91deg). Our system output, harvest sugar content (106deg), is located between the 2 axis.

The principal components are orthogonal, meaning no significant linear correlation between system output (harvest sugar) and the available input variables is to be found.

Once the axis are interpreted, let us have a look to the sample location in this new space. The examples are distributed according to a East-West gradient: their coordinates on the first axis, bottom part of figure 6, roughly correspond to their location in the field as underlined by figure 7.

According to PCA results, we may conclude that the first axis separates the different examples according to a soil-plant-vigour interaction criterion.

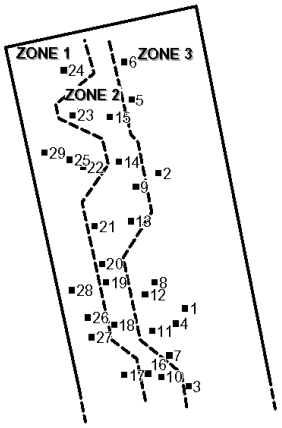


Figure 7. Field's map of the three zones corresponding to the three groups identified on the PCA.

The West side of the field is characterized by low vigour, low yield and high soil resistivity while the East side is characterized by the opposite.

To confirm this hypothesis an independent expert analysis was carried out.

b- Expert analysis of maturity

The expert works for the californian vineyard industry and has been developping extensive viticultural research program for over 6 years .

Expert has no prior knowledge of the Spanish experimental field and was only presented with the raw data after their acquisition. Based on a global analysis as well as on statistical computations over the three groups, his main conclusions are summarized as follows:

1. The field presents an increasing vigour (i.e. an increasing plant vegetative development) from the western zone (zone 1), to the eastern zone (zone 3). This is shown by the increasing values from West to East of all the variables except the sugar content and the pH. As expected

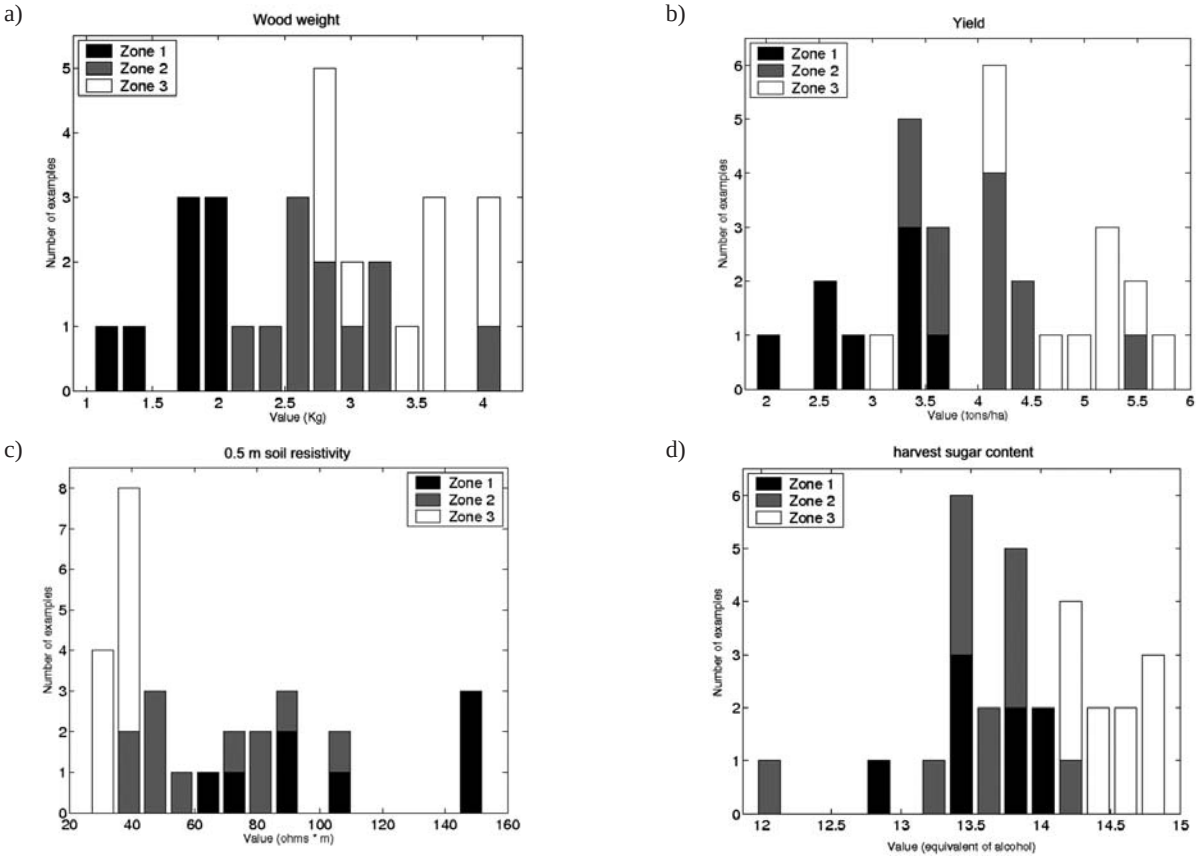


Figure 8. Some zone labelled distributions

when vigour level is high, we observe for each sampling date that pH values decrease from West to East. An intermediate central zone for vigour appears for some variables, as shown on figure 8a, and to a lesser extent, 8b. This vigour gradient is related to an increasing water supply gradient following the East-West direction (i.e. inversely related to an increasing soil resistivity gradient). Figure 8a shows that vine wood weight (an indirect measurement of leaf area) increases from zone 1 to zone 3. Figure 8c shows that soil resistivity decreases from West (zone 1) to East (zone 3) according to a gradient similar to leaf area spatial variations. Relationships between soil resistivity, vine development and berry ripening variations have been recently reported by Goulet and Barbeau (2006). Our data are in agreement with this report.

2. High water availability and high leaf area development are often associated with a slower sugar accumulation rate and lower sugar content at harvest, as recently reported by Santesban and Royo (2006). For that reason, we expect to observe an increasing gradient for sugar content at harvest from zone 3 to zone 1, inversely related to the increasing vine leaf area gradient from zone 1 to zone 3.

3. However, figure 8d shows that zone 3 is clearly the one containing the highest sugar content values. Sugar accumulation results from highly variable mechanisms. One assumption was formulated to explain, at least partly, this phenomenon.

In zone 3, high soil water availability at the beginning of season (low soil resistivity) stimulated vine growth. In this area, vigourous vines developed the greatest leaf area (figure 8a). This observation is in agreement with reports from Bindi *et al.* (2005) and Pellegrino *et al.* (2006) about the positive effect that higher soil moisture availability has on leaf area development. The greatest yield are also obtained on vines located in zone 3 (figure 8b). Higher fruit load has been reported by Naor *et al.* (1997) to increase vine water stress. For that reason, as the season progresses, and as the soil dries out, vines with a high transpirative area and a high yield become more susceptible to water stress. A high evaporative demand towards the end of the ripening period, particularly during hot vintages like 2003, made the vines with high leaf area and high yield more vulnerable to drought. As a result of drought and vine higher water stress level, berry dehydration is observed at the bunch level. The water loss experienced by the berry leads in turn to an increase in berry sugar content. Unfortunately, because time and spatial variations of soil moisture content were not recorded throughout the ripening period, it is not possible to validate this assumption.

It is generally accepted that moderate water stress controls leaf area development (Bindi et *al.*, 2005) and has a positive impact on fruit characteristics (Koundouras *et al.*, 2006). For those reasons a higher sugar content (and a higher fruit quality) is generally expected to be coming from areas showing a moderate level of leaf area development.

In the analysis of sugar content variations in zone 3, we can distinguish 2 phases during the ripening period.

First, the rate of sugar accumulation between September 1st and September 8th is the slowest in zone 3. Higher yield and potentially higher levels of soil water available at the beginning of September, possibly account for the slow rate of sugar accumulation. Negative effect of high water availability and high yield on sugar content and sugar accumulation rate have been recently reported by Koundouras *et al.* (2006) and Santesban and Royo (2006). Our data support these observations.

Second, after September 8th and until October 6th, we observed the fastest rate of sugar accumulation in zone 3. This rapid change in sugar accumulation rate can be attributed to the berry loss of water. As explained earlier, berries coming from the most vigourous vines are more likely to experience berry dehydration symptoms towards the end of the season.

Finally, the expert gives a possible mechanism for the elaboration of harvest sugar content. He attempts to explain why sugar is not linearly correlated with time stable parameters like soil resistivity spatial structure and especially the vigour.

There are two distinct ways leading to a faster sugar accumulation rate within the berries. The first way is the most classic: a low to moderate vigour level, associated with a lower yield lead to a faster sugar accumulation rate. The second way results from a berry drying phenomenon: a faster sugar accumulation rate is observed in response to berry concentration by water loss, exacerbated by vigorous early season growth.

2. Interpretable fuzzy rule induction

The goal of this section is to extract knowledge, formalized as a set of interpretable fuzzy rules, from the dataset.

The result analysis particularly focuses on:

- System ability for non-linearity management;
- Rules relevancy, compared to the explanations given by the expert.

In order to build a predictive system, physiological measurements taken on harvest date are not used. Thus, the remaining available variables are either time stable characteristics, or physiologic data assessed at least one month before the harvest.

Several FIS can be designed, according to different algorithm parameters. Let us focus on a simple one made up of only two input variables, exposed leaf area (Ela) and 1/09/03 sugar content (1/09/03sc), the output being the harvest sugar content (Hsc).

a- FIS characteristics

Let us first underline the weak level of linear correlation between the three variables involved in the system:

- (Ela, Hsc): r=0.406;
- (Ela, 1/09/03sc): r=-0.062;
- (1/09/03sc, Hsc): r=0.562;

Both the input variables and their number of linguistic terms are automatically selected by a refinement algorithm (Guillaume and Charnomordic, 2004) available in FisPro (Guillaume *et al.*, 2002). The goal is to design a compact and accurate system.

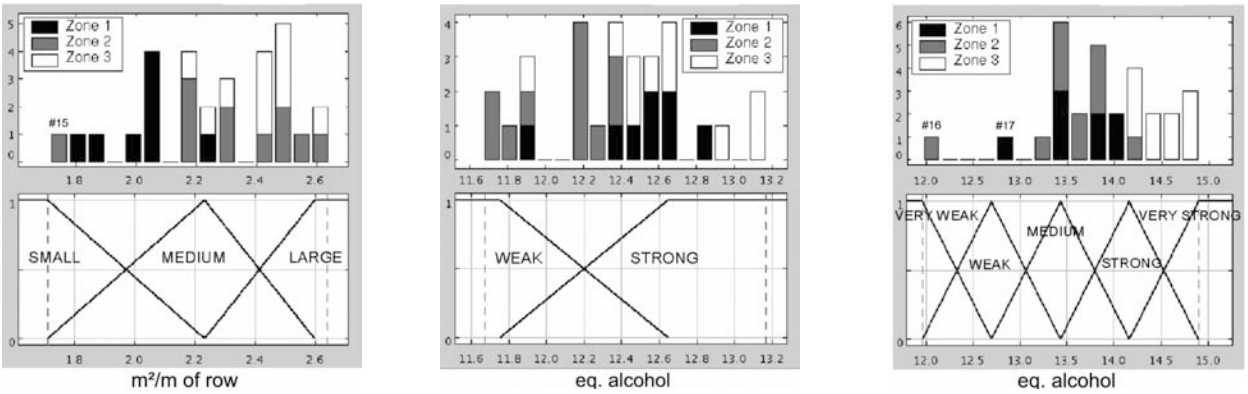
The first selected variable is exposed leaf area. This is consistent with the expert analysis: this emphasizes the importance of the vigour gradient to explain the final sugar content. The exposed leaf area is assumed to be related to this gradient, at least between West and East. It is interesting to note that vigour also influences the PCA result, by defining the first axis. This highlights once again the fact that this problem is not uni-dimensional. The second selected input is 1/09/03 sugar content variable, which is orthogonal to the first axis of the PCA. This selection is discussed in the next section.

The exposed leaf area variable is partitioned into three linguistic terms, while only two are chosen for the 1/09/03 content. The output universe is divided in a five term regular grid. The partitions, joined to the corresponding data distribution, are plotted in figure 9.

The fuzzy set limits are meaningful to an expert, each of the fuzzy sets can be assigned a linguistic label:

- Exposed leaf area: small, medium, large;
- 1/09/03 sugar content: weak, strong;
- Vintage Harvest sugar content: very weak, weak, medium, strong, very strong.

The output histogram (figure 9 (c)) shows that the two lowest values (examples #16 and #17) are quite isolated. We can expect some difficulties to obtain good interpolations to reproduce these values, as the induction process aims to generate general rules.



(a) Exposed leaf area (b) 1/09/03 sugar content (c) Harvest sugar content

Figure 9. Input and output variable partitions and histograms.

Rules

Rules Display

Rule	Active	IF Ela	AND 1/09/03sc	THEN Hsc
1	✔	medium	strong	very strong
2	✔	large	strong	very strong
3	✔	medium	weak	weak
4	✔	large	weak	strong
5	✔	small	strong	medium
6	✔	small	weak	medium

Figure 10. The induced rule base.

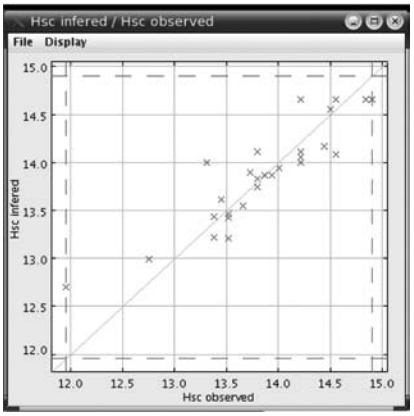


Figure 11. FIS Observed/Inferred plot.

The rule base, figure 10, is made up of six rules, all the possible rules have been designed. The performance index, *PI* , defined in equation 2, is 0.275 equivalent alcohol degree with a maximal error of 0.758. This high value is due, as expected, to the outlier (11.96 degree) that is not properly interpolated, as shown by the inferred/observed plot in figure 11.

b- Evaluation of extracted knowledge

Each rule is interpretable. To make a rule base analysis, let us consider rule conclusions only.

In this evaluation, rule #6 is not considered since it is a specific rule significantly activated by only one example (#15), with a very strong matching degree. Example#15 is the minimal value of the dataset for the exposed leaf area and is quite atypical in the eastern part of the field. It activates only rule #6. This example could correspond to a measurement error. If this rule and this example are both removed, the FIS performance doesn't change.

Rules #1, #2 and #4 have their conclusions defined by the labels “very strong” or “strong”. Rules #2 and #4 have an antecedent defined by “large” for exposed leaf area. Rule #1 cannot be analyzed alone and will be discussed later.

Rule#5 leads to the medium harvest sugar content conclusion. It is the only rule defined by the “small” label for exposed leaf area. Again, this is consistent with the expert analysis, who found the same link between exposed leaf area and the output.

Rule #3 is the only one leading to a “weak” sugar content, and is defined by the “medium” label for exposed leaf area. Rule #3 probably applies to a transition zone where two different mechanisms of sugar accumulation are observed. In this “transition zone” the 2 different mechanisms of sugar accumulation leading to a high sugar content are observed.

The expert suggested that differences of exposed leaf area were related to differences in sugar content. Note that most of this information is given in a very similar way both by the expert and by the linguistic rules. This semantic agreement validates the whole induction process which includes input variables selection, fuzzy partitioning design as well as rule generation.

To go beyond a simple linguistic analysis, the spatial influence of the rule are mapped (always skipping rule #6). Our purpose is to confirm that as expected, expert zones correspond to rule zones. The rules are examined in turn, representing their influence zones. To build a given rule map we interpolate matching degrees using a simple inverse distance method.

Then the resulting map is thresholded, grey zones include values higher than the threshold.

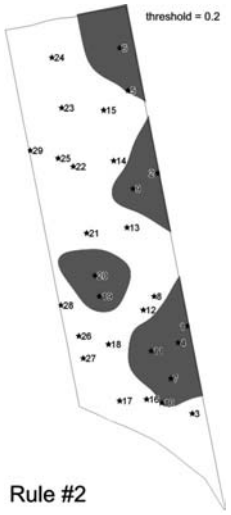


Figure 12. Rule 2 influence map.

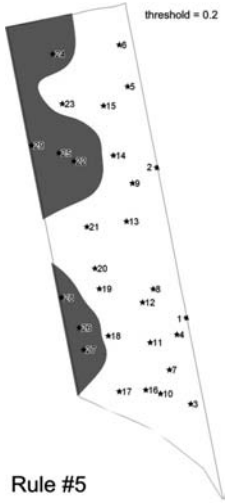


Figure 14. Rule 5 influence map.

1. Rule #2:

The rule is "If Ela is *large* and 1/09/03sc is *strong* then Hsc is *very strong*".

Figure 12 shows that the spatial influence of the rule corresponds primarily to the eastern part of the field . The rule influence map matches the high vigour/high sugar content zone mentioned by the expert. As previously noticed this area corresponds to a particular behaviour of this vineyard. (the activation zone into the western part of the field is an artifact interpolation due to the lack of points in this zone).

2. Rule #4:

The rule is "If Ela is *large* and 1/09/03sc is *weak* then Hsc is *strong*".

Compared to rule #2, accumulation of sugar seems to be faster in the area where rule #4 applies since the strong harvest sugar content comes from a weak sugar content one month before harvest (one month before harvest, sugar content is strong with rule #2). This rule, as shown by figure 13, is specific to a small group of sites (with a high matching level), where exposed leaf area values are the greatest of the dataset. On those sites, sugar content is low on 1/09/03. The negative effect of high vigour level on sugar accumulation rate probably accounts for the lower initial sugar content. However, due to their larger exposed leaf area, those sites illustrate best the effects of the drying symptoms leading to a high sugar content at harvest as noted by the expert analysis.

3. Rule #5:

The rule is "If Ela is *small* and 1/09/03sc is *strong* then Hsc is *medium*".

Rule #5 applies to the Western part of the field for which Ela is small (figure 14, the zone corresponding

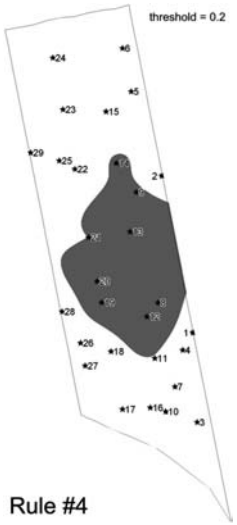


Figure 13. Rule 4 influence map.

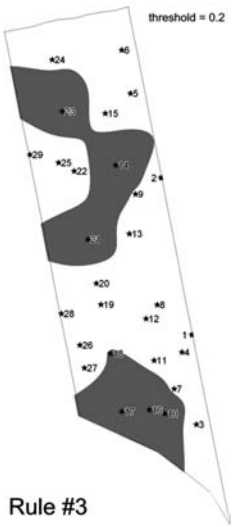


Figure 15. Rule 3 influence map.

to “low vigour” and “high sugar content”. This zone matches the low vigour/high sugar content zone underlined by the expert.

4. Rule #3:

The rule is “If Ela is *medium* and 1/09/03sc is *weak* then Hsc is *weak*”.

This rule is activated mostly into the center part of the field (figure 15), confirmed by the expert analysis. This zone was also identified by the expert. Compared to the western part of the field (rule #5), it follows the common observance of higher vigour leading to a ripening delay when no berry dehydration occurs.

5. Rule #1: The rule is “If Ela is *medium* and 1/09/03sc is *strong* then Hsc is very *strong*”.

This rule is the most difficult one to interpret. Figure 16 shows that this rule applies to both sides of the field corresponding approximately to zones 1 and 3 of the PCA.

To understand its meaning, let us examine figure 17. It shows the Ela distribution jointly with the variable fuzzy partition. We observe that the fuzzy set corresponding to medium Ela includes the whole range of data points. It means that each sample has a non-null membership degree to this fuzzy set. The samples corresponding to PCA zone 1 (western part of the field) are labelled with a “medium” Ela “from down”. “From down” means that the Ela from those vines also belong to the label “low” Ela. An illustration is given on figure 17. We see that vines labelled “medium-” Ela have Ela value lower than the value of the triangular membership function vertex. Similarly, vines from PCA zone 3 are labelled “medium” Ela “from up”. “From up” means that the Ela from those vines also belong to the label “high” Ela. For that reason, the

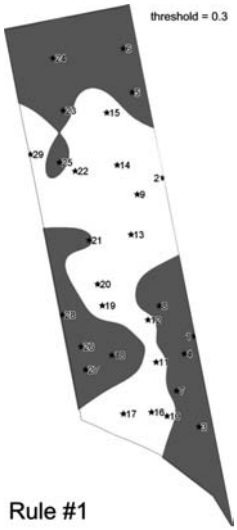


Figure 16. Rule 1 influence map.

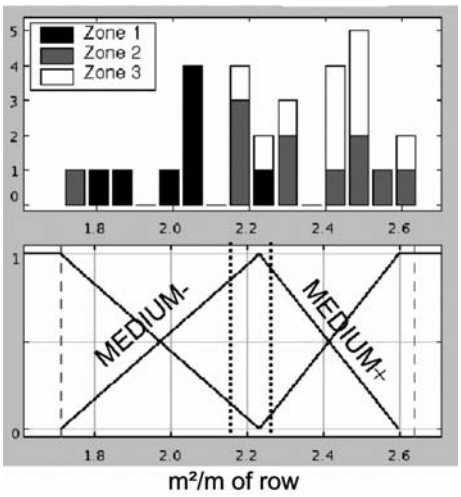


Figure 17. Focus on Ela distribution and the medium membership function.

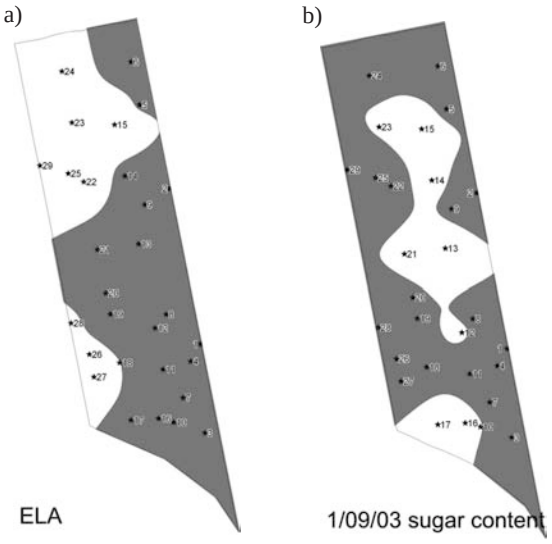


Figure 18. Within-field layout of Ela values (a) and 1/09/03 values (b).

corresponding vines are reported on the figure under the label “medium+”.

However, the data distributions plotted in figure 9 also show that most of zone 1 and 3 samples have a strong 1/09/03 sugar content. For Ela variable and 1/09/03 sugar content variable, the maps in figure 18 display the difference between the value measured in the field and a reference value. For Ela, the reference value corresponds to the medium membership function vertex. For 1/09/03 sugar content, the reference value corresponds to the intersect of weak and strong membership functions. The colored zones correspond to positive differences.

The vines labelled with “medium” Ela could also belong to “small” (West area) or “large” (East area) Ela

J. Int. Sci. Vigne Vin, 2007, **41**, n°1, 19-31

labels. As a consequence, vine with "*medium-*" Ela and firing rule #1, also fire rule #5 since they also belong to the *small* Ela label. The same is true for vine with "*medium+*" Ela and firing rule #1. Those vines also fire rule #2 since they also belong to the large Ela label. Rules #5 and #2 premises only differ in the Ela variable and the two Ela antecedents are overlapping when rule #1 applies.

This highlights that rules #5, #2 and #1 are not independent at all. Due to the membership function overlap, a given input is likely to activate a set of different rules. The inferred output is the result of individual rule output combination.

CONCLUSION

The goal of this work was to test the use of fuzzy rule induction methods on precision viticulture data. These methods are interesting in this particular area since they provide:

- An easy way to analyze multidimensional dataset. No significant skills in statistics or data analysis are required. This fits with growers and agronomists who want to focus on the data and their explanations;

- Semantic rules that are easy to understand and to manage to professionals.

This study focuses on precision viticulture data from a vineyard located in Spain. Fuzzy rule induction is used to predict the harvest quality (harvest degree) from data that were collected at the within field level before harvest. The data set includes different kinds of variables related to vine, soil but also to grape quality (sugar content, pH, titrable acidity). These last ones are available at different times during the ripening.

Besides the rule induction process, a classical expert analysis is independently performed on the data. The relevancy of our approach is assessed by a qualitative comparison of both analysis.

Results show that the machine-learning process allows to find the same trends as the expert. The extracted rules are simple to analyze as they are defined by semantic labels, similar to the ones used by expert reasoning.

The studied vineyard is particularly interesting. It involves three different zones that are difficult to identify using a classical linear data analysis (PCA), each one corresponding to a specific maturation process. This complicated and non-linear phenomenon is clearly highlighted by the fuzzy induced rules. The fuzzy system presented in this paper is quite small: three variables and six rules. Nevertheless, the rule base gives the main keys to understand the whole system while it provides a relevant prediction of the harvest quality. This preliminary work shows that fuzzy rule induction may be a relevant tool for precision agriculture data interpretation.

Dealing with spatial data, a fuzzy system has to include spatial reasoning. Further work is needed to manage the link between rules and their spatial location. This is a prerequisite to make sure the extracted rules also fit within field tractable zones.

Acknowledgements: The authors would like to thank the EVENA (Estación de Viticultura y Enología de Navarra) and the Julian Chivite winery for providing them with the data used in this work.

REFERENCES

Bindi M., Bellesi S., Orlandini S., Fibbi L., Moriondo M., Sinclair T., 2005. Influence of Water Deficit Stress on Leaf Area Development and Transpiration of Sangiovese Grapevines Grown in Pots. *Am. J. Enology Vitic.*, **56** (1), 68-72.

Bouchon-Meunier B. and Marsala C., 2003. *Logique floue, principes, aide à la décision*. Lavoisier.

Bramley R.G.V., Hamilton R.P., 2004. Understanding variability in winegrape production systems. 1. within vineyard variation in yield over several vintages. *Aust. J. Grape Wine Research*, **10**, 32-45.

Drummond S.T., Sudduth K.A., Joshi A., 2000. Predictive ability of neural networks for site-specific yield estimation. *In: Proceed. second international geospatial Iinformation in agriculture and forestry conference*, Lake Buena Vista, January 10-12, Florida.

Dubois D., Prade H., 2000. *Fundamentals of fuzzy sets*. Kluwer Academic Publishers.

Dzeroski S., Grbovic J., Walley W. J., Kompare B., 1997. Using machine learning techniques in the construction of models. ii. data analysis with rule induction. *Ecological Modelling* **95**, 95-111.

Espinosa J., Vandewalle J., 2000. Constructing fuzzy models with linguistic integrity from numerical data-afreli algorithm. *IEEE Transactions on Fuzzy Systems*, **8** (5), 591-600.

Glorennec P.-Y., 1999. *Algorithmes d'apprentissage pour systèmes d'inférence floue*. Editions Hermès, Paris.

Goulet E., Barbeau G., 2006. Contribution of soil electric resistivity measure*ments to the studies on soil/grapevine water relations. J. Int. Sci. Vigne Vin*, **40** (2), 57-69.

Guillaume S., 2001. Designing fuzzy inference systems from data: an interpretability-oriented review. *IEEE Transactions on Fuzzy Systems*, **9** (3), 426-443.

Guillaume S., Charnomordic B., 2004. Generating an interpretable family of fuzzy partitions. *IEEE Transactions on Fuzzy Systems*, **12** (3), 324-335.

Guillaume S., Charnomordic B., Lablée J.-L., 2002. *Fispro: An open source portable software for fuzzy inference systems*. http://www.inra.fr/Internet/Departements/MIA/M/fispro

Koundouras S., Marinos V., Gkoulioti A., Kotseridis Y., van Leeuwen C., 2006. Influence of vineyard location and vine water status on fruit maturation of nonirrigated cv.

Agiorgitiko (*Vitis vinifera* L.). Effects on wine phenolic and aroma components. *J. Agric. Food Chem.*, **54** (14), 5077-5086.

Naor A., Gal Y., Bravdo B., 1997. Crop load affects assimilation rate, stomatal conductance, stem water potential and water relations of field-grown Sauvignon blanc grapevines. *J. Exp. Botany*, **48**, 1675-1680.

Pellegrino A., Goze E., Lebon E., Wery J., 2006. A model-based diagnosis tool to evaluate the water stress experienced by grapevine in field sites. *Europ. J. Agronomy*, **25** (1), 49-59

Ruspini E.H., 1969. A new approach to clustering. *Information and Control*, **15**, 22-32.

Santesteban L.-G., Royo J.-B., 2006. Water status, leaf area and fruit load influence on berry weight and sugar accumulation of cv. 'Tempranillo' under semiarid conditions. *Scientia Horticulturae*, **109** (1), 60-65

Shatar T.M., Mcbratney A.B., 1999. Empirical modeling of relationships between sorghum yield and soil properties. *Precision Agriculture*, **1**, (3), 249-276.

Stamp J., 2003. *Partial rootzone drying*. Wine Business Monthlyn 01/10/2003.

Taylor J., Tisseyre B., Praat J-P., 2005. Bottling Good Information: Mixing Tradition and Technology in vineyards. *Frutic' 05 Symposium*, September 12-16, Montpellier France, 719-736.

Tisseyre B., Taylor J., Ojeda H., 2006. New technologies to characterize spatial variability in viticulture. *Proceedings of the VI international congress on terroir*, 204-217.

Valente de Oliveira J., 1999. Semantic constraints for membership functions optimization. IEEE Transactions on Systems, Man and Cybernetics. Part A 29 (1), 128-138.

White R.E., 2003. *Soils for fine wines*. Oxford University Press, New York.

Zadeh L.A., 1965. Fuzzy sets. *Information and Control*, **8**, 338-353.

