

## Combining a sensor software with statistical analysis for modeling vine water deficit impact on grape quality

Aurélie Thébaut, Thibault Scholash, Brigitte Charnomordic, Nadine Hilgert

### ► To cite this version:

Aurélie Thébaut, Thibault Scholash, Brigitte Charnomordic, Nadine Hilgert. Combining a sensor software with statistical analysis for modeling vine water deficit impact on grape quality. 2014. <hal-00863992v2>

**HAL Id: hal-00863992**

**<https://hal.inria.fr/hal-00863992v2>**

Submitted on 6 Jan 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Combining a sensor software with statistical analysis for modeling vine water deficit impact on grape quality

Aurélie Thébault<sup>a</sup>, Thibaut Scholasch<sup>b</sup>, Brigitte Charnomordic<sup>a</sup>, Nadine Hilgert<sup>a</sup>

<sup>a</sup>*INRA-SupAgro, UMR 729 MISTEA, F-34060 Montpellier, France*

<sup>b</sup>*Fruition Sciences, 34 Montpellier, France*

---

## Abstract

This work proposes a methodology using temporal data and domain knowledge in order to analyze a complex agronomical feature, namely the influence of vine water deficit on grape quality. Raw temporal data are available but they are not directly usable to estimate vine water deficit. The methodology associates advanced techniques in computer science and statistics. A preliminary step is required to determine if the amount of water effectively used by the vine is sufficient or not. This step necessitates an ecophysiological model, based on expertise. The expertise is first formalized in an ontology, under the form of concepts and relationships between them, and then used in conjunction with raw data and mathematical models to design a software sensor. Next the software sensor outputs are put in relation to product quality, assessed by quantitative measurements. This relation is analyzed by regression trees and advanced data analysis methods, such as functional data regression. The methodology is applied to a case study involving an experimental design in French vineyards. The temporal data consist of sap

---

*Email address:* [bch@supagro.inra.fr](mailto:bch@supagro.inra.fr) (Brigitte Charnomordic)

*Preprint submitted to Computer and Electronics in Agriculture Tuesday 10<sup>th</sup> December, 2013*

flow measurements, and the goal is to explain fruit quality parameters (sugar concentration and weight), using vine's water variations at key stages of vine phenological development. The results are discussed, as well as the method genericity and robustness.

*Keywords:* vine water stress, functional data analysis, ontology, expert knowledge, grape quality, regression tree, temporal data

---

## 1. Introduction

In modern Agronomy, the recent progress of sensors provides a lot of data, among them many temporal data. This opens new challenges, such as the proper calibration of these sensors, and the use of temporal data to establish relationships with product characteristics and quality. These relationships are not easy to determine because of the high variability of biological material. This can be compensated by the integration of expertise, as Agronomy is a domain that has always relied as much on experience than on science. Nevertheless, for domain knowledge to be effectively used in collaboration with mathematical models and data, an expertise formalization step is required.

Our objective in this paper is to show the interest of a formalized data and knowledge-based approach to study a complex agronomical phenomenon, namely the influence of vine water deficit on grape quality. Grape quality analytical measurements are available and well established. In contrast, vine water deficit cannot be directly measured and requires a preliminary step to relate the amount of water effectively used by the vine, in order to determine if this amount falls short of some reference amount. Typically, the *reference*

19 amount is the maximal amount of water a vine can use.

20 Various methods exist to characterize the level of water deficit experi-  
21 enced by the plant as reviewed by Jones (2004). Tissue water status can  
22 be assessed visually or by measurements of vine water potential. However,  
23 both methods have serious drawbacks. The lack of precision of visual obser-  
24 vations often leads to yield reduction before visible symptoms occurs. The  
25 pressure chamber method used to measure water potential is slow and labour  
26 intensive, especially for predawn measurement, and is unsuitable for automa-  
27 tion. In a production context, collecting predawn leaf water potential is not  
28 a practical solution. It is a destructive method, which must be performed  
29 before sunrise and is sensitive to vapor pressure deficit, making interpreta-  
30 tion difficult, see Rodrigues et al. (2012). In addition, measurements done  
31 with pressure chambers are very dependent on atmospheric conditions and  
32 vine phenological stage, see Olivo et al. (2009); Williams and Baeza (2007);  
33 Rodrigues et al. (2012); Santesteban et al. (2011). Other plant sensor-based  
34 monitoring approaches for estimating water deficit, like trunk diameter fluc-  
35 tuations, have been reported as unsuccessful for irrigation scheduling, see  
36 Montoro et al. (2011).

37 Thus, as of today, sap flow sensors are the only commercially available  
38 method to measure automatically and continuously systemic plant water use,  
39 see Ferreira et al. (2012). Sap flow sensors indirectly measure changes in  
40 stomata conductance and have recently become available. The main advan-  
41 tage of sap flow measurements is to allow automatic and continuous mea-  
42 surement of water flowing through the plant, which is directly related to  
43 transpiration, see Escalona et al. (2002); Jones (2004); Cifre et al. (2005);

44 Zhang et al. (2011). However, sap flow is a complex phenomenon. The sen-  
45 sitive measurement technique requires a complex instrumentation and tech-  
46 nical expertise for the definition of irrigation control thresholds, see Ginestar  
47 et al. (1998). Expert knowledge is necessary to convert raw data into useful  
48 transformed data, i.e. water courses, by designing a software sensor. To the  
49 best of our knowledge, no such attempt to design a sap flow software sensor  
50 has been done yet.

51 Once these data transformations are validated, it is possible to study  
52 the influence of vine water deficit on grape characteristics. The existence of  
53 relationships between vine water deficit and fruit composition has already  
54 been reported in the literature, see des Gachons et al. (2005); Koundouras  
55 et al. (2006); Van Leeuwen et al. (2009). These studies are limited to the  
56 study of vine water status scalar measurements. In the present paper, a  
57 proposal is made to use water courses, that opens the way to a range of new  
58 studies.

59 We will first show how a formalized data and knowledge-based approach  
60 can be useful to design a software sensor. Knowledge formalization will be  
61 done by using ontologies, which take increasing importance in the field of  
62 Life Sciences, see Villanueva-Rosales and Dumontier (2008); Thomopoulos  
63 et al. (2013), for their ability to model and structure qualitative domain  
64 knowledge.

65 In a second step, water use trajectories will be put in relation to grape  
66 quality indicators such as Berry Weight or Sugar Concentration, using recent  
67 data analysis tools and formalized knowledge. Innovative data analysis tools  
68 include functional data analysis that offers the possibility to use curve (func-

69 tional) data instead of scalar data. Functional data analysis has not been  
70 much used in life sciences yet, see Ullah and Finch (2013), though it could  
71 be of particular interest in the Vine and Wine Industry, and more generally  
72 for modern Agronomy.

73 The modeling task is divided into two independent parts: software sen-  
74 sor design and temporal data analysis. If the sensor design procedure were  
75 different, this would not affect the validity of the data analysis methodology.

76 The methodological work is illustrated by a case study, involving an ex-  
77 perimental design on several vineyards in the Languedoc region (France).

78 The paper is organized as follows: Section 2 presents the material and  
79 methods. It is divided into four parts. The first part gives some elements  
80 about data and the second one presents ontology-based formalization. The  
81 software sensor design, that relies on the use of mathematical models, data  
82 and formalized knowledge, is described in the third part. The illustrative  
83 example shows how it is possible to transform raw sap flow data into vine  
84 water deficit courses. The fourth part describes the data analysis methods  
85 used for analyzing the software sensor output in relation to product quality.  
86 Section 3 presents and discusses the results for vine water deficit estimation  
87 and its relationship with grape composition (Sugar Concentration, Berry  
88 Weight). Some concluding remarks and perspectives are given in Section 4.

## 89 **2. Material and methods**

90 In this section, we propose to follow four steps:

- 91 • to describe the experimental design with its input and output variables;
- 92 • to formalize ecophysiological knowledge using an ontology;

- 93 • to design a software sensor using formalized knowledge, a mathematical  
94 model, and data;
- 95 • to relate software sensor output to product quality using decision trees  
96 and functional data analysis.

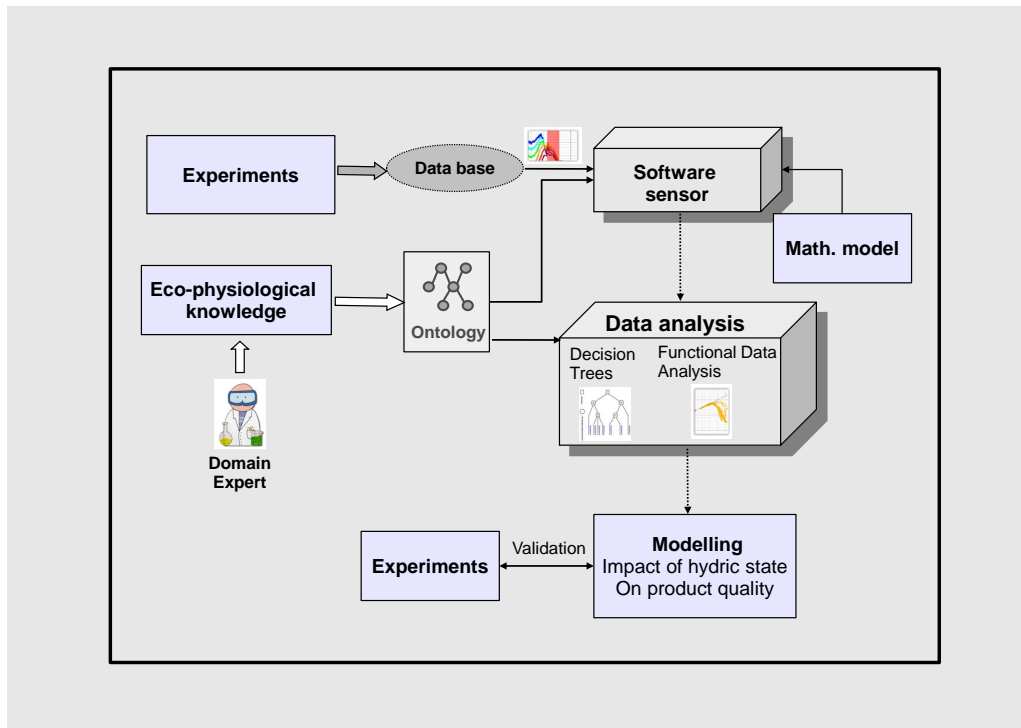


Figure 1: Outline of the proposed modeling approach.

97 Expert knowledge plays an essential part in the modeling process, and  
98 we focus on providing an efficient way to separate the data-based statistical  
99 procedures from the qualitative knowledge-based assumptions.

100 The outline of the approach is given in Fig.1. Experiments feed a data  
101 base. The software sensor integrates data from the data base, an ontology

102 and a mathematical model. Its outputs can be analyzed using data analy-  
103 sis. This analysis also calls for ecophysiological knowledge, essentially about  
104 the phenological stages. Therefore, the ontology is used at two different le-  
105 vels: software sensor design and data analysis supervision. Data analysis is  
106 performed on the basis of two complementary perspectives for determining  
107 relations between software sensor temporal output and product quality. The  
108 first line of work is to design scalar explanatory variables, by summarizing a  
109 period of interest in compliance with the ecophysiological knowledge. These  
110 variables can then be used as input to decision trees. The second line of work  
111 is to use recent advances in functional data analysis, so that the inputs to  
112 the statistical model are the temporal data as a whole.

113 In the following, the approach is illustrated with a software sensor to  
114 estimate vine water courses and their relation to grape quality. Nevertheless,  
115 the proposed methodology is generic in many aspects, and could be useful in  
116 the development of other decision support tools in Agronomy, provided that  
117 expertise and temporal data are available.

### 118 *2.1. Experimental design*

119 Data used in this paper come from a multi-site experiment located in the  
120 south of France. The same experimental design was set up in seven sites  
121 across the Languedoc Roussillon region in order to test for the effects of  
122 vine water deficit status on grape potential and wine quality in contrasted  
123 environmental conditions. In total, vine water deficit status was followed  
124 over 16 vine plots, each planted with one of the following varieties: Merlot,  
125 Cabernet-Sauvignon, Grenache or Chardonnay.

126 To get a wider range of vine water statuses during the season, an irrigation



127 treatment was applied for two years on each of the eight site-variety com-  
128 binations. The irrigation treatment consisted of two modalities, replicated  
129 twice, yielding 32 experimental subplots. In the non irrigated subplots, vines  
130 only received natural precipitations during the growing season while in the  
131 irrigated subplots, vines received regular extra-amounts of water through  
132 drippers line (emission rate from 2 to 4 l.h<sup>-1</sup>, 1 to 2 drippers per plant).

133 Several kinds of data, collected according to an experimental design, are  
134 available: local meteorological data, vine water deficit related measurements,  
135 phenological state assessments, as well as grape quality analyzes.

#### 136 *2.1.1. Meteorological data*

137 Hourly meteorological data on wind speed (km.h<sup>-1</sup>), minimal, maximal  
138 and mean air temperature (°C), air humidity (%), solar radiation (W.m<sup>-2</sup>)  
139 and amounts of precipitations (mm) were extracted from local meteorological  
140 stations for each site.

#### 141 *Transformed data*

142 Hourly vapor pressure deficit (*VPD*) and reference atmospheric evapora-  
143 tive demand (potential evapotranspiration  $ET_{ref}$ ) were calculated according  
144 to methodologies referred to as FAO-56, see Allen et al. (1998). Calculation  
145 of reference atmospheric evaporative demand ( $ET_{ref}$  in mm.d<sup>-1</sup>) is based on  
146 Penman-Monteith formula.

147 Daily meteorological data were obtained from hourly data after a trapeze  
148 integration. Thermal time, *i.e.* the accumulation of growing degree days  
149 (GDD) from April 1<sup>st</sup>, was calculated by daily integration of mean air tem-  
150 perature minus a base temperature of 10°C, which is considered as the sim-

151 plest model to estimate vine phenology, see Parker et al. (2011).

### 152 *2.1.2. Phenological data*

153 The main phenological phases (budbreak, bloom, nouaison, veraison)  
154 were estimated visually in each experimental plot when 50% of the plants  
155 reached the stage. Bloom was observed when 50% of the clusters had the  
156 cap off. Nouaison was defined using the bloom stage, according to local ex-  
157 pert knowledge (see Section 2.2.1). Veraison dates were recorded when 50%  
158 of the fruit had turned red.

### 159 *2.1.3. Vine water status data*

160 Vine water status was monitored by two kinds of measurements: discrete  
161 measurements of leaf water potential at predawn ( $\Psi_b$ , or predawn LWP) and  
162 continuous measurements of sap flow.

#### 163 *Leaf water potential at predawn*

164 LWP measurements were conducted every week from the end of June to  
165 the end of August with a pressure chamber at predawn (between  $\approx 3.00$  am  
166 and  $\approx 5.00$  am).

#### 167 *Sap flow*

168 The energy balance method (Sakuratani, 1981) was used to measure sap  
169 flow with Sap IP system (Dynamax, Houston, TX, USA). There is one variety  
170 per vineyard site. The vineyard site is divided into 2 irrigation treatments.  
171 Two vineyard rows were selected. One row represents one irrigation treat-  
172 ment. In each selected row, 2 vines were equipped with one sensor. Each  
173 sensor measured vine sap flow rate every 15 minutes. The 2 selected vines  
174 were within 25 meters of each other within the same row.

175 Sap flow rates measured on each vine were averaged on an hourly basis  
176 within each row. Total sap flow of each vine was calculated as the product  
177 of sap flux density and cross sectional sap wood area at the measurement  
178 point. Various expert methods were applied to filter out nighttime, weak  
179 and erroneous signals. Sap flow measurements were scaled at the plant level  
180 according to plant leaf area estimates corresponding to each sensor. The  
181 daily sap flow assumed to measure daily vine transpiration was computed by  
182 adding all hourly sap flow rates measured during the day. The volumetric  
183 flux per vine ( $\text{g}\cdot\text{h}^{-1}$ ) was converted into  $\text{mm}\cdot\text{h}^{-1}$  taking into account the  
184 respective area of ground per vine. Daily vine transpiration will be noted  
185  $T(t)$ .

#### 186 *2.1.4. Fruit composition quality data*

187 Starting two weeks before harvest, fruit was sampled for each irrigation  
188 treatment in each vineyard. Fruit data was collected at three different dates.  
189 Fruit composition analysis focused on berry weight (g), sugar concentra-  
190 tion ( $\text{g}\cdot\text{l}^{-1}$ ), acidity ( $\text{g}(\text{H}_2\text{SO}_4)\cdot\text{l}^{-1}$ ), anthocyanins and assimilable nitrogen  
191 ( $\text{mg}\cdot\text{l}^{-1}$ ).

#### 192 *2.2. Formalizing knowledge*

193 In this section, our aim is to show how ontologies can be used to formalize  
194 domain knowledge. In information science, an ontology formally represents  
195 knowledge as a set of concepts within a domain, and the relationships between  
196 pairs of concepts.

197 Ontologies are becoming increasingly popular, due to the great amount of  
198 available (complex) data and to the need for model (qualitative) knowledge

199 and structural information. This need first arose out of the development of  
200 the World Wide Web. However, there are still very few attempts to combine  
201 ontologies and statistical or data-driven models. This could be particularly  
202 useful in Life Sciences and Agronomy, see Villanueva-Rosales and Dumontier  
203 (2008); Thomopoulos et al. (2013); Destercke et al. (2013).

204 The main incentives for using ontologies, see Guarino et al. (2009), are  
205 the following ones:

- 206 1. To share a common understanding of structured information, as advocated  
207 in Musen (1992);
- 208 2. To explicit the specificities of domain knowledge;
- 209 3. To identify ambiguous or inappropriate model choices.

210 For the present work, a specific ontology has been built, in order to for-  
211 malize the concepts and relations required to design a vine water deficit  
212 indicator and to analyze its impact on grape quality.

213 The general class diagram of the ontology, called Ontology of Vine Water  
214 Stress (OVWS), is shown on Figure 2 as a Unified Model Language (UML)  
215 diagram. It is composed of concepts, represented as rectangular boxes, and  
216 of relations, represented by arrows. Formally, the ontology  $\Omega$  is defined as a  
217 tuple  $\Omega = \{\mathcal{C}, \mathcal{R}\}$  where  $\mathcal{C}$  is a set of concepts and  $\mathcal{R}$  is a set of relations.

218 Let us comment the main concepts and relations.

### 219 2.2.1. Concepts

220 In this ontology, four kinds of primary concepts were defined: *Variable*,  
221 *Condition*, *Constraint* and *ShiftStage*. All other concepts are sub-concepts  
222 of these primary ones and linked to them by a *subsumption* relation, as ex-

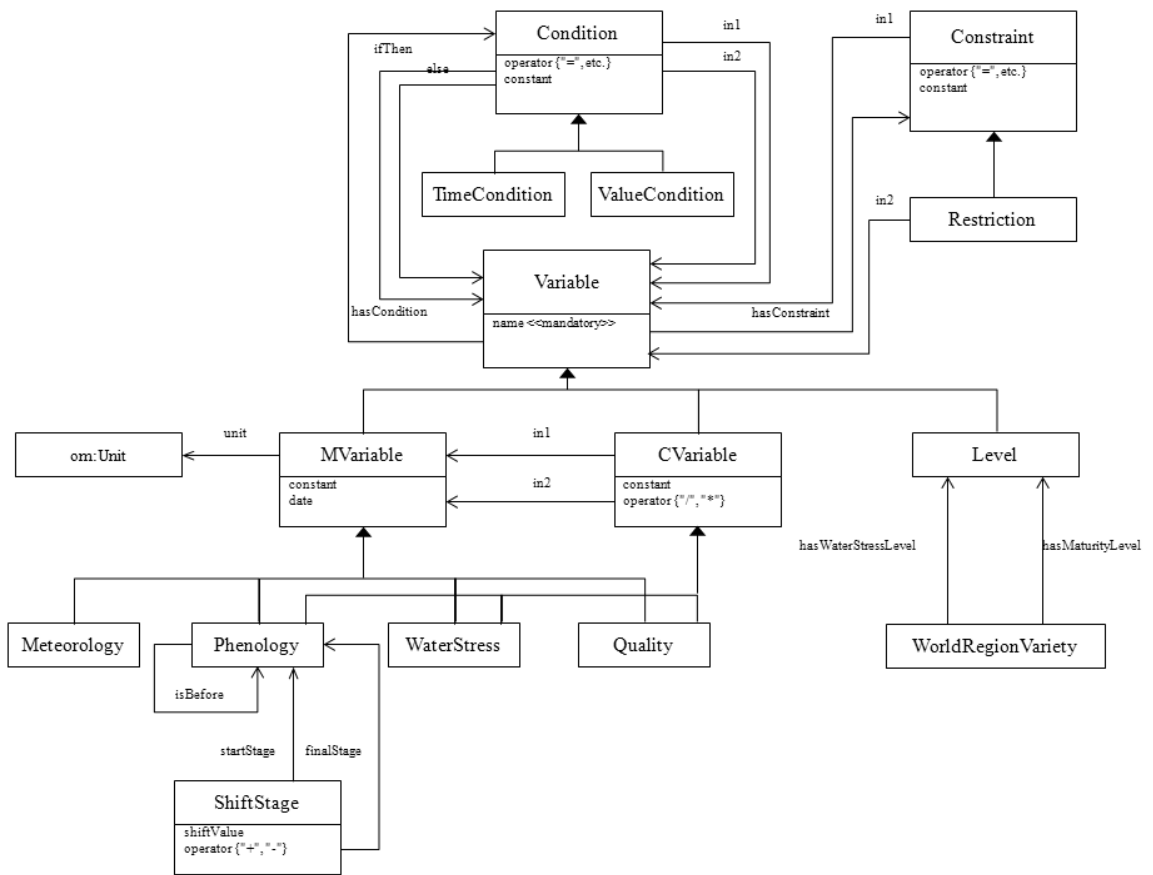


Figure 2: Class diagram of the ontology of Vine Water Stress.

223 plained in Section 2.2.2. For instance, in Figure 2, *Meteorology*, *Phenology*,  
224 *WaterStress* and *Quality* are sub-concepts of *Variable*.

- 225 • All *variables* must have a name, they can have a date, a unit and a  
226 default value. The units are taken from *OM*, an ontology of units of  
227 measures and related concepts, see Rijgersberg et al. (2013).
- 228 • The *Condition* concept is defined with a comparison operator and two  
229 operands. It will be used together with the *hasCondition* ( $\mathcal{HCO}_c$ ) re-  
230 lation, defined in Section 2.2.2.
- 231 • The *Constraint* concept is defined with a comparison operator and  
232 one operand. It will be used together with the *hasConstraint* ( $\mathcal{HCS}_c$ )  
233 relation, defined in Section 2.2.2. The *Restriction* concept is a sub-  
234 concept of *Constraint*, and is a specific two-fold constraint.
- 235 • The *ShiftStage* concept is proposed in order to determine a phenological  
236 stage from another one. This is the case for the *Nouaison* stage, which  
237 is not generally observed. Its date can be estimated by shifting the  
238 *Bloom* date by  $k$  *GDD*, where  $k$  can be variety-dependent. *Nouaison*  
239 and *Bloom* are instances of the *Phenology* concept.

#### 240 2.2.2. Relations

241 On Fig. 2, there are two kinds of arrows: thick-headed arrows and regular  
242 ones. The former correspond to the *subsumption* relation, and the latter to  
243 the other relations. In that last case, the arrow label gives the relation name,  
244 for instance *hasCondition*.

- 245 • The *subsumption* relation, also called the ‘kind of’ relation and denoted  
246 by  $\preceq$ , defines a partial order over  $\mathcal{C}$ . Given a concept  $c \in \mathcal{C}$ , we denote  
247 by  $\mathcal{C}_c$  the set of sub-concepts of  $c$ , such that:

$$\mathcal{C}_c = \{c' \in \mathcal{C} | c' \preceq c\}. \quad (1)$$

248 For example, in Figure 2, let us consider the concept  $c = Variable$ .  
249 We have  $\mathcal{C}_{Variable} = \{MVariable, CVariable, Level\}$ , where *MVariable*  
250 represents a measurement available in a data base, *CVariable* a vari-  
251 able calculated following a given method, and *Level* a constant value  
252 depending on some other concepts.

- 253 • The *subsumption* relation can be multiple. For instance a *Phenological*  
254 *concept* can be such as  $c \preceq CVariable$  or  $c \preceq MVariable$ .
- 255 • The *isBefore* relation allows to represent temporal precedence. It is  
256 very important for checking the consistency of the phenological stage  
257 dates, where *bloom* has to occur before *veraison*, and so on.
- 258 • The HasCondition ( $\mathcal{HCO}_c$ ) relation, where  $c$  represents the concept on  
259 which the condition is to be applied, is used together with a condition.
- 260 • Similarly, the HasConstraint ( $\mathcal{HCS}_c$ ) relation allows the application of  
261 a *constraint* on the  $c$  concept.

262 In Section 2.3.2, examples will be given to illustrate the interest of the  
263 ontology for designing the software sensor.

264 The ontology is modeled using the Web Ontology Language (OWL). OWL  
265 is a semantic markup language for publishing and sharing ontologies on the

266 World Wide Web, which is specified using W3C<sup>1</sup> recommendations. The  
267 use of OWL allows reusing ontologies developed elsewhere, for instance the  
268 Ontology of units of Measure (*OM*)<sup>2</sup>.

### 269 *2.3. Design of the software sensor for vine water deficit estimation*

270 Based on the knowledge formalized in the ontology given in Fig.2 and on  
271 a mathematical model, established by Ferreira et al. (2012), a software sensor  
272 is required to transform raw data from sap flow sensors into a significant vine  
273 water deficit estimator, denoted by  $Ks(t)$ .

274 The software sensor is a relatively complex information system that per-  
275 forms different functions and associates various technologies. Its design can  
276 benefit from using a conceptual framework including several viewpoints, such  
277 as the ones proposed by a “4+1” viewpoint set, introduced by Kruchten  
278 (1995) or RDM-OP approach (Reference Model for Open Distributed Pro-  
279 cessing), described in Raymond (1995).

280 For instance, the RDM-OP framework defines a set of five viewpoints:  
281 enterprise, information, computational, engineering and technology. Among  
282 them, the information view describes the way that the architecture stores,  
283 manipulates, manages, and distributes information. The computational view,  
284 presented in Fig.3, contains an object-oriented model of the functional struc-  
285 ture of the system, with a particular focus on interfaces and interactions.  
286 Each component (rectangle) is a modular part of the system, interfaces are  
287 represented by connectors with circles, and dependency between components

---

<sup>1</sup><http://www.w3.org/TR/>

<sup>2</sup><http://www.wurvoc.org/vocabularies/om-1.6/>



288 is illustrated by dashed arrows.

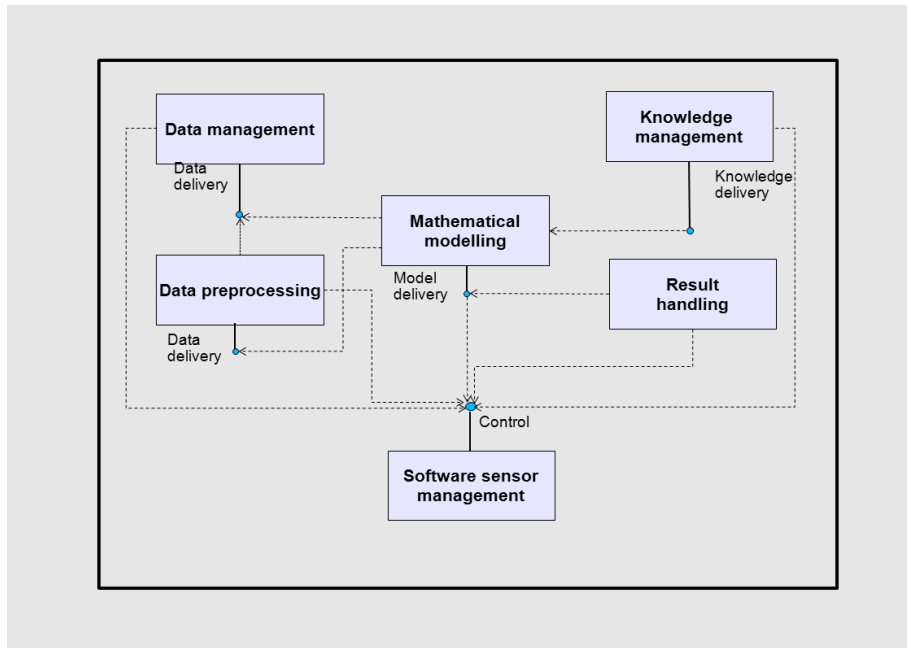


Figure 3: Software sensor computational view.

289

290 Note that each component can be modeled independently, using a suitable  
291 language (sql, OWL, R). The communication between the various compo-  
292 nents can be implemented by a high level interface, written in Python, PHP  
293 or Java, according to the technology viewpoint. In its present implementa-  
294 tion, the software sensor is available as a desktop application to the members  
295 of the *Pilotype* project (see Acknowledgments).

296 More details about the components of the desktop application are given  
297 in Fig.4 and the various steps to follow for estimating the vine water deficit  
298 ( $K_s$ ) are detailed below.

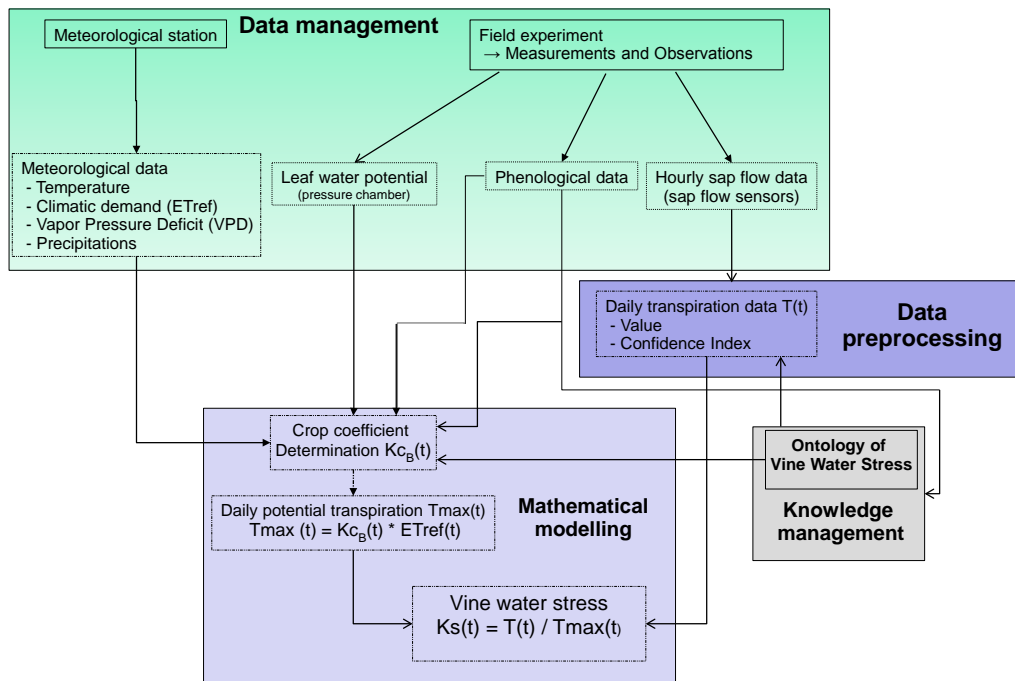


Figure 4: Software sensor component details for the vine water deficit ( $K_s$ ) estimation.

300 *2.3.1. Sap flow under limiting soil water condition: computation of  $Ks$*

301  $Ks$  is the ratio between actual and maximum crop transpiration, defined  
302 as:

$$Ks(t) = \frac{T(t)}{T_{max}(t)} \quad (2)$$

303 It accounts for the decline in vine water use due to soil moisture deficit.  $Ks$   
304 represents the level of daily vine water use by reference to its maximal level.  
305  $Ks = 1$  reflects a situation when maximal level of vine water use is fully  
306 reached/ satisfied. When  $Ks < 1$ , the maximal level of vine water use is not  
307 reached. Daily vine water use is limited and  $Ks$  level indicates some level of  
308 water deficit. Arbitrarily we characterize this situation as a 'stress'. When  
309  $Ks = 0$ , stress is maximal.

310 Allen et al. (1998) have presented a general proposal for estimating  $Ks$ .  
311 Ferreira et al. (2012) have reported results showing the variations of specific  
312  $Ks$  in vineyard subjected to contrasted soil moisture regimes. Functions for  
313 vineyards, from field experiments, are not generally available.

314 In the vine context, in Eq.2,  $T$  is the daily measured transpiration from  
315 sap flow and  $T_{max}$  is the daily maximal vine transpiration obtained under  
316 dry soil condition (meaning no cover crop) when soil moisture is non limiting,  
317 defined as in Allen et al. (1998).

$$T_{max}(t) = Kc_B(t) ET_{ref}(t) \quad (3)$$

318  $ET_{ref}$  is the reference evapotranspiration and  $Kc_B$  a coefficient linearly  
319 related to the leaf area index (LAI) or to the fraction of ground coverage, see  
320 Picón-Toro et al. (2012). Consequently a site-specific determination of  $Kc_B$

321 is necessary for each vineyard to account for differences due to canopy size  
322 and planting density.

323 *2.3.2. Sap flow under non limiting soil water condition : computation of dry*  
324 *soil  $K_{CB}$*

325  $K_{CB}(t)$  is vine specific and varies with leaf area development. When  
326  $K_{CB}(t)$  is multiplied by  $ET_{ref}(t)$ , it yields an estimate of plant maximal  
327 transpiration, which is the volume of vine water use in absence of soil moisture  
328 deficit (see Eq.3). We propose to use formalized concepts and relations based  
329 on expertise, all of them implemented in the OVWS ontology. We divided  
330 the  $K_{CB}(t)$  profile into two main growth stages:  $L_{dev}$  and  $L_{mid}$  as presented  
331 in Fig.5. This profile is derived from the FAO segmented crop profile for 2  
332 growing stages (development period and mid-season period) as reported by  
333 Allen and Pereira (2009).  $L_{dev}$  corresponds to the period during which leaf  
334 area is growing at a fast rate, linearly with thermal time (the *grand growth*  
335 *period*).  $L_{mid}$  corresponds to the period during which leaf area does not grow  
336 anymore (because of natural shoot growth cessation or due to mechanical  
337 hedging cutting away the growing points).

338 To determine  $K_{CB}(t)$ , two hypotheses on the curve shape are assumed:

$$K_{CB}(t) = f(t) \text{ for } t < t_{K^*} \quad (4)$$

$$K_{CB}(t) = K^* \text{ for } t \geq t_{K^*} \quad (5)$$

339 where  $f(t)$  is assumed to be linear in  $t$ , and  $t_{K^*}$  is the breakpoint for which  
340  $K_{CB}$  reaches the plateau  $K^*$ . The key point is to set  $t_{K^*}$ , or indifferently  $K^*$ .  
341 According to Eq.3, we make the hypothesis that, in the absence of water  
342 deficit,  $K^*$  is defined as:

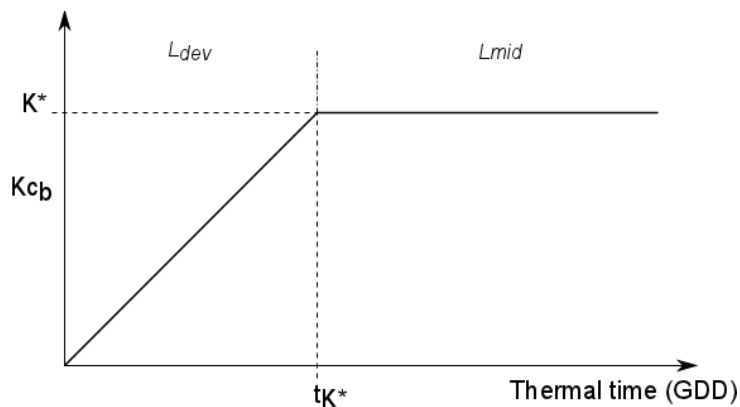


Figure 5: Theoretical curve of  $K_{CB}$  evolution during the season.

$$K^* = \frac{T(t_{K^*})}{ET_{ref}(t_{K^*})} \quad (6)$$

343 Using the OVWS ontology defined in Section 2.2, the following rules are  
 344 set up to automatically define a limited number of potential options for  $t_{K^*}$ .  
 345 The interest of having the rules and concepts defined in an ontology is two-  
 346 fold: *i*) they have to be completely explicit, *ii*) they can evolve independently  
 347 of the numerical procedures.

348 1. *Selection based on phenology*

349 A linear relationship exists between  $K_{CB}$  variations and leaf area index  
 350 (LAI) or the fraction of ground covered by the vine, see Ferreira et al.  
 351 (2012). We thus assume that peak  $K_{CB}$  (i.e.  $K^*$ ) is reached when  
 352 LAI stops increasing. Consequently, the search period for  $K^*$  has been  
 353 limited to the period between budbreak and veraison.

354 These two concepts are defined as sub-concepts of the Phenology con-  
 355 cept, itself being a sub-concept of  $MVariable$ . The period limitation is  
 356 instantiated by two *TimeConditions*, applied onto the  $K_{CB} \preceq CVariable$

357 concept, the  $\mathcal{HCO}_c$  relation where the *Condition* is characterized by a  
358 comparison operator  $\leq$  (resp.  $\geq$ ) and the *Veraison* (resp. *Budbreak*)  
359 concept.

360 2. *Selection based on predawn leaf water potential*

361 Conditions of maximal soil moisture availability could be inferred from  
362 predawn leaf water potential measurements, associated with a confi-  
363 dence interval derived from VPD. A rule was set so that  $K^*$  has to be  
364 reached before the first day at which predawn LWP measurement re-  
365 veals a water deficit level limiting shoot elongation. The levels to which  
366 predawn LWP characterizes that limiting effect can be defined by the  
367 stakeholder, or else set in agreement with a standard level, based on a  
368 region or/and variety.

369 This is implemented in the ontology by the *WorldregionVariety* and  
370 *level  $\preceq$ Variable* concepts.

371 3. *Selection based on meteorology*

372 Transpiration measurement through sap flow is sensitive to climatic  
373 conditions, mainly light and *VPD*. To account for sensitivity of transpi-  
374 ration measurements to  $VPD(t)$ , a filtering rule was set to remove com-  
375 puted  $K_{CB}(t)$  obtained in situations of heat spikes, defined as period  
376 with *VPD* greater than a given level, set to 3.5 k.Pa in the present case  
377 study.

378 The rule is implemented using a  $\mathcal{HCS}_c$  relation, applied onto the  
379  $K_{CB} \preceq CVariable$  concept, where the *Constraint* is characterized by a  
380 comparison operator  $\leq$  and the  $VPD \preceq Meteorology \preceq MVariable$  con-  
381 cept.

382 4. *Selection based on curve shape*

383 By definition,  $K^*$  is reached when the ratio  $\frac{T(t)}{ET_{ref}(t)}$  reaches a maxi-  
384 mum during a few days (as  $T(t)=T_{max}(t)$  and  $K_{cB}(t)=K^*$ ) and then  
385 decreases (as  $T(t)<T_{max}(t)$  due to limiting soil water conditions while  
386  $K_{cB}(t)=K^*$ ). As such, potential options for  $t_{K^*}$  have been defined at  
387 points with a null first derivative and a negative second derivative. This  
388 selection is implemented using two concepts:  $\dot{K}^*$  and  $\ddot{K}^*$ , both such as  
389  $\preceq WaterStress \preceq CVariable$ , and a  $\mathcal{HCS}_c$  relation with a *Constraint*  
390 characterized by a comparison operator  $\leq \epsilon$  or  $\leq 0$ .

391 *User selection based on expert knowledge.*

392 The analysis of  $\frac{T(t)}{ET_{ref}(t)}$  curve shape, associated with all previous rules  
393 based on phenology, meteorology and predawn leaf water potential leads to  
394 the proposal of a small finite set of  $t_{K^*}$  candidates.

395 The final choice is left to the stakeholder who is the best aware of the  
396 management practices or particular uncontrolled events that could have in-  
397 terfered with vine growth (irrigation, leaf removal, trellis system...) and  
398 therefore with the  $K_{cB}$  curve.

399 2.4. *Relating software sensor output to product quality*

400 The software sensor output consists of temporal data, and various meth-  
401 ods can be used to study the relationships between these data and product  
402 quality. Two complementary lines of work, based on statistical methods, are  
403 explored in the present work. The first one consists of extracting significant  
404 scalar parameters from the temporal data and using them as input to deci-  
405 sion trees, in order to provide the most discriminant features. The second

406 one uses functional data analysis, that gives the possibility to model the tem-  
407 poral data impact on product quality as a whole. However, curve analysis is  
408 a recent research topic, with relatively few methods available, in comparison  
409 with classical data analysis.

410 This section is divided into three parts. The first part describes how  
411 to use the formalized knowledge for extracting significant scalar parameters  
412 from the temporal data. The other two parts give some elements necessary  
413 to understand the statistical methods that will be used: decision trees and  
414 functional data analysis.

#### 415 *2.4.1. Extracting scalar parameters from software sensor output*

416 Meaningful scalar parameters can be extracted from temporal courses  
417 determined by the software sensor outputs. In many cases, expert knowledge  
418 can be the support of such extraction procedures. In the case of vine water  
419 courses, this can be achieved by taking into account important phenological  
420 periods, which are defined as concepts in the ontology (see Fig.2).

421 Three periods were first defined according to phenological stages: the  
422 whole season, the pre-veraison period, which goes from the nouaison stage  
423 to the veraison stage, and the post-veraison period which ranges from the  
424 veraison stage to the harvest date. In a second step, the post-veraison period  
425 was divided by taking into account the maturity stage of berries, which al-  
426 lowed to add a fourth period ranging from veraison to maturity. Maturity  
427 stage is reached when the ratio between Sugar Concentration and acidity in  
428 grapes yields a given threshold, defined according to variety.

429 Using trapeze integration under  $Ks$  curves over these four periods, the  
430 continuous  $Ks(t)$  curve was summarized into four new variables corresponding



431 to the cumulative amount of stress encountered by the vine over these peri-  
432 ods: *NouHarv*, *NouVer*, *VerHarv* and *VerMat*. Table 1 gives the summary  
433 of these four aggregated variables for each plot. Since all these aggregated  
434 variables are based on the area under the curve, the lower their value, the  
435 stronger the water stress over the considered period.

#### 436 2.4.2. Decision trees as interpretable models

437 Decision tree algorithms are well established learning methods in super-  
438 vised data mining and statistical multivariate analysis. They allow to display  
439 non linear relationships between features and their impact on a response vari-  
440 able, in a compact way.

441 Decision trees can handle classification problems or regression cases, de-  
442 pending on the nature of the response variable. Note that the CART family,  
443 see Breiman et al. (1984), based on binary splits, is mostly used by statisti-  
444 cians. There is another tree family, called ID3, see Quinlan (1986, 1993), al-  
445 lowing non binary splits and mostly used by artificial intelligence researchers.

446 We recall here the principle of the regression case, where the response  
447 variable is numerical.

448 Input to regression decision trees consists of a collection of  $N$  train-  
449 ing cases, each having a tuple of values for a set of  $P$  input variables,  
450 and one continuous output variable  $(\mathbf{x}_i, \mathbf{y}_i) = (x_{1,i}, x_{2,i} \dots x_{P,i}, y_i)$ . An in-  
451 put  $X_p$  ( $p = 1 \dots P$ ) is continuous or discrete and takes its values  $(x_{p,i})_{i=1 \dots N}$   
452 on a domain  $\mathcal{X}_p$ . The goal is to learn from the training cases a recursive  
453 structure (taking the shape of a rooted tree) consisting of (i) leaf nodes la-  
454 beled with a mean value and a standard deviation, and (ii) test nodes (each  
455 one associated to a given variable) that can have two or more outcomes, each

Site	Variety	Irrigation	NouHarv	NouVer	VerHarv	VerMat
LB-CS	Cabernet-S.	$i_0$	1165.6	704.6	432.6	432.6
		$i_1$	1117.8	713.8	375.0	347.6
OUV-Mer	Merlot	$i_0$	742.1	457.2	278.3	163.9
		$i_1$	1233.8	814.7	403.4	247.9
StGER-Mer	Merlot	$i_0$	608.9	470.9	131.6	129.4
		$i_1$	808.3	473.4	327.0	214.5
PR-Mer	Merlot	$i_0$	655.4	381.8	250.2	199.5
		$i_1$	722.7	398.3	313.4	266.2
StSAU-Char	Chardonnay	$i_0$	695.3	442.6	241.9	213.6
		$i_1$	693.8	414.0	265.3	253.7
RIE-Gre-Chm	Grenache	$i_0$	651.1	465.2	169.9	169.9
		$i_1$	620.0	390.9	209.5	209.5
RIE-Gre-Chp	Grenache	$i_0$	512.3	380.0	123.0	NA
		$i_1$	963.9	580.2	362.0	362.0
PIO-Gre	Grenache	$i_0$	677.2	458.4	212.0	152.3
		$i_1$	887.4	514.4	363.8	180.8

Table 1: Values of aggregated variables for each site-variety-irrigation treatment combination (i) over the entire season *NouHarv*, (ii) before veraison *NouVer*, (iii) after veraison *VerHarv* and (iv) from veraison to maturity *VerMat*.

456 of these linked to a sub-tree.

457 On a given node, the algorithm examines in turn all available variables,  
458 and selects the variable that most effectively splits the set of samples into  
459 subsets improving the separation between output values. Once (and if) a  
460 variable is selected, a new test node is created that splits on this variable,  
461 and the procedure is recursively applied on each (new) node child. At each  
462 node, the algorithm stops when no more variables are available, or if there is  
463 no improvement by splitting further: the node then becomes a leaf.

464 Decision trees are easily interpretable for a non-expert in statistical or  
465 learning methods, and facilitate exchanges with the domain expert. A low  
466 complexity, see Ben-David and Sterling (2006), is essential for the model  
467 to be interpretable, as confirmed by Miller's conclusions, see Miller (1956),  
468 relative to the *magical number* seven.

469 Well-known drawbacks of decision trees are the sensitivity to outliers and  
470 the risk of overfitting. To avoid overfitting, cross-validation is included in the  
471 procedure and to gain in robustness, a pruning step usually follows the tree  
472 growing step, see Quinlan (1986); Breiman et al. (1984); Quinlan (1993).

473 In this work, we used CART-based trees. In that case, the splitting  
474 criterion is based on finding the one predictor variable (and a given threshold  
475 of that variable) that results in the greatest change in explained deviance (for  
476 Gaussian error, this is equivalent to maximizing the between-group sum of  
477 squares, as in an ANOVA). This is done using an exhaustive search of all  
478 possible threshold values for each predictor. The implementation used for  
479 decision trees is the R software, described in R Development Core Team

480 (2009), with the *rpart* package<sup>3</sup>.

481 Specifying variety, *NouVer*, *VerHarv* and *VerMat* as explanatory vari-  
482 ables, we performed decision trees on maximum values of grape quality fea-  
483 tures over the season.

#### 484 2.4.3. Functional data analysis

485 Functional linear regression is an approach to model the relationship  
486 between a scalar dependent variable  $Y$  and a functional predictor  $X(t)$ , a  
487 function of a real variable  $t$  (time for example). The model is written as

$$Y_i = \beta_0 + \int X_i(t)\beta(t)dt + \varepsilon_i, \quad i = 1, \dots, n \quad (7)$$

488 where  $\varepsilon_i$  is a random error,  $\beta_0$  is the intercept of the model and  $\beta(t)$  is the  
489 coefficient function, both unknown and to be estimated from independent  
490 observations  $(X_i(t), Y_i)_{i=1, \dots, n}$ . In this model,  $\beta(t)$  determines the effect of  
491  $X_i(t)$  on  $Y_i$ . For example,  $X_i(t)$  has a greater effect on  $Y_i$  over regions of  $t$   
492 where  $|\beta(t)|$  is large. On the opposite,  $X_i(t)$  has no effect on  $Y_i$  over regions  
493 of  $t$  where  $\beta(t)$  is zero. Estimating  $\beta(t)$  in Eq.7 has given rise to an increasing  
494 literature in the last decade, see for example Ramsay and Silverman (2005).  
495 A common approach involves projecting  $\beta$  and the  $X_i$ 's in a  $p$ -dimensional  
496 basis function where  $p$  is large enough to capture the unknown variations of  
497  $\beta$ , but small enough to regularize the fit. Such techniques are not sufficient  
498 to produce estimates of  $\beta(t)$  that are exactly zero in the regions of  $t$  where  
499  $X_i(t)$  has no effect on  $Y_i$ .

500 Recently, James et al. (2009) introduced new estimators that are both

---

<sup>3</sup><http://cran.r-project.org/web/packages/rpart/index.html>

501 interpretable, flexible and accurate. The method, called “Functional Linear  
502 Regression That’s Interpretable” (FLRTI), is based on a particular basis  
503 function and variable selection techniques. The time-period is divided into  
504 a fine grid of points  $(t_j)_{j=1,\dots,p}$ . The  $\beta$  function is assumed to be exactly zero  
505 over some time periods and exactly linear over the remaining periods, period  
506 location being unknown. The reason behind the first assumption is that all  
507 the  $X_i(t)$  observations, for varying  $t$ , are not of equal importance to explain  
508 the response  $Y_i$  since  $X_i(t_j)$  has no effect on  $Y_i$  when  $\beta(t_j) = 0$ . Hence the  
509  $\beta$  function is assumed to be sparse. The second assumption is made for  
510 obtaining an easily interpretable  $\beta$  function, it is implicitly equivalent to the  
511 assumption that the second derivative of  $\beta(t)$  will be zero over these regions  
512 of  $t$ , that is, the second derivative is assumed to be sparse.

513 These assumptions will constraint the estimation of the regression model  
514 (Eq.7), which corresponds to a penalized regression in sparse models, with a  
515 number of time grid points  $p$  much larger than the number of observations  $n$ .  
516 To estimate the  $\beta$  function at each point  $t_j$ , it is necessary to minimize the  
517 mean squared error criterion subject to a regularity constraint. The Dantzig  
518 selector is the solution to this problem used by the authors of FLRTI. The  
519 whole method is implemented in an R function available on J. Gareth’s web  
520 page<sup>4</sup>. Finally, two tuning parameters have to be fixed, a penalty term  $\sigma$  and  
521 a weight  $\omega$ . The penalty term is part of the Dantzig selector procedure. The  
522 largest the  $\sigma$ , the more the form-related constraint is enforced. The weight  
523  $\omega$  impacts the relative number of zeros of the  $\beta$  function. A weight equal to

---

<sup>4</sup><http://www-bcf.usc.edu/gareth/research/flrti>

524 0 indicates that only the linear form constraint is respected, no assumption  
525 is made on the sparsity of  $\beta$ .

526 A cross-validation algorithm is also proposed to optimize the choice of  
527  $\sigma$  and  $\omega$ . The cross-validation procedure aims at estimating optimal values  
528 for  $\sigma$  and  $\omega$ , from two sets of possible values for them  $(\sigma_k)_k$  and  $(\omega_l)_l$ . The  
529 principle is to divide the data set into  $N_f$  folds (typically 10). All folds but  
530 one are used to train the estimation process with each combination of  $(\sigma_k, \omega_l)$ .  
531 The excluded fold is used to test the estimated model, yielding an error for  
532 each  $(\sigma_k, \omega_l)$ . This is repeated until all folds have been used once for testing.  
533 At the end, we obtain  $N_f$  errors for each combination  $(\sigma_k, \omega_l)$ , whose mean  
534 yields a cross-validated error for each  $(\sigma_k, \omega_l)$ . The optimal choice of  $\sigma$  and  
535  $\omega$  is the couple with the smallest cross-validated error.

### 536 3. Results

537 In this section, we first present the results of the  $Ks(t)$  estimation using  
538 the software sensor. In a second step, we study the relationship between  
539  $Ks(t)$  and grape quality features, using the methods described in Section 2.  
540 In the following, we will refer to irrigated treatments with  $i_1$ , and to non  
541 irrigated ones with  $i_0$ .

#### 542 3.1. Vine water stress course $Ks(t)$ estimation

543 Sap flow data require a pre-treatment, including sensor selection and  
544 signal smoothing. Sap flow sensors have only been used recently in European  
545 vineyards. Thus, calibration protocols are not established yet and therefore  
546 sensors can still be unreliable. Consequently, a selection step is required.

547 Sensor reliability has been assessed on the basis of the number of incorrect  
548 hourly measurements resulting from expert filtering methods. A sensor was  
549 considered reliable when less than 5% of the hourly data were filtered out. For  
550 each variety-irrigation combination, the mean daily vine transpiration was  
551 calculated as the mean of daily measures from reliable sensors, which helped  
552 limiting the variability in plant transpiration measurements. However, one  
553 of the major drawback of sensor selection was the potential lack of reliable  
554 measurements on a daily basis.

555 To capture important patterns in daily sap flow data, while leaving out  
556 noise and extreme variations (daily peak), sap flow courses were smoothed  
557 with the central moving average method with a five day window. This  
558 smoothing allowed the removal of missing values and extreme peaks.

### 559 3.1.1. $K^*$ determination

560 Regarding all site-variety combinations, the knowledge-based algorithm  
561 for  $t_{K^*}$  determination proposed from 5 to 9 candidates (resp. from 4 to 8  
562 candidates) in the non-irrigated  $i_0$  (resp. irrigated  $i_1$ ) treatments. Most of  
563 the dates proposed by the mathematical algorithm were in accordance with  
564 expert knowledge, so allowing the expert to choose  $t_{K^*}$  within the algorithm  
565 suggestions (Fig.6). The results are given in Table 2. Fig.6 illustrates the  
566 results for the Grenache variety at the Piolenc site.

567 The validity of the  $K^*$  determination procedure can be assessed according  
568 to different points. First, the results regarding  $K^*$  determination based on the  
569 coupling of mathematical algorithms and expert knowledge were consistent  
570 with existing literature. Indeed, most of  $t_{K^*}$  occurred between 600 and 700  
571 GDD after budbreak (Table 2), which is in accordance with  $t_{K^*}$  reported in

Site	Variety	Irrigation	$K^*$	$t_{K^*}$ (GDD)	First irrigation(GDD)
La Baume	CS	i0	20.3	677.9	
		i1	32	698.6	1268.3
Pech Rouge	Merlot	i0	19.4	614.5	
		i1	26,6	614.5	610.4
St Gervasy	Merlot	i0	69.3	625.3	
		i1	85.1	669.4	844
Ouveillan	Merlot	i0	37.1	829.2	
		i1	21	1005.1	939.7
Piolenc	Grenache	i0	44.3	530.1	
		i1	58.1	530.1	864.4
Rieux	Grenache+	i0	43.4	600.5	
		i1	29.1	594.5	789.5
Rieux	Grenache-	i0	29.2	580	
		i1	46.1	580	789.5
St Sauveur	Chardonnay	i0	43.3	642	
		i1	54.4	749.7	777.2

Table 2: Values of basal crop coefficients  $K^*$  and dates ( $t_{K^*}$  in GDD) at which they were estimated in the site-variety-irrigation treatment combinations during season 2012.



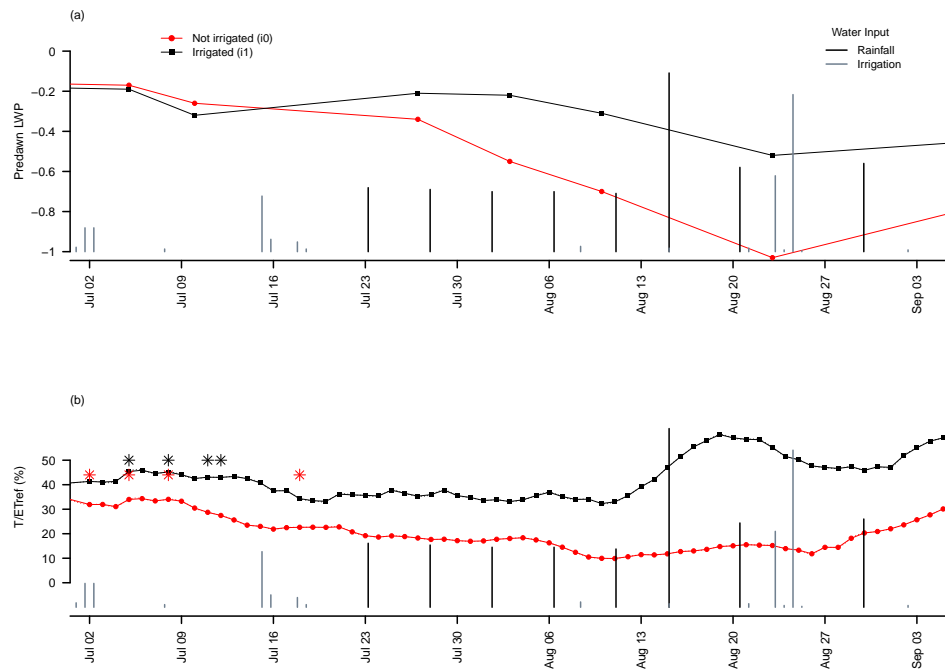


Figure 6: (a) Vine water status as indicated by leaf water potential and (b) determination of  $K^*$  and  $t_{K^*}$  (stars) based on expert knowledge following the mathematical algorithm suggestions for non irrigated (red bullets) and irrigated (black squares) treatments on the Grenache variety at Piolenc site.

572 Picón-Toro et al. (2012) from a 3 year study in western Spain on Tempranillo,  
 573 and in FAO-56, see Allen et al. (1998), that respectively reported  $t_{K^*}$  around  
 574 650 GDD and 555-592 GDD after budbreak.

575 *3.1.2. Maximal transpiration and  $K_s$  estimation*

576 Following determination of  $K^*$  and  $t_{K^*}$ ,  $Kc_B(t)$  was calculated over all  $t$   
 577 values (Eq.4). Its variation for a Grenache variety is plotted on Fig.7, both  
 578 in calendar time (a) and thermal time since budbreak (b).

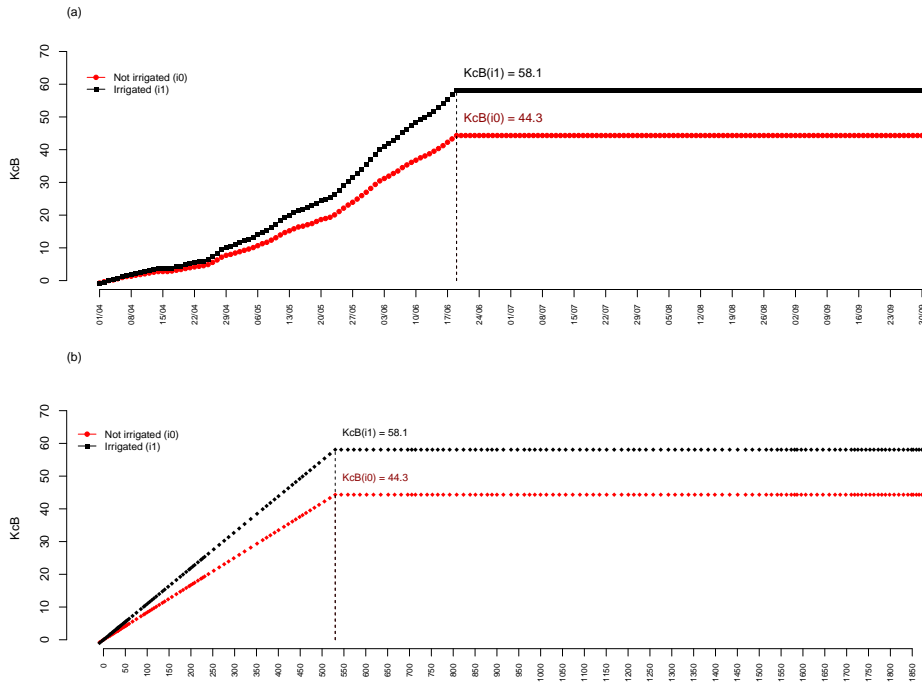


Figure 7: Evolution of vine basal crop coefficient ( $Kc_B$ ) during the season at Piolenc site with Grenache variety, from budbreak to harvest. (a) x-scale in Julian days - (b) x-scale in GDD since budbreak.

579  $Kc_B(t)$  was then used to calculate the daily vine maximal transpiration  
580 ( $T_{max}$ ), according to Eq.3. Finally,  $Ks(t)$  was calculated as the daily ratio  
581 of measured transpiration by reliable sensors over potential transpiration  
582 (Eq.2).

583 Figure 8 shows vine water status according to both indicators: (a) Predawn  
584 LWP and (b)  $Ks$  during the season 2012 in a Grenache variety of the Languedoc-  
585 Roussillon region.

### 586 3.2. Relationships between vine water stress $Ks(t)$ and grape quality

587 As explained in Section 2.4,  $Ks(t)$  can be used in two different ways,  
588 either summarized as a series of scalar values, or as a whole. The way to  
589 summarize  $Ks(t)$  is detailed in Section 2.4.1. Scalar values and  $Ks(t)$  will be  
590 put in relation to grape quality at harvest time, by the respective use of (i)  
591 regression trees and (ii) functional data analysis. The studied grape quality  
592 features include *i*) Berry Weight and *ii*) Sugar Concentration in berries. For  
593 interpreting the results, note that  $Ks(t)$  is inversely related to water deficit.

#### 594 3.2.1. Regression trees

595 Aggregated variables over periods can be used as explanatory variables  
596 in regression trees to detect and prioritize the periods critical to changes in  
597 grape quality. We studied the effects of *NouVer*, *VerHarv*, *VerMat* and vari-  
598 ety on the two components of grape quality cited above. The corresponding  
599 regression trees are displayed in Fig.9 and Fig.10, together with the distri-  
600 bution of values at terminal nodes, represented by boxplots. Table 3 shows  
601 the gaining in deviance for each splitting step during the tree generation.  
602 The number of available samples being small (16), the minimum number of

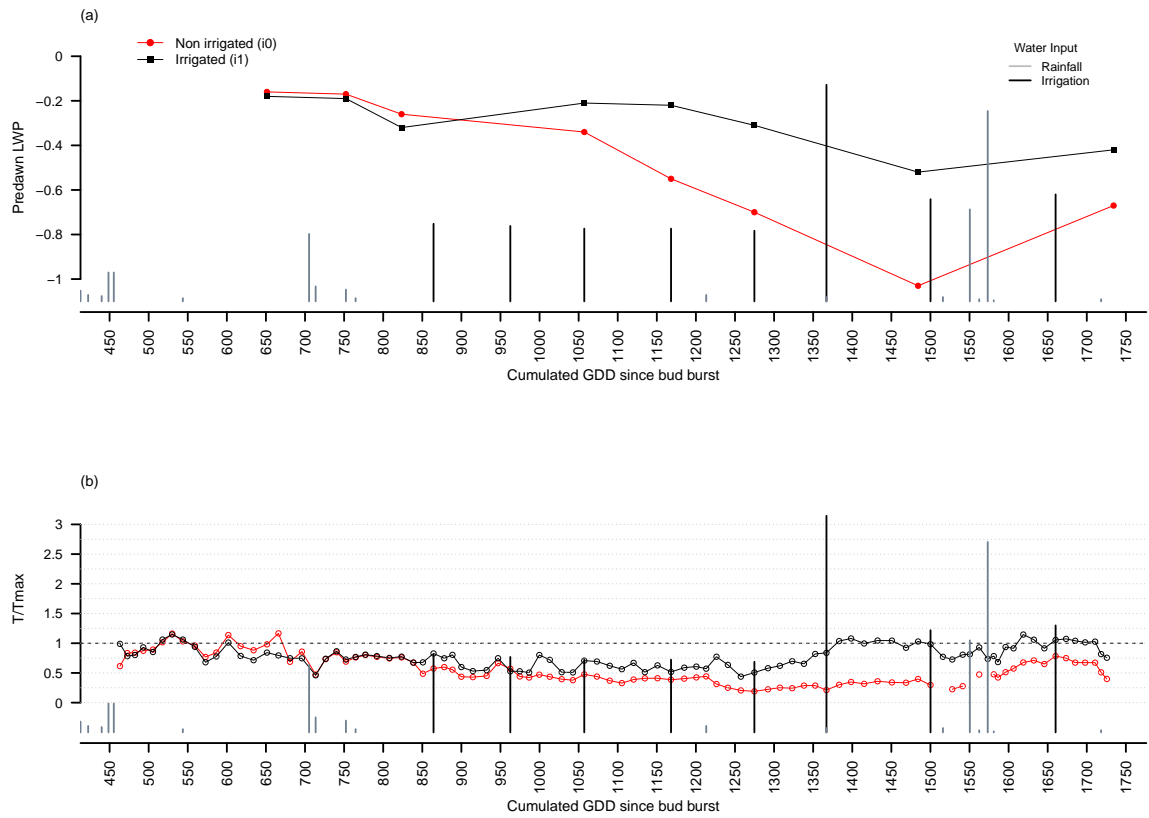


Figure 8: Water deficit during 2012 millesim in Grenache variety at Piolenc site assessed by (a) Predawn LWP and (b) vine water stress indicator  $K_s$ .

603 observations in any terminal node was set to 1. That is not sufficient to  
 604 support prediction with a good confidence level, but is still interesting for  
 605 summarizing the data.

Regression tree	split 1	split 2	split 3	split 4
Berry weight	0.64	0.63	0.69	
Sugar concentration	0.44	0.39	0.87	0.33

Table 3: Gain in deviance during regression tree generation.

606 According to Fig.9, Berry Weight seems to be mostly affected by the  
 607 variety (Fig.9). Grenache variety significantly yields heavier berries. The  
 608 second split for all varieties is done on the post-veraison water stress only  
 609 (either *VerHarv* or *VerMat*). The more severe is water deficit *post veraison*,  
 610 the smaller is the Berry Weight.

611 Regarding Sugar Concentration, regression trees show that it is affected  
 612 by water stress in both pre-veraison *NouVer* and post-veraison *VerHarv* pe-  
 613 riods (Fig.10). The first discriminant variable on Sugar Concentration is the  
 614 post-veraison water stress (*VerHarv*, Fig.10). The first split shows that a  
 615 higher post-veraison water stress leads to a lower Sugar Concentration.

616 The left branch resulting from the first split shows that the next dis-  
 617 criminant variable is again the post-veraison stress *VerHarv*, so enhancing  
 618 the effect of the previous split. Lastly, pre-veraison water stress (*NouVer*)  
 619 can exacerbate the decrease in Sugar Concentration (as shown at the tree  
 620 bottom).

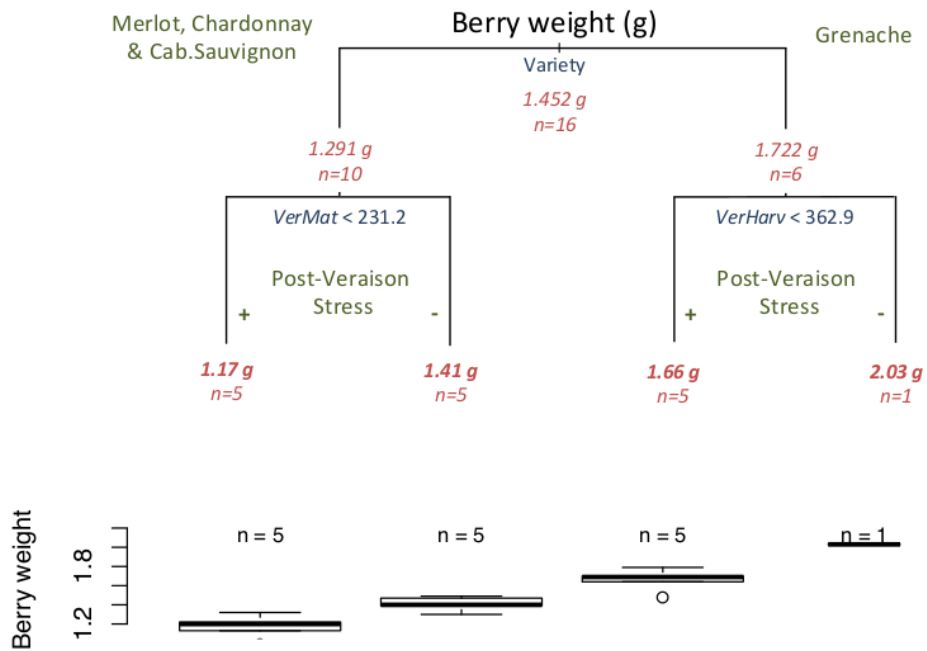


Figure 9: Regression tree explaining Berry Weight (g) using scalars summarizing the three periods, i.e. pre-veraison (*NouVer*), and post-veraison either until maturity (*VerMat*) or harvest (*VerHarv*). Boxplots showing the distribution of values at terminal nodes are displayed below the tree.

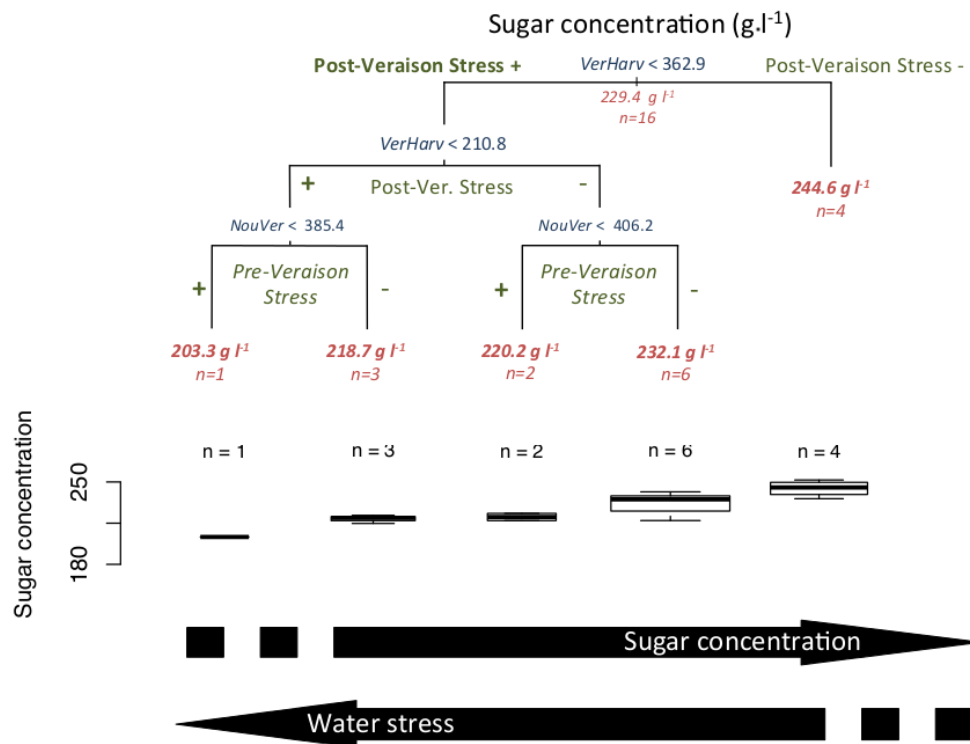


Figure 10: Regression tree explaining Sugar Concentration in berries ( $\text{g}\cdot\text{l}^{-1}$ ) using scalars summarizing the three periods, i.e. pre-veraison (*NouVer*), and post-veraison either until maturity (*VerMat*) or harvest (*VerHarv*). Boxplots showing the distribution of values at terminal nodes are displayed below the tree.

621 *3.2.2. Functional data analysis*

622 Using a continuous indicator of water deficit enables the use of the whole  
623 season water deficit curve to explain berry composition. This in turn is likely  
624 to promote a more precise monitoring of vine water needs according to the  
625 targeted fruit composition. Using the FLRTI method, described in James  
626 et al. (2009), we analyzed the effects of the vine water deficit over the season  
627 on Berry Weight and Sugar Concentration in berry at harvest.

628 Regarding Berry Weight, the results showed no significant effect of  $Ks(t)$ .  
629 This was confirmed by applying a testing procedure designed to test the  
630 nullity of the  $\beta$  function in a generic functional linear model with scalar  
631 output like the one given in Eq.7. The literature on such tests is scarce. We  
632 applied the one introduced in Hilgert et al. (2013), which has the particularity  
633 of not requiring any prior knowledge on the  $\beta$  function. A p-value of 0.7 of  
634 the procedure was estimated by Monte-Carlo simulations (with 10 000 runs).  
635 The fact that  $Ks(t)$  has no significant effect on Berry Weight might be due  
636 to the non taking into account of the variety effect in the model, which is  
637 very important to explain Berry Weight. On top of that, it may be possible  
638 that the level of water deficit is not severe enough to induce changes in Berry  
639 Weight or that the timing of water deficit happens too late in the season to  
640 have an effect at limiting berry size, see Ojeda et al. (2002). Decision trees  
641 highlighted an effect of post veraison water deficit on Berry Weight, but as  
642 a minor effect compared to the variety influence.

643 The functional data analysis on the effect of  $Ks(t)$  on Sugar Concentra-  
644 tion yields an estimation of the  $\beta(t)$  coefficient function, that is displayed in  
645 Fig.11.  $\beta_0$ , the intercept in Eq.7, is estimated at 178.3 g.l<sup>-1</sup>. The tuning pa-



646 rameters are indicated in the legend. The goodness-of-fit of the estimated  $\beta$   
647 curve is measured by a  $R^2$  value, equal to 0.7. Since this coefficient measures  
648 the percentage of variation of the data explained by the fitted model, a  $R^2$   
649 equal to 0.7 is a rather high value in the context of penalized regression. A  
650 p-value of 0.02 of the testing procedure was estimated, in the same way than  
651 for Berry Weight. Residuals, plotted in Fig.12, showed a good repartition  
652 when plotted against predicted values, and no tendency. So the  $Ks(t)$  curve  
653 appears to be a relevant variable to explain the Sugar Concentration. Let us  
654 also note that parameters were obtained following a ten-fold cross validation.  
655 A sensitivity analysis to small  $\sigma$  and  $\omega$  variations showed a good robustness  
656 of the model, with three main peaks always located in the same time periods  
657 across the different varieties. These three main peaks are labeled (1), (2)  
658 and (3). Each of them corresponds to a significant effect of  $Ks$  on Sugar  
659 Concentration, which can be positive (peaks (1) and (3)) or negative (peak  
660 (2)).

661 Peaks (1) and (3) are positive, which implies a rise in Sugar Concentra-  
662 tion. During these periods, the stronger the  $Ks$  value, the higher the rise.  
663 As  $Ks$  varies inversely with water deficit, it means that the lower the water  
664 deficit during these periods, the higher the rise in Sugar Concentration. The  
665 effect is twice as strong for the peak (1) than for the peak (3).

666 Regarding the time period, peak (1) appears to be located before pre-  
667 veraison whereas peak (3) occurred during pre-veraison. By contrast, peak  
668 (2) has a negative effect on Sugar Concentration. During this period located  
669 within the grand growth phase, a low water deficit decreases the Sugar Con-  
670 centration. This can be reformulated as follows: the higher the deficit during

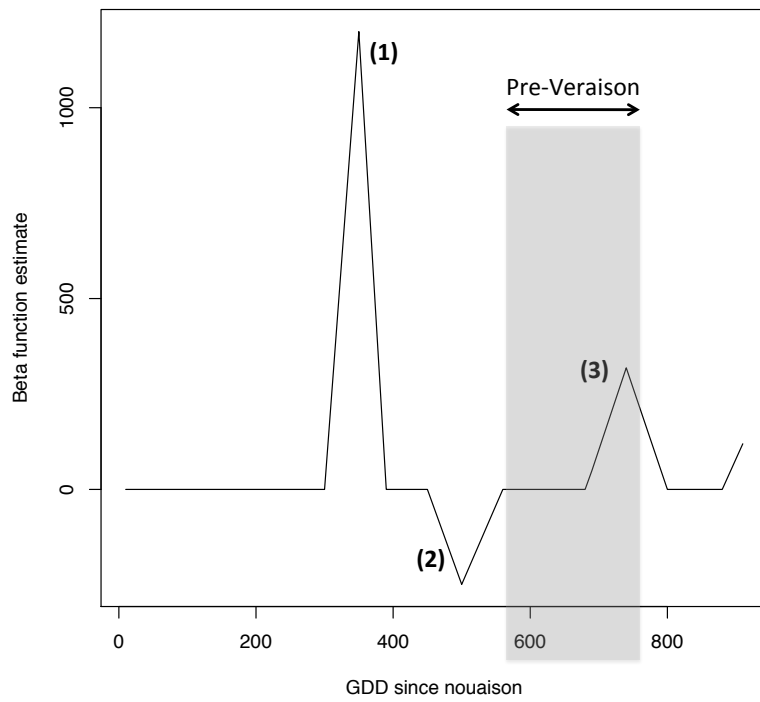


Figure 11: Beta function evolution over time (see Eq. 7), for explaining Sugar Concentration at harvest. Abscissae are in GDD. The values of  $\sigma = 0.05$  and  $\omega = 0.95$  have been found by cross-validation.

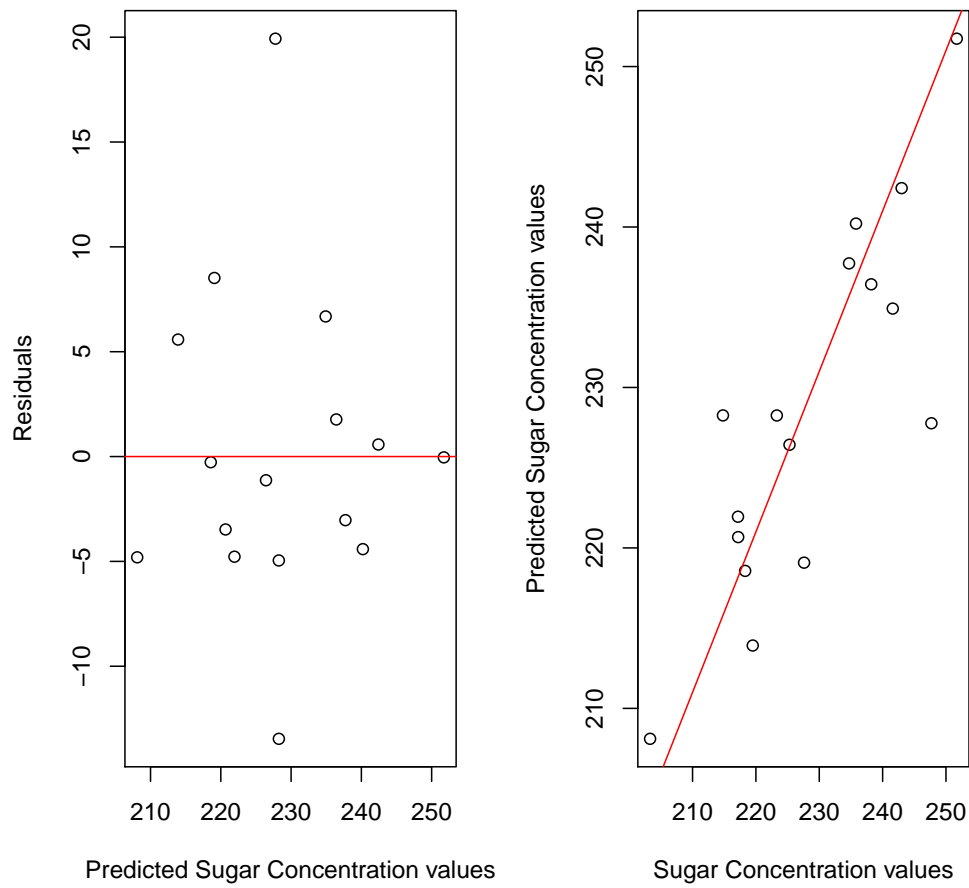


Figure 12: Plots of the residuals and the predicted Sugar Concentration values associated with the Beta function estimation. Abscissae are in  $g.l^{-1}$ .

671 that period of grand growth (i.e. before  $K^*$  is reached), the lower the Sugar  
672 Concentration. These results are consistent with the ones obtained by using  
673 decision trees (section 3.2.1, Fig.10), but more informative regarding the time  
674 period of interest.

#### 675 4. Conclusion

676 The work presented in this paper used formalized knowledge and math-  
677 ematical models to design a software sensor from raw data and relate its  
678 temporal output to product quality. The proposed approach has been ap-  
679 plied to the case of a vine water deficit indicator, and its relation to two  
680 grape quality variables: Berry Weight and Sugar Concentration.

681 Results provide a number of meaningful insights.

682 First of all, the software sensor key point, which is the determination of  
683  $K^*$ , seems reasonably consistent with the literature. From an agronomical  
684 point of view, this allows to effectively work at plot scale, and to offer decision  
685 support for irrigation, as a function of each plot characteristics.

686 The use of an ontology allows to separate expert knowledge and numerical  
687 models. It makes it much easier to build a generic model, that is both evo-  
688 lutive and adaptable over time as knowledge progresses or climate changes.

689 Contrary to a data base, an ontology schema adds semantics to the data  
690 structure, allowing automatic reasoning, using logical properties, such as  
691 reflexivity or transitivity.

692 The ontology presented here has a moderate complexity level: only four  
693 kinds of primary concepts, and five types of relations. This is still sufficient to  
694 express many mathematical conditions and dependencies, going well beyond

695 the scope of the present case study. There may however be cases where new  
696 concepts and relations are necessary, and the ontology can easily be enriched  
697 when needed.

698 Second, the two-fold proposal for data analysis appears to be a good  
699 means of exploiting such temporal data as provided by the water deficit in-  
700 dicator  $Ks(t)$ . The results show that the water deficit has an effect on grape  
701 quality. Their analysis confirmed already known facts about the vine phys-  
702 iological response according to the variety and the irrigation effect. Thus  
703 our results are comforting the validity of the  $Ks$  indicator, and therefore  
704 the level of confidence and reliability in the software sensor design proce-  
705 dure. Functional data analysis highlighted critical periods for vine and berry  
706 development, regarding final quality features.

707 On one hand, the knowledge-based extraction of meaningful summary  
708 features over phenological periods of interest allowed to feed these features  
709 as input to decision trees. This confirmed the primordial effect played by the  
710 variety on Berry Weight determination. On the other hand, functional data  
711 analysis made it possible to use the water stress curve ( $Ks(t)$ ), as a whole,  
712 to explain Sugar Concentration. This will in the future allow more precise  
713 monitoring of vine water needs according to a targeted product.

714 Note that we did not take account of the variety factor in functional  
715 data analysis. This would require a covariance analysis model adapted to  
716 functional data, which was not possible in this study as the number of data  
717 per variety was not sufficient.

718 These results show the complementarity of both approaches: the first  
719 one performs dimensional reduction by summarizing features which requires

720 expert assumptions, the second one handles the continuous temporal data,  
721 without any reduction, but it needs more numerous data to be efficient.

722 Applied perspectives of this work include the study of the relationship  
723 between vine water stress and other more complex quality features. In par-  
724 ticular new chemical analyses make it possible to follow the aroma develop-  
725 ment in berries over time, which is assumed to be very sensitive to the vine  
726 water status.

727 Our approach is innovative in more than one aspect. Even if the software  
728 sensor had a different design, the same advanced methodology could still  
729 be applied to analyze the temporal data. Beyond the present case study,  
730 the proposed methodology has a high genericity level, for the applied fields  
731 of Agronomy and Environment. It could be used in many cases when raw  
732 data have to be transformed by software sensors to be meaningful, or when  
733 temporal data have to be analyzed in depth.

## 734 **Acknowledgments**

735 The research leading to these results has received funding from the Pilotype  
736 Program, funded by OSEO innovation and the Languedoc Roussillon regional  
737 council. The authors would like to thank all members of the project for  
738 their help and advices: Les Grands Chais de France, Alliance Minervois, Les  
739 Vignerons du Narbonnais, INRA Pech Rouge, INRA SPO, INRA MISTEA,  
740 IFV Rhône Méditerranée, SupAgro Montpellier (ITAP), Fruition Sciences,  
741 Nyseos. Finally, we wish to particularly thank Nicolas Saurin (INRA), Denis  
742 Caboulet, Jean-Christophe Payan and Elian Salançon (IFV) for providing  
743 the vine and wine-related data.

744 **References**

- 745 Allen, R. G., Pereira, L. S., Sep. 2009. Estimating crop coefficients from  
746 fraction of ground cover and height. *Irrigation Science* 28 (1), 17–34.  
747 URL <http://link.springer.com/10.1007/s00271-009-0182-z>
- 748 Allen, R. G., Pereira, L. S., Raes, D., Smith, M., 1998. Crop evapotranspi-  
749 ration: Guidelines for computing crop water requirements. *Irrigation and*  
750 *Drainage Paper No. 56*. FAO, Rome, Italy.
- 751 Ben-David, A., Sterling, L., 2006. Generating rules from examples of human  
752 multiattribute decision making should be simple. *Expert Syst. Appl.* 31 (2),  
753 390–396.
- 754 Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. *Classification and*  
755 *regression trees*. Wadsworth, Belmont, CA 1.
- 756 Cifre, J., Bota, J., Escalona, J., Medrano, H., Flexas, J., Apr. 2005. Physio-  
757 logical tools for irrigation scheduling in grapevine (*Vitis vinifera* L.). An  
758 open gate to improve water-use efficiency? *Agriculture, Ecosystems &*  
759 *Environment* 106 (2-3), 159–170.  
760 URL <http://linkinghub.elsevier.com/retrieve/pii/S0167880904002956>
- 761 des Gachons, C. P., Leeuwen, C. V., Tominaga, T., Soyer, J.-P., GaudillÁire,  
762 J.-P., Dubourdieu, D., 2005. Influence of water and nitrogen deficit on fruit  
763 ripening and aroma potential of *vitis vinifera* l cv sauvignon blanc in field  
764 conditions. *Journal of the Science of Food and Agriculture* 85 (1), 73–85.
- 765 Destercke, S., Buche, P., Charnomordic, B., jan. 2013. Evaluating data relia-  
766 bility: An evidential answer with application to a web-enabled data ware-

767 house. Knowledge and Data Engineering, IEEE Transactions on 25 (1), 92  
768 -105.

769 Escalona, J., Flexas, J., Medrano, H., 2002. Drought effects on water flow,  
770 photosynthesis and growth of potted grapevines. *Vitis* 41, 57–62.

771 Ferreira, M. I., Silvestre, J., Conceição, N., Malheiro, A. C., Jun. 2012. Crop  
772 and stress coefficients in rainfed and deficit irrigation vineyards using sap  
773 flow techniques. *Irrigation Science* 30 (5), 433–447.  
774 URL <http://www.springerlink.com/index/10.1007/s00271-012-0352-2>

775 Ginestar, C., Eastham, J., Gray, S., Iland, P., 1998. Use of sap flow sensors  
776 to shedule vineyard irrigation. 1. effects of post-veraison water deficit on  
777 water relation, vine growth, and yield of shiraz grapevines. *Am. J. Enol.*  
778 *Vitic.* 49, 413–420.

779 Guarino, N., Oberle, D., Staab, S., 2009. What is an ontology? In: Staab,  
780 S., Studer, R. (Eds.), *Handbook on Ontologies*. International Handbooks  
781 on Information Systems. Springer Berlin Heidelberg, pp. 1–17.

782 Hilgert, N., Mas, A., Verzelen, N., 2013. Minimax adaptive tests for the  
783 functional linear model. *The Annals of Statistics* 41 (2), 838–869.

784 James, G. M., Wang, J., Zhu, J., 2009. Functional linear regression that’s  
785 interpretable. *The Annals of Statistics* 37, 2083–2108.

786 Jones, H. G., Nov. 2004. Irrigation scheduling: advantages and pitfalls of  
787 plant-based methods. *Journal of experimental botany* 55 (407), 2427–36.  
788 URL <http://www.ncbi.nlm.nih.gov/pubmed/15286143>



- 789 Koundouras, S., Marinos, V., Gkoulioti, A., Kotseridis, Y., van Leeuwen, C.,  
790 2006. Influence of vineyard location and vine water status on fruit mat-  
791 uration of nonirrigated cv. agiorgitiko (*vitis vinifera* l.). Effects on wine  
792 phenolic and aroma components. *Journal of Agricultural and Food Chem-*  
793 *istry* 54 (14), 5077–5086.
- 794 Kruchten, P., Nov. 1995. The 4+1 view model of architecture. *IEEE Softw.*  
795 12 (6), 42–50.  
796 URL <http://dx.doi.org/10.1109/52.469759>
- 797 Miller, G. A., 1956. The magical number seven, plus or minus two: Some  
798 limits on our capacity for processing information. *Psychological Review*  
799 63, 81–97.
- 800 Montoro, A., Fereres, E., Lopez-Urrea, R., Manas, F., Lopez-Fuster, P., 2011.  
801 Sensitivity of trunk diameter fluctuations in *vitis vinifera* l. tempranillo and  
802 cabernet sauvignon cultivars. *American Journal of Enology and Viticulture*  
803 63(1), 85–93.
- 804 Musen, M., 1992. Dimensions of knowledge sharing and reuse. *Computers*  
805 *and Biomedical Research* 25 (5), 435–467.
- 806 Ojeda, H., Andary, C., Kraeva, E., Carbonneau, A., Deloire, A., 2002. Influ-  
807 ence of pre-and postveraison water deficit on synthesis and concentration  
808 of skin phenolic compounds during berry growth of *vitis vinifera* cv. shiraz.  
809 *American Journal of Enology and Viticulture* 53 (4), 261–267.
- 810 Olivo, N., Girona, J., Marsal, J., 2009. Seasonal sensitivity of stem water

811 potential to vapour pressure deficit in grapevine. *Irrigation Science* 27,  
812 175–182.

813 Parker, A., De Cortázar-Atauri, I., Van Leeuwen, C., Chuine, I., 2011. Gen-  
814 eral phenological model to characterize the timing of flowering and verai-  
815 son of *vitis vinifera* l. *Australian Journal of Grape and Wine Research* 17,  
816 206–216.

817 Picón-Toro, J., González-Dugo, V., Uriarte, D., Mancha, L. a., Testi, L.,  
818 Jun. 2012. Effects of canopy size and water stress over the crop coefficient  
819 of a Tempranillo vineyard in south-western Spain. *Irrigation Science*  
820 30 (5), 419–432.  
821 URL <http://www.springerlink.com/index/10.1007/s00271-012-0351-3>

822 Quinlan, J., 1986. Induction of decision trees. *Machine learning* 1 (1), 81–106.

823 Quinlan, J., 1993. *C4. 5: programs for machine learning*. Morgan Kaufmann.

824 R Development Core Team, 2009. *R: A Language and Environment for Sta-*  
825 *tistical Computing*. R Foundation for Statistical Computing, Vienna, Aus-  
826 tria, ISBN 3-900051-07-0.  
827 URL <http://www.R-project.org>

828 Ramsay, J. O., Silverman, B. W., 2005. *Functional data analysis*, 2nd Edition.  
829 Springer Series in Statistics. Springer, New York.

830 Raymond, K., 1995. Reference model of open distributed processing (rm-  
831 odp): Introduction. In: *Proceedings of the IFIP TC6 International Con-*  
832 *ference on Open Distributed Processing*. pp. 3–14.

- 833 Rijgersberg, H., van Assem, M., Top, J. L., 2013. Ontology of units of mea-  
834 sure and related concepts. *Semantic Web* 4 (1), 3–13.
- 835 Rodrigues, P., Pedroso, V., Gouveia, J. P., Martins, S., Lopes, C., Alves, I.,  
836 2012. Influence of soil water content and atmospheric conditions on leaf  
837 water potential in cv. touriga nacional deep-rooted vineyards. *Irrigation*  
838 *Science* 30, 407–417.
- 839 Santesteban, L. G., Miranda, C., Royo, J. B., 2011. Suitability of pre-dawn  
840 and stem water potential as indicators of vineyard water status in cv.  
841 Tempranillo. *Australian Journal of Grape and Wine Research* 17 (1), 43–  
842 51.  
843 URL <http://dx.doi.org/10.1111/j.1755-0238.2010.00116.x>
- 844 Thomopoulos, R., Destercke, S., Charnomordic, B., Iyan, J., Abécassis, J.,  
845 Feb. 2013. An iterative approach to build relevant ontology-aware data-  
846 driven models. *Information Sciences* 221, 452–472.
- 847 Ullah, S., Finch, C. F., 2013. Applications of functional data analysis: A  
848 systematic review. *BMC Medical Research Methodology* 13 (1), 43.  
849 URL <http://www.biomedcentral.com/1471-2288/13/43>
- 850 Van Leeuwen, C., Tregoat, O., Choné, X., Bois, B., Pernet, D., Gaudillère,  
851 J.-P., et al., 2009. Vine water status is a key factor in grape ripening and  
852 vintage quality for red bordeaux wine. How can it be assessed for vineyard  
853 management purposes? *J. Int. Sci. Vigne Vin* 43 (3), 121–134.
- 854 Villanueva-Rosales, N., Dumontier, M., 2008. Modeling life science knowledge  
855 with owl 1.1. In: *Proceedings of OWL'08*.

- 856 Williams, L., Baeza, P., 2007. Relationships among ambient temperature and  
857 vapor pressure deficit and leaf and stem water potentials of fully irrigated  
858 field-grown grapevines. *Am. J. Enol. Vitic.* 58, 2.
- 859 Zhang, Y., Kang, S., Ward, E. J., Ding, R., Zhang, X., Zheng, R., 2011.  
860 Evapotranspiration components determined by sap flow and microlysime-  
861 try techniques of a vineyard in northwest China: Dynamics and influential  
862 factors. *Agricultural Water Management*.  
863 URL <http://dx.doi.org/10.1016/j.agwat.2011.03.006>