Data Mining for Rules Extraction to Virgin Olive Oil Optimization Process

A. Jiménez Márquez
Instituto Andaluz de Investigación y Formación Agraria, Pesquera, Alimentaria y de la Producción Ecológica (IFAPA), Junta de Andalucía Centro ‘Venta del Llano’
PO. Box 50, E-23620 Mengibar, Jaén. Spain

G. Beltrán Maza
Instituto Andaluz de Investigación y Formación Agraria, Pesquera, Alimentaria y de la Producción Ecológica (IFAPA), Junta de Andalucía Centro ‘Venta del Llano’
PO. Box 50, E-23620 Mengibar, Jaén. Spain

ABSTRACT
From historical database of virgin olive oil elaboration process and data mining techniques can be extracted useful information of the process to its optimization. In this work data from an analytical and process variables database, collected trough four crop years, was used and a spurious data filtration and a tree decisional J48 algorithm based protocol was applied. The results obtained allow see the oil mill behavior under different conditions of process variables accord two objectives: the performance by the oil loss in olive cake and oil quality by the oil polyphenols content. This information extracted can allow the building of rules for variables combination to get the objectives.

General Terms
Discovering Patterns, Machine Learning, Data Anlaysis

Keywords
Data Mining, Olive Oil, EVOO Process, Optimization, Olive Cake, Polyphenols

1. INTRODUCTION
At present, many industrials and other process are widely computerized. The rapid spread of computer with an improvement in the speed and data storage has enabled the implementation and use of the automation system and new communication technologies that generate a large volume of data, which are a great value by the information they contain. This information is collected on stored databases of historical time series from continuous or discrete measurement and data process obtained automatically by sensors or operator hand. These databases are a great source of knowledge, but the large number of variables process, operating conditions and other system perturbations can difficult its extraction. In this sense, data mining techniques can be used in order to extract and generate previously unknown process knowledge for use to improve the performance process, product quality and environmental improvement.

Data mining can be considered as a multidisciplinary statistic, artificial intelligence, data visualization technique for extracting useful information from large data set [1][2]. Data mining extracts the patterns in the data and the associations or relationships between the variables by using different statistical and mathematical techniques in order to creation of models to find out the hidden information. This technique is very applied at different sectors [3]: economic science [4], biological and medical science [5], industrial process [6] or new technologies [7] and by its philosophy can be useful in optimization of the virgin olive oil elaboration process.

The Virgin olive oil elaboration process is a multistage transformation process where olive fruits are crushing, kneading and centrifuged for oil physical separation from solids and vegetation water. Each stage is characterized by certain process variables and its values have a great impact on performance and quality of oil obtained [8]. They are a large number of these variables and its regulation is very difficult by its multifactorial relationship. At present, this regulation is carried out by manual and subjectively procedures in basis on mill technician knowledge from visual controls. The last envelope of SCADA systems on oil mills it’s a great evolution in this sector and a large data volume are generated from on-line sensors, but this information is poorly used for process information and its optimization. In this sense, data mining, can aid by the analysis of the historical databases for extracting certain rules of behavior, process variable hierarchical structures, decisional trees, and other information that they are useful for manual optimization and necessary for ‘soft computing’ system control implementation as ‘Fuzzy Logic’, ‘Artificial Neuronal Network’ [9][10] or ‘Expert system’.

In this work, from four crop year, a historical process database was employed for analyze the viability of data mining to extracting information about process. Two objectives were considered: fat on dried matter in olive cake (for industrial performance) and total polyphenols content (for oil quality).

2. MATERIAL AND METHODS

2.1 Oil mill.
The work was developed using the continuous system of two phases installed in “Venta del Llano” IFAPA Centre [11]. This is a traditional continuous system “SC-90” (Pieralisi SA) and this is a system constituted basically by a hammer crusher with mobile grids and regulated hammer rotation speed, a horizontal three bodies thermomalaxer, of 500 kilograms of capacity each one, a solid-liquid horizontal centrifugal decanter (HCD) prepared for two exits and a centrifugal plate separator for the oil clarification. The HCD owns a nominal capacity of 45 t/day, with a bowl of 120 centimeters in length and 35.3 centimeters of diameter, which turns to 3500 RPM. In Figure 1, a schematic oil mill plant is detailed.

All the process is integrated in a computerized control system, type SCADA. Along of the productive process they are installed different sensors and actuators, which allow the process data visualization in real time, as well as the immediate action on the handling variables of the elaboration system, which allows carry out a control in real time of the olive oil elaboration process.
3. RESULTS
3.1 Data pretreatment.
The data selected for datamining analysis provide a wide range of values. Process and analytical variables they represent the more common situations and regulation conditions to oil mill optimization. Table 1 shows the range of these.

Table 1. Variables and its value range used from historical database of process.

| Process Variables Denomination | Range |
|--------------------------------|-------|
| Process:                      |       |
| Sieve diameter (mm)           | Sd    | 5, 6, 7 |
| Hammer mill speed (RPM)       | Hms   | 2000 - 3000 |
| Paste temperature malaxer (°C)| Ptm   | 9.0 - 36.7 |
| Paste kneading time (min)     | Pkt   | 35 - 175 |
| MTN addition (%)              | MTN   | 0 - 1 |
| Water addition DCH (L h⁻¹)    | WaDCH | 0 - 200 |
| Paste flow DCH (kg h⁻¹)       | PIDCH | 460 - 1300 |
| Oil point overflow DCH (mm)   | OgoDCH| 98, 99, 100 |

Analytical:

|          | Denotation | Range |
|----------|------------|-------|
|          | Ocf (%)    | 38.41 - 65.80 |
|          | Ocm (%)    | 15.10 - 30.87 |
|          | Ocf (%)    | 55.83 - 77.30 |
|          | Ocm (%)    | 1.66 - 5.92 |

Olive oil total polyphenols (mg kg⁻¹) OoTPLF 180 - 7227

Two objective output were considered. For performance industrial the fat loss on olive cake and for virgin olive oil quality the total polyphenols content. Olive cake is a byproduct which fat content indicate a good or poor elaboration process and it is verified by the olive cake fat on dried matter (OcfDM). Therefore, polyphenols are a very important components related with nutritional and sensorial properties of virgin olive oils.

Both objectives were subdivided at nominal classes from the practical and traditional experience on oil mill, and accord to olive fruit cultivar, agricultural area and machinery. Thus, OcfDM and OoTPLF were subdivided in four classes: good, reasonable, poor, bad and low, mid, high, upper, respectively. The range for each class is show in Table 2.

Table 2. Class for objective parameters.

| Class                          | Range |
|--------------------------------|-------|
| Olive cake fat on dried matter (OcfDM) |       |
| Good                           | < 7%  |
| Reasonable                     | 7% < > 8.5% |
| Poor                           | 8.5% < > 10% |
| Bad                            | > 10% |

| Olive oil total polyphenols (OoTPLF) |       |
| Low                             | < 450 mg kg⁻¹ |
| Mid                             | 450 mg kg⁻¹ -> 1250 mg kg⁻¹ |
| High                            | 1250 mg kg⁻¹ -> 2500 mg kg⁻¹ |
| Upper                           | > 2500 mg kg⁻¹ |

3.2 Data mining for industrial performance.

Decision tree is a supervised learning method that allows assigns class labels to a pattern set by developing models for each class. For classification, the dataset was split on the basis of discrete decisions using the thresholds of the attributes values, the tree search variables and cutoff that best discriminate the groups to classify. It is one of the most widely used classification method and typical on industrial
with high process variables. In this work the pattern set were constituted by all plant process variables and analytical olive oil variables: olive fruit moisture (Ofm) and olive fruit fat (Off). The olives fruits characteristics have critical information necessary to understand the behavior of olive oil elaboration process and the multifactorial relationship between the variables. The application of J48 algorithm to training set, and nominal class defined for OcDM, allows building a model whose validation show a high correct classification percentage, with value above of 73 %. In Table 3 the confusion matrix from validation set is showed. The correct classification for each class is 62.5%, 62.5%, 67.2% and 80.8% for Good, Reasonable, Poor and Bad, respectively. Good and Bad are widely discriminated, Reasonable and Poor they have interface values with their neighbor class. On create range of OcDM class, the values gap with the neighbor classes is small, they are values that they could be in both range of OcfDM class, the values gap with the neighbor classes is characterized by the olive fruit moisture, the paste temperature from malaxer, the olive fat on Dried Matter.

Table 3. Confusion Matrix obtained for Olive Cake fat on Dried Matter.

| Classified as | Good | Reasonable | Poor | Bad |
|---------------|------|------------|------|-----|
| Good          | 10   | 6          | 0    | 0   |
| Reasonable    | 6    | 20         | 5    | 1   |
| Poor          | 2    | 12         | 41   | 6   |
| Bad           | 0    | 0          | 15   | 63  |

From generated decision tree, showed in Figure 2, its can observed as the first four levels are characterized by the olive fruit moisture, the paste temperature from malaxer, the coaduyant addition and fruit milling conditions: sieve diameter and hammer rotation velocity. Olive fruit moisture is the first nodule discriminant with a splitter value of 55.85 %.

![Figure 2: Decision tree for Olive Cake fat on Dried Matter. Abbreviation as Table 1 and Table 2.](image-url)
The knowledge of moisture data is essential for the optimization process because on decanter they are a solid-liquid-liquid separation by the centrifugal forces. The water content on olive paste generates an intermediate ring hydraulic between the light oil phase and heavy solid matter phase. The thickness of this ring hydraulic is defined by the water fruit content and define the oil output overpoint from decanter, its moisture and its solid organic matter content, which have an important relationship with the oil loss in the cake [10]. Furthermore, the olive oil elaboration process takes place throughout several months of winter. Throughout this period, the fruit has important changes in its characteristics and composition as a result of its ripeness state. At the beginning of crop year we have fruits with a high water content, which decreased to low water content, at the end of the campaign. This situation also occurs at the same time of picking olives, as a result of the coming into fruits from different zones and parcels (zones of mountain or countryside, dry or irrigation farming, different varieties) in the oil mill. The ‘Ofm’ cutoff, established in decision tree, is a typical moisture value at beginning of crop year, only at exceptional years with high rain precipitations can become to higher values of 65 %, as observed in Table 1. This cutoff split two initial situations: unusual olive fruit moisture (>55.85%) and ordinary olive fruit moisture range (<55.85%). For first branch, when unusual, important information is extracted: only if ‘Ofm’ > 55.85% and pHDC < 800 kg h⁻¹ and Ptm < 21.8°C can be expected a fat loss in cake ‘Reasonable’ (OcfDM < 8.5%), in the other conditions a high fat loss will be produced. At the considered process variables, this oil mill machinery can’t work well when high ‘Ofm’ is considered, new process variables must be found for this situation.

When the normal ‘Ofm’ range is considered, temperature of paste out from malaxer is the next decisional node. The heating of the paste is the traditionally applied solution for loss fat reduction in olive cake, the temperature has effect on oil viscosity and provides the oil exit from organic matter and the kneading and slight water addition were carried out. The sieve diameter and hammer rotation velocity effects on OcfDM are the next important discriminant nodes; thus, when sieve diameter is ≤5mm, the hammer should rotate less 2700 r.p.m. and kneading time not greater than 75 min. for obtain ‘Reasonable’ results. If sieve diameter is greater, ‘Ofm’ define water addition or hammer rotation; thus, when olive fruit is dehydrated (Ofm ≤ 46.5%), mainly ‘Reasonable’ and ‘Good’ results are obtained when water addition is applied. In normal situation of olive fruit moisture (Ofm > 46.5%) low levels of water addition and short kneading times (around 60 min) generate ‘Reasonable’ results, but ‘Good’ results can be expected for kneading time around 90 min. and high kneading temperature.

As can be seen, the analysis of decision tree allows the extracting of certain rules to optimization the performance of process in basis to minimal fat loss in olive cake, the resume of these rules agrees with the usually practices carried out in olive mills and experimental result obtained from different researchers [12][13][14]. Flow paste reduction when high fruit moisture, water addition when low fruit moisture, temperature kneading increase when high fat loss are taking place, rpm and particle size reduction on milling when cool oil mill work, they are some real regulations observed.

In Figure 3 is showed the variability of olive cake fat on dried matter. The results were obtained when certain process variables, extracted from assays carried out, are considered while maintaining all other constant. This is in accordance with real assays and rules extracted. Significant decrease of fat in olive cake is observed when reduction of flow paste to DCH, increase of paste temperature, increase of time kneading and slight water addition were carried out. The sieve diameter and hammer rotation velocity effects on OcfDM allows to see the importance of this stage where it is necessary to adjust the hammer rotation on sieve diameter to avoid the formation of emulsions that decrease the industrial performance.

### 3.3 Data mining for olive oil quality

As above, when 484 algorithm was applied to training set and nominal class defined for ‘OoTPLF’, the building model on validation allows a very high correct classification percentage, with value above of 85 %, as its can see in the confusion matrix showed in Table 4. There is a high percentage of correct classification in each class: 85.2%, 88.9%, 71.9% and 95.5% for ‘Low’, ‘Mid’, ‘High’ and ‘Upper’ respectively, this means that the process variables used to build model are sufficient for explained the process behavior in olive oil polyphenol content. The polyphenols they are an important components of virgin olive oil whose availability is possible by the

![Figure 3. Data values of Olive cake fat on Dried Matter (OcfDM) and standard deviations (sd) for experimental conditions extracted from assays historical data base of oil mill.](image)
physical procedures applied for oil extraction from olive fruit. Antioxidant, healthy and sensorial contributions are the effects more known of them but its quantity on the oil is very affected by its oil-water diffusion constant. Water addition, temperature, interphases contact times are some the principal factors that they can affect their content in oil. The ‘Upper’ and ‘High’ class should be considered as objectives for the oils that output from DCH since the later clarification stage give rise to a decrease on content of them on finished product.

The tree decisional, building from J48-classifier (Figure 4), show the paste flow to DCH, the oil output point (level, diaphragm), the paste temperature and water addition as the main cutoffs in the first levels. These variables are related with both sedimentation (ts) and residence times (tr) [9][10], which affect to oil-water diffusion constant of polyphenols components.

### Table 4. Confusion Matrix obtained for Total Polyphenols in virgin olive oil.

| Classified as | Low   | Mid  | High | Upper |
|--------------|-------|------|------|-------|
| Low          | 23    | 2    | 2    | 0     |
| Mid          | 3     | 72   | 5    | 1     |
| High         | 0     | 9    | 41   | 7     |
| Upper        | 0     | 1    | 0    | 21    |

The first discriminant variable was the flow paste to DCH, with a cutoff around of 760 kg h⁻¹. This process variable is usually inversely correlated to the paste kneading time: the high paste flows they are often associated to kneading times shorts. But these comportments can be mechanically changed and both variables becoming at independents, high flows paste and high times kneading can coexist. From decisional tree (Figure 4), at low paste flows (<760 kg h⁻¹) and hammer crusher cold working (Ptm<19°C), not water addition is necessary for high polyphenols content. If ‘Ptm’ is around of 19°C, for very low flows (<650 kg h⁻¹), the oil output point from DCH should be far away of rotational axis of decanter (open the diaphragm) for high contents. The ‘upper’ and ‘high’ polyphenols content can been see when low paste flow, low or null water addition, total or midpoint of diaphragm level and high kneading times have been set.

When they are a normal and high paste flows, the diaphragm level has a significant contribution. This parameter is related with the ‘ts’ and ‘tr’ and useful volume of DCH. When diaphragm is open (≥99 mm) a DCH useful volume reduction and a minor sedimentation time is established. High moisture and organic matter impurities in oil from DCH can be seen. When closed (98 mm) a more time contact between interphases is originates and a light heating of paste (<25 °C) along with low water addition and middle times kneading, make it possible to ‘upper’ and ‘high’ class polyphenols. A larger quantity of samples classified on low ‘OoTPLF’ levels can been seen when opening diaphragm is considered. The larger samples classification into ‘high’ and ‘upper’ class was observed if water addition is carried out and the temperature

---

**Figure 4. Decision tree for Olive oil Total Polyphenols.**

**Abbreviation as Table 1 and Table 2.**
paste is on 22 -30 °C interval, the paste flow not higher to 1000 kg h⁻¹ and there is a high velocity rotational of hammer mill.

From these results it can be noted as the total polyphenols content is very influenced by the paste flow vs time kneading, the temperature paste and oil out overpoint vs water addition. So, when low paste flow working high level of polyphenols can be obtained if paste temperature is under 25ºC, there is low water addition and diaphragm is closed. For greater paste flows working, heating of paste, opening the diaphragm and slight water addition is necessary to increase quantity of them.

In this situation the oil out from DCH has high levels of moisture and organic matter, which contributes to total content of polyphenols in oil.

As with the previous parameter, these rules are in accordance with the real practices in oil mills and result obtained from different researchers that quantified the ‘OoPLF’ in oils clarified and filtered [12][13][14]. Figure 5, show some results extracted from historical data base. Effects of time kneading, paste flow and paste temperature are showed. For high kneading times when flow paste to DCH decrease an increase of ‘OoTPLF’ is observed, if paste temperature on kneading increase an increase of polyphenols is also observed, if kneading times increase and paste temperature is higher to 28ºC a decrease of polyphenols is observed, but if this kneading temperature is lower to 25ºC a slight increase of them can be induced.

**Figure 5. Data values of Olive oil Total Polyphenols (OoTPLF) and standard deviations (sd) for experimental conditions**

| Experimental conditions | Assays | Olive oil Moisture (%) | Olive Fat (%) | Slad(mm) | RPM | Ptm (ºC) | Pkt(min > 28ºC) | Pkt(min < 25ºC) |
|-------------------------|--------|------------------------|---------------|----------|-----|----------|----------------|----------------|
| FDCH                    | 59.09±0.41| 20.61±0.28              | 7             | 2000    | 28  | 125-175  | 0              | ----           |
| Ptm                     | 56.02±2.19| 19.02±0.79              | 6             | 2500    | --- | 60       | 75             | 800            |
| Pkt(> 28ºC)             | 49.03±0.29| 19.99±0.84              | 6             | 2700    | 28  | 34       | ---            | 30             |
| Pkt(< 25ºC)             | 55.02±0.67| 19.06±0.39              | 6             | 2500    | 21  | 25       | ---            | 45             |

4. CONCLUSIONS
The availability of historical data bases of the main process variables from an oil mill, can allow discover knowledge about of dynamic process behavior, in a give ‘Almazara’, when techniques of data mining are applied.

Through the process objectives definition, sensors and actuator data collection, data filtering by spurious rejections, data mining application and evaluation of discovered knowledge in contrast to the knowledge that the mill technician has on real situations, can be extracted the accurate information to build rules of behavior that these process variables for a better automatization and optimization of the EVOO process by smart control techniques.

Expert knowledge in the virgin olive oil elaboration process it becomes essential for a suitable evaluation of rules obtained and its reduction to the most significant in order to they can be used on building two-way decisional fuzzy matrices, on building neural network models or other intelligent system for the simulation and control of the oil mill process. A industrial performance improvement and better oil quality control in fact can be expected by using the new technologies of ‘Big Data’, data processing and cloud computing from the large amount of data that actual EVOO mills generates which can be translate into better use of resources and a more sustainable process.

5. ACKNOWLEDGMENTS
This work was supported by P10-AGR6429 Excellent Project: ‘Modelling and Optimization of Virgin Olive Oil Elaboration Process’. Project 1’, funded by ‘Ministerio de Ciencia e Innovación’ and ‘Consejería de Economía, Innovación y Ciencia de la Junta de Andalucía’, Spain. Special thanks to Juan Torres and Justo Cardenas by their collaborations in oil mills and all the laboratory personnel.

6. REFERENCES
[1] Adamo, J.M. 2001 Data Mining for Association Rules and Sequential Patterns.Sequential and Parallel Algorithms. Springer, New York.
[2] Cios, K.J., Pedrycz, W., Swiniarski, R.W., Kurgan, L.A. 2007 Data mining: A knowledge discovery approach. Springer, New York.
[3] Silwattananusar, T., Tuamsuk, K. 2012 Data Mining and Its Applications for KnowledgeManagement : A Literature Review from 2007 to 2012. International Journal of Data Mining & Knowledge Management Process (IJDKP) 2(5), 13-24.
[4] H. Wang, H., Wang, S. 2008 A knowledge management approach to data mining process for business intelligence. Industrial Management & Data Systems 108(5), 622-634.
[5] Ting, S.L., Shum, C.C., Kowk, S.K., Tsang, A.H.C., Lee, W.B. 2009 Data Mining in Biomedicine: Current Applications and Further Directions for Research. Journal Software Engineering & Applications 2, 150-159.
[6] Soroush-Rohanizadeha, S., Bameni-Moghadam, M. 2009 A Proposed Data Mining Methodology and its
Application to Industrial Procedures. Journal of Industrial Engineering 4, 37-50.

[7] Last, M., Szczepaniak, P.S., Volkovich, Z., Kandel, A. 2006 Advances in Web Intelligence and Data Mining. Studies in Computational Intelligence, Vol. 23. Springer, New York.

[8] Hermoso, M., González, J., Uceda, M., García-Ortíz, A., Morales, J., Frias, L., Fernández, A. 1998 Virgin olive oil of quality. Elaboration by tow-phases process. 61/98. Consejería de Agricultura y Pesca. Junta de Andalucía. Sevilla. Spain.

[9] Jiménez, A., Aguilera, M.P., Uceda, M., Beltrán, G. 2009 Neural network as tool for virgin olive oil elaboration process optimization, Journal of Food Engineering, 95(1), 135-141.

[10] Jiménez, A., Aguilera, M.P., Beltrán, G., Uceda, M. 2008 A sensor-software based on artificial neural network for the optimization of olive oil elaboration process, Sensor and Actuator B, 129, 985-990.

[11] Jiménez, A., Hermoso, M., Uceda, M. 1995 Extraction of virgin olive oil by two-phases continuous system. Influence of different variables of process on certain parameters related to oil quality. Grasas y Aceites, 4-5 (29), 299-303.

[12] Di Giovacchino, L., Sestilli, S., Di Vicenzo, D. 2002 Influenze of olive processing on virgin olive oil quality. European Journal of Lipids Science and Technology, 104, 587-601.

[13] A.M. Linarejos-Garcia, A.M., Gomez-Rico, A., Salvador, M.D., Fregapane, G. 2009 Influence of maxalation conditions on virgin olive oil yield, overall quality and composition. European Food Research and Technology, 228, 112-120.

[14] Aguilera, M.P., Beltrán, G., Sanchez-Villasclaras, S., M. Uceda, Jiménez, A. 2010 Kneading olive paste from unripe ‘Picual’ fruits: I. Effect on oil process yield. Journal of food Engineering, 97, 533-538.