

Sorting of White Wines from Minho Region: An Experimental Analysis with CPP-TRI Method

Glauco Barbosa da Silva

Universidade Federal Fluminense (UFF)

Rua Passo da Pátria, 156 Campus Praia Vermelha Bloco D - sala 309 São Domingos, Niterói, RJ

Brasil CEP: 24.210-240

glaucos@id.uff.br

Phelipe Medeiros da Rocha

Universidade Federal Fluminense (UFF)

Rua Passo da Pátria, 156 Campus Praia Vermelha Bloco D - sala 309 São Domingos, Niterói, RJ

Brasil CEP: 24.210-240

pheliperocha@id.uff.br

Helder Gomes Costa

Universidade Federal Fluminense (UFF)

Rua Passo da Pátria, 156 Campus Praia Vermelha Bloco D - sala 309 São Domingos, Niterói, RJ

Brasil CEP: 24.210-240

heldergc@id.uff.br

ABSTRACT

The present article aims to analyze the CPP-TRI method through the sorting task in the Vinho Verde (white wine) dataset with 4898 instances. For this, from an initial experiment, the results achieved are presented, analyzed and compared with the results from SVM technique. The justification for the present work stems from the recent proposition of the CPP-TRI method, which still has few results evaluated when compared with methods more widespread in the literature.

KEYWORDS. CPP-TRI. MCDA. SORTING.

Paper topics: BDA – Big Data e Analytics

1. Introduction

Driven by the benefits it brings to health and based on a science (Oenology) responsible for observing and following the entire production process, from the planting of the grapes, through ageing, until the bottles reach the world's shelves; wine is a product increasingly appreciated by a wide range of consumers.

Faced with such a favourable scenario and in order to support its growth, the wine industry has invested in new technologies for the wine production and sales processes. In this context, key elements are certification and quality assessment. In general, certification assesses: the results of physical-chemical laboratory tests, which determine density, alcohol content and pH, among others; and sensory tests, resulting from expert evaluation.

According to [Smith e Margolskee, 2006], the description of human taste perception is influenced by four factors: salinity, acidity, sweetness and bitterness. Conjugating these factors makes the classification of wines an extremely complex and difficult task, which has attracted many researchers to analyze the relationship between physicochemical and sensorial tests. [Souza-Coutinho et al., 2020] claim that the recognition of wine quality is not an easy task in a complex beverage in which preference is driven not only by sensory factors, but also by demographic, cultural or psychological aspects. [Ferreira e Loureiro, 2019] highlight the complexity of the relationship between physicochemical and sensory tests and state that there are few studies with Portuguese wines.

In this paper, a dataset with 4898 instances of Vinho Verde (white), where each instance is composed by the results of the physicochemical and sensory tests of the sample, is classified by the CPP-TRI method. The data are available in the University of California, Irvine - Machine Learning Repository [Bache e Lichman, 2013] and were collected between May 2004 and February 2007. The proposed preference modelling, based on the available analytical data from the certification stage, seeks to contribute to the improvement of the decision-making process.

The justification for the present work stems from the recent proposition of the CPP-TRI method, which still has few results evaluated when compared with methods more widespread in the literature. CPP-TRI has been applied to problems in different areas such as: in the classification of municipalities regarding the supply of daycare [Sant'Anna et al., 2013]; in the classification of courses by multiple evaluators [Sant'Anna et al., 2016]; in the classification of federation units based on the economically occupied population [(Silva et al., 2016)]; and in the selection and prioritization of a set of bibliographic data [Treinta et al., 2014].

As related studies, [Cortez et al., 2009] proposed a Support Vector Machine(SVM) based approach to classify this same dataset and [Silva e Costa, 2017] analyzed a dataset of red wines.

The present paper aims to present an experimental analysis of the Probabilistic Preference Composition Method(CPP-TRI). From the results of physicochemical tests of a sample of Vinho Verde (white) - Minho Region, a data set is classified by the CPP-TRI method. Then, the results obtained by the CPP-TRI method are analyzed and compared to the results obtained in a related work that used the SVM technique for the classification of the same dataset.

This paper is organized in five sections, as follows: Section 2 describes the CPP-TRI method; Section 3 presents the characteristics of the data set; Section 4 applies the CPP-Tri method and analyzes the results; and Section 5 presents the considerations.

2. CPP-TRI Method

The Probabilistic Composition of Preferences method in its Tricotomic mode (CPP-TRI) is defined as an approach to the problem of sorting, which consists in allocating alternatives(instances) in classes predetermined by reference profiles(vector of evaluations

according to each criterion). It is based on the probabilistic transformation proposed by [Sant'Anna et al., 2012], similarly to the transformation of crisp numbers into pertinence intervals of Fuzzy Set Theory [Zadeh, 1965].

The CPP-TRI method has as its main characteristic to take into account the inaccuracies existing in the assessments according to each criterion; it seeks to associate numerical measures to probability distributions. In general, it is assumed, besides normal distributions, identical distribution and independence among the perturbations that cause imprecision in the measures; the data of the available evaluations in each criterion are used to estimate the variances of these normal distributions [Sant'Anna et al., 2012]

[Sant'Anna et al., 2015] claim that an advantage of modelling with probability distributions is to free the decision maker from the task of assigning weights or determining boundary parameters.

Once the evaluations are transformed into probability distributions, the preference model is constructed by the probability distance measures(δ), calculated for all j classes (ordinates), as the difference between the probabilities of being above the j class profile and of being below the j class profile. For any alternative (instance) a_i ; a_i is compared with the reference profile of each class j ; resulting in one probability of being above class j and another probability of being below class j . a_i is allocated to the nearest class, i.e., the class j that has the smallest absolute value of (δ).

3. Data Features

The sample analyzed is composed of 4898 instances of Vinho Verde (white) – DOC - <https://www.vinhoverde.pt/pt/denominacao-de-origem-controlada> – with characteristics of the producing region of northwestern Portugal (Minho). To provide a better understanding of the evaluations and the classification of the wines, some considerations about their main components are highlighted, without the pretension of exhausting the subject.

Among the components present in laboratory tests, [Waterhouse, 2020] describes the main ones, as follows:

- a) Acidity - is a fundamental property, which mainly contributes to taste and resistance to microbial infection; it affects taste and pH, which influence color, oxidation stability and, consequently, the overall shelf life of a wine. Some acids arise as a result of the fermentation process of yeast and/or bacteria, i.e. these parameters are relevant in wine quality. Total acidity is divided into two groups: **volatile acidity** and **fixed acidity**. The volatile acidity defines the wine's spoilage, consisting largely of acetic acid (vinegar).
- b) Citric acid - is an organic acid that has many uses in wine production; often used as a natural preservative. In others added to wines to increase acidity and give a "fresh" taste. The disadvantage of adding citric acid is microbial instability, since bacteria use citric acid in their metabolism, it can increase the growth of unwanted microbes.
- c) Residual sugar - is the amount of sugar remaining after fermentation. The main reasons for this are: the yeast's inability to convert sugar into alcohol; the interruption of fermentation by the presence of a "heavier" alcohol; and the significant drop in temperature, which ends up killing the yeast. The residual sugar is balanced by acidity, alcohol content, and tannins.
- d) Chloride (sodium chloride) - is the amount of salt in wine.
- e) Sulfur dioxide - is a preservative present in all wines.
- f) Density - can be used as a quick check measure of homogeneity and consistency.

- g) pH - describes how acidic or basic a wine is, ranging on a scale of 0 (very acidic) to 14 (very basic). Most wines are between 3 and 4.

For the eleven main sample parameters from the laboratory tests, and considered as criteria for evaluation, table 1 presents a consolidation of the measures of interest.

Table 1. *measures of interest*

criteria	Description	Min	Max	Mean	Std- Dev
g ₁	Fixed acidity	3.8	14.2	6.855	0.844
g ₂	Volatile acidity	0.08	1.1	0.278	0.101
g ₃	Citric acid	0.0	1.66	0.334	0.121
g ₄	Residual sugar	0.6	65.8	6.391	5.072
g ₅	Sodium chloride	0.009	0.346	0.046	0.022
g ₆	Free Sulfur dioxide	2.0	289.0	35.08	17.007
g ₇	Total Sulfur dioxide	9.0	440.0	138.361	42.498
g ₈	Density	0.987	1.039	0.994	0.003
g ₉	pH	2.72	3.82	3.188	0.151
g ₁₀	Sulphates	0.22	1.08	0.49	0.114
g ₁₁	Alcohol	8.0	14.2	10.514	1.231

In sensory tests, the quality attribute is derived from the average of the evaluations of at least three experts, who assign scores ranging from 0 (very bad) to 10 (excellent). For each instance, the experts' evaluation is registered in the Quality attribute; seven classes were identified for the analyzed set. The distribution of the instances according to the experts' evaluation (quality) is presented in figure 1.

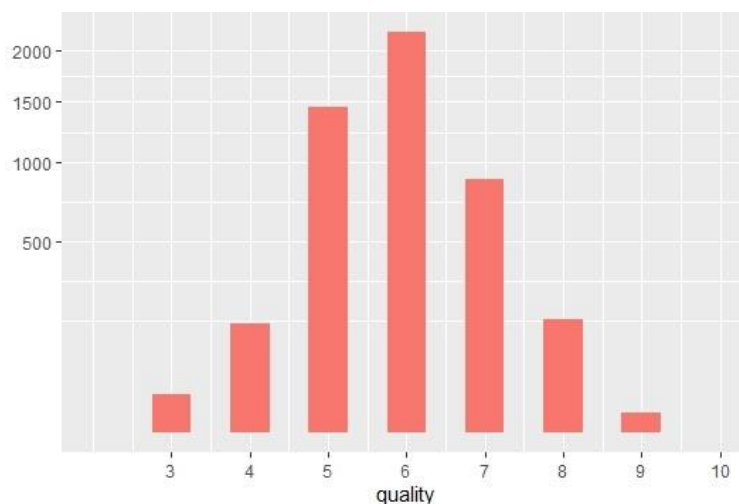


Figure 1. Distribution of instances per attribute Quality.

	Classes							total
	3	4	5	6	7	8	9	
Instances	20	163	1457	2198	880	175	5	4898

4. Application of Method

In this section the CPP-TRI method is applied to the dataset. As proposed by [Sant'Anna, 2014], in formulating a classification problem, we adopt:

- ✓ $G = (g_1, \dots, g_k)$, as a set of k criteria.
- ✓ $A = (a_1, \dots, a_k)$, as a vector of dimension R_k that holds the evaluations according to each k criteria for the instance to be classified.
- ✓ $C = (C_1, \dots, C_j)$, is a set of j classes, ordered from worst to best, so that an instance is as better the higher the index of the class it is rated.

Based on the laboratory tests, for each class C_j a representative class profile is established - (C_{jk}) , where each dimension of the profile is defined by the mean and standard deviation for each g_k criteria; the representative profile for each class is presented in table 2.

Table 2. Representative Profiles by Class.

Class		g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
C9	Mean	7.42	0.298	0.386	4.12	0.027	33.4	116	0.991	3.308	0.466	12.18
	Stdev	0.983	0.058	0.082	3.759	0.007	13.428	19.824	0.003	0.083	0.093	1.013
C8	Mean	6.657	0.277	0.327	5.671	0.038	36.72	126.166	0.992	3.219	0.486	11.636
	Stdev	0.819	0.108	0.085	4.263	0.013	16.204	33.006	0.003	0.152	0.147	1.28
C7	Mean	6.735	0.263	0.326	5.186	0.038	34.126	125.115	0.992	3.214	0.503	11.368
	Stdev	0.756	0.091	0.079	4.298	0.011	13.245	32.743	0.003	0.158	0.13	1.247
C6	Mean	6.838	0.261	0.338	6.442	0.045	35.651	137.047	0.994	3.189	0.491	10.575
	Stdev	0.837	0.088	0.119	5.167	0.02	15.736	41.286	0.003	0.151	0.113	1.148
C5	Mean	6.934	0.302	0.338	7.335	0.052	36.432	150.905	0.995	3.169	0.482	9.809
	Stdev	0.838	0.1	0.141	5.328	0.026	18.146	44.086	0.003	0.141	0.098	0.847
C4	Mean	7.129	0.381	0.304	4.628	0.05	23.359	125.279	0.994	3.183	0.476	10.152
	Stdev	1.081	0.173	0.164	4.164	0.026	20.391	52.754	0.002	0.164	0.118	1.003
C3	Mean	7.6	0.333	0.336	6.393	0.054	53.325	170.6	0.995	3.187	0.475	10.345
	Stdev	1.725	0.141	0.081	5.317	0.046	69.421	107.758	0.003	0.21	0.12	1.224

In CPP-TRI method, [Sant'Anna, 2014] establishes that no criterion a profile of a higher class may present an evaluation lower than that of a profile of a lower class by that same criterion and by at least one criterion the profile of the higher class presents evaluation higher than that of the profile of the lower class.

With a brief analysis of table 2, it is possible identifying that profiles are not ordered as it is required in [Sant'Anna, 2014] to apply the CPP-Tri algorithm. Thus, the category profiles are handled to avoid attributes in different directions, i.e., all attributes (profile) in a highest category are better performance (at least equal performance) than a lowest category.

Algorithm to Eliminate Inversions in Representative Profiles*

Algorithm to adjust Representative Profiles –(Provided by Santanna, A.P)

Objective:

for all i_1, i_2 , with $i_2 > i_1$,

the coordinates of the profiles of a class i_1 must be all below the coordinates of the profiles of the classes i_2 .

Notation.

For any criterion, i the initial value of the criterion in the profile of class i is denoted by P_i and the final value (ie after elimination of the inversion) by Q_i .

Approach:

1) Identifying to be the central class (C) the class with more observations (instances).

If there is more than one class in this condition and there is inversion in some criterion between them, use the arithmetic mean of the inverted values in place of them.

2) For classes i with $i < C$: $Q_i = \text{minimum}(Q_i + 1, P_i)$.

3) For classes i with $i > C$: $Q_i = \text{maximum}(Q_i - 1, P_i)$.

So, applying the *Algorithm to adjust Representative Profiles*, from central class(C_6), new representative profiles are established as shown in table 3.

Table 3. New representative Class profiles (adjusted)

Class		g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
C9	Mean	- 6.657	0.298	- 0.326	- 4.12	-0.027	-33.4	-116	- 0.991	3.308	0.503	12.18
	Stdev	0.983	0.058	0.082	3.759	0.007	13.428	19.824	0.003	0.083	0.093	1.013
C8	Mean	- 6.657	0.277	- 0.326	- 5.186	-0.038	-34.126	-125.115	- 0.992	3.219	0.503	11.636
	Stdev	0.819	0.108	0.085	4.263	0.013	16.204	33.006	0.003	0.152	0.147	1.28
C7	Mean	-6.735	0.263	-0.326	-5.186	-0.038	-34.126	-125.115	-0.992	3.214	0.503	11.368
	Stdev	0.756	0.091	0.079	4.298	0.011	13.245	32.743	0.003	0.158	0.13	1.247
C6 (Central)	Mean	-6.735	0.263	-0.326	-5.186	-0.038	-34.126	-125.115	-0.992	3.214	0.503	11.368
	Stdev	0.837	0.088	0.119	5.167	0.02	15.736	41.286	0.003	0.151	0.113	1.148
C5	Mean	-6.934	0.261	-0.338	-7.335	-0.052	-36.432	-150.905	-0.995	3.169	0.482	9.809
	Stdev	0.838	0.1	0.141	5.328	0.026	18.146	44.086	0.003	0.141	0.098	0.847
C4	Mean	-7.129	0.261	-0.338	-7.335	-0.052	-36.432	-150.905	-0.995	3.169	0.476	9.809
	Stdev	1.081	0.173	0.164	4.164	0.026	20.391	52.754	0.002	0.164	0.118	1.003
C3	Mean	-7.6	0.261	-0.338	-7.335	-0.054	-53.325	-170.6	-0.995	3.169	0.475	9.809
	Stdev	1.725	0.141	0.081	5.317	0.046	69.421	107.758	0.003	0.21	0.12	1.224

For a class C_j , on the k -th criterion, we have:

$A_{ijk+} \leftarrow$ the probability that the instance a_i presents a value above the representative profile of class C_{jk} .

$$A_{ij+} = \prod_k A_{ijk+} \quad (1)$$

$A_{ijk-} \leftarrow$ the probability that the instance a_i presents a value below the representative profile of class C_{jk} .

$$A_{ij-} = \prod_k A_{ijk-} \quad (2)$$

The procedure for associating an instance to a given class, classification, is based on the difference between the probabilities of being above or below each class, probability distance (δr). The instance is classified in the class in which the probabilistic distance (δr) is minimal.

$$\delta r = |A_{ij+} - A_{ij-}| \quad (3)$$

In the calculation of probabilities, [Sant'Anna, 2014] points out that if the variance is small, a "single" criterion may be sufficient to determine the pertinence of an instance to the class. That is, the variance must be large enough so that the probability associated with the criterion is not too small and makes the other criteria irrelevant, which is, avoid the veto power of a single criterion.

From the dataset of Green(white) Wines[Bache and Lichman, 2013] and the representative profiles of the classes, files `profilesbyclass.txt`(mean) and `stdev.txt`(standard deviations), an experiment for the calculation of probabilities, probabilistic distances and classification was conducted with the R code, processed in the *Rstudio* tool, as presented below:

```
## i = instance index ( 1 -- 4898)
## j = classe index ( 1 -- 7)
## k = criterion index ( 1 -- 11)
WinesMod <- read.csv2("/WinesMod.csv")
stdevbyclass<-read.table("/stdevbyclass.txt", quote="", comment.char="")
profilebyclass<-read.table("/profilebyclass.txt",quote="", comment.char="")
A<-WinesMod
B<- profilebyclass
D<- stdevbyclass
acima<-A
abaixo<-A
Pacima<- A
Pabaixo<- A
Pacima<-Pacima[1:nrow(A), 1:nrow(B)]
Pabaixo<-Pabaixo[1:nrow(A), 1:nrow(B)]
C<-Pacima
Classe<-Pabaixo
for (i in 1:4898)
```

```

{
for (j in 1:7)
{
  for (k in 1:11)
  {
    abaixo[j,k]<-pnorm(B[j,k],A[i,k],D[j,k])
    acima[j,k]<-1-pnorm(B[j,k],A[i,k],D[j,k])
  }
  Pabaixo[i,j]<-prod(abaixo[j,])
  Pacima[[i,j]]<-prod(acima[j,])
}
}
C<-abs(Pacima-Pabaixo)
for (i in 1:nrow (A))
{
  for (j in 1:nrow (B))
  {
    Classe[i,j]<-ifelse (abs(C[i,j])=min(abs(C[i,])), 10-j, ":")
  }
}
View(Classe)
write.csv(Classe, file="ClasseWhite.csv")

```

The results of the classification of the instances by the CPP-TRI method are summarized in Table 4 and Table 5 presents the confusion matrix as follows.

Table 4: Distribution of the instances classified with the CPP-TRI.

Class	3	4	5	6	7	8	9	total
Instances	1200	304	819	347	632	229	1367	4898

Table 5. Confusion Matrix.- CPP-TRI.

Original Class	CPP-TRI- Predictions							
	C3	C4	C5	C6	C7	C8	C9	
C3	4	0	4	1	2	1	8	20
C4	25	7	17	6	22	3	83	163
C5	513	112	276	61	181	37	277	1457
C6	525	148	394	169	300	112	550	2198
C7	112	32	99	95	107	60	375	880
C8	21	5	28	15	19	15	72	175
C9	0	0	1	0	1	1	2	5
	1200	304	819	347	632	229	1367	

6. The accuracy and precision of the CPP-TRI method, by class, are consolidated in Table

Table 6. Accuracy and Precision by class.

	C3	C4	C5	C6	C7	C8	C9
Accuracy (%)	20	4.29	18.94	7.69	12.16	8.57	40
Precision(%)	0.3	2.3	33.7	48.7	16.93	6.55	0.15

From results of white wines classification with CPP-Tri, about 11.84% of instances are classified in the original category and about 35,44% in the original or closest category. In effect, within this domain the $T = 1.0$ distance is accepted as a good quality control process.

In the analogous experiment by Cortez et al. [2009], a classifier based on Support Vector Machine(SVM), with different tolerance levels, the results achieved were about 30% with $T=0$ and 90% with $T=1.0$ - superior to results from CPP-TRI. However, for classes with smaller number of instances, the hit rate was higher for CPP-TRI. The SVM model did not classify any instance in class 3 and did not get a hit for class 9.

Although the CPP-TRI results were below the SVM classifier, it is possible that better results can be achieved. Cortez et al. [2009] suggest that not all attributes are relevant. Thus, the first action demanded was to evaluate the strength of the association between two attributes, Pearson's correlation(r).

Then, strong associations are identified like: Fixed acidity – pH; Residual Sugar – Total Sulfur Dioxide; Residual Sugar – Density; Residual Sugar – Alcohol; Free Sulfur Dioxide – Total Sulfur Dioxide; Total Sulfur Dioxide – Density; Total Sulfur Dioxide – Alcohol; Density – Alcohol.

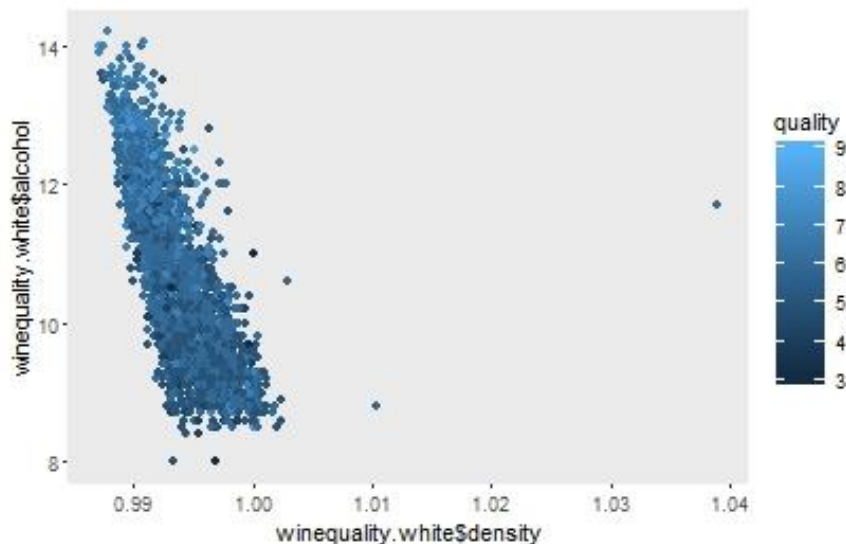


Figure 2. Density – Alcohol ($r = -0.780$).

5. Conclusions

The present work analyzed the ordered classification of Vinho Verde(white) wines with the CPP-TRI method. The task of wine classification is quite complex and difficult since it combines physicochemical and sensory parameters. The results achieved by the method were compared to the results obtained in a previous work that used the SVM technique for the classification of the same dataset.

Considering the Majority Class with 2198 instances, which represents about 45% of all instances, the complexity of the data set can be corroborated by the result obtained by the SVM technique that is a widely spread technique. On the other hand, it is relevant to highlight the result obtained by the CPP-TRI for classes with fewer instances.

Although the results are not superior to SVM, it is believed that the CPP-TRI classification may have been influenced (and compromised) by the nature of the data: the classes are not balanced and there is correlation among the attributes.

It is suggested that in future work, an attribute selection step be included before the method is applied.

Referências

- Bache, K. & Lichman, M. (2013). *Repository, UCI Machine Learning*. Irvine, CA: University of California.
- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4). <https://doi.org/10.1016/j.dss.2009.05.016>
- Ferreira, M. M., & Loureiro, V. (2019). *O Vinho Sentido*. Plátano.
- Parracho Sant'Anna, A., Faria, F., & Gomes Costa, H. (2013). Aplicação da Composição Probabilística e do Método das K-MÉDIAS à Classificação de Municípios Quanto à Oferta de Creches. *Cadernos Do IME - Série Estatística*, 34(1). <https://doi.org/10.12957/cadest.2013.15771>
- Sant'Anna, A. P. (2014). *Probabilistic Composition of Preferences, Theory and Applications*. Springer International Publishing.
- Sant'Anna, A. P., Costa, H. G., & Pereira, V. (2015). CPP-TRI: A sorting method based on the probabilistic composition of preferences. *International Journal of Information and Decision Sciences*, 7(3), 193–212. <https://doi.org/10.1504/IJIDS.2015.071372>
- Sant'Anna, A. P., Costa, H. G., Nepomuceno, L. D. O., & Pereira, V. (2016). A probabilistic approach applied to the classification of courses by multiple evaluators. *Pesquisa Operacional*, 36(3), 469–485. <https://doi.org/10.1590/0101-7438.2016.036.03.0469>
- Sant'Anna, Annibal Parracho, Costa, H. G., & Pereira, V. (2012). CPP-Tri: um método de classificação ordenada baseado em composição probabilística. *Relatórios de Pesquisa Em Engenharia de Produção (UFF)*, 12(8).
- Silva, G. B. da, Barros, M. D. de, & Costa, H. G. (2016). *Abordagem Híbrida Multicritério – Mineração de Dados Aplicada a Classificação de Unidades da Federação com Base na População Economicamente Ocupada*. <https://doi.org/10.5151/marine-spolm2015-140483>
- Silva, G. B. da, & Costa, H. G. (2017). Análise Experimental do Método de Composição Probabilística de Preferências em seu modo Tricotômico (CPP - TRI) na Classificação Ordenada de Vinhos da Região do Minho. *Anais Do XLIX SBPO*. <http://www.sbp2017.iltc.br/pdf/169519.pdf>
- Smith, D. V., & Margolskee, R. F. (2006). Making Sense of Taste. *Scientific American*, 16(3s). <https://doi.org/10.1038/scientificamerican0906-84sp>
- Souza-Coutinho, M., Brasil, R., Souza, C., Sousa, P., & Malfeito-Ferreira, M. (2020). Consumers Associate High-Quality (Fine) Wines with Complexity, Persistence, and Unpleasant Emotional Responses. *Foods*, 9(4), 452. <https://doi.org/10.3390/foods9040452>
- Treinta, F. T., & Farias Filho, José Rodrigues; Sant'Anna, Annibal Parracho; Rabelo, L. M. (2014). Metodologia de pesquisa bibliográfica com a utilização de método multicritério de apoio à decisão. *Production*, 24(3). <https://doi.org/10.1590/S0103-65132014000300002>
- Waterhouse, A. (2020). *Whats in Wine?*
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3). [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)