# Highlights

**???Grape Quality and its Link to Regional Differences in the Australian Winegrowing Industry**

Author

- ???
- ???
- ????
- ????

# ???Grape Quality and its Link to Regional Differences in the Australian Winegrowing Industry

Author[1,1,1]

## Abstract

## 1. Introduction

The Australian wine-growing industry is a rich and diverse landscape that is separated into multiple Geographical Indicator Regions. Each region describing unique reputations, qualities and varietals of wine produced there. While a great deal has been done regarding individual regional properties and traits, there has been little statistical insight into broader regional comparisons; due to a lack of cross-regional and in-depth data sources (Keith Jones, 2002; Knight et al., 2019). In this study we use Classification Trees to compare regional differences and how these differences relate to sustainable practices.

A vineyard's region predetermines several physical parameters, such as: climate, geology and soil; making location a widely considered key determinant of grape yield and quality (Abbal et al., 2016; Agosta et al., 2012; Fraga et al., 2017). Through the use of classification trees this study aims to highlight the key differences in sustainable practices at a regional level and how these practices relate to the different grades of grape quality.

## 2. Methods

### 2.1. Data

The Australian wine industry is divided into 65 regions, known as a Geographical Indicator Regions (GI Region). Each GI Region is used to describe different unique localised traits of vineyards across Australia; with each having its own mixture of climatic and geophysical properties (Halliday, 2009; Oliver et al., 2013; SOAR et al., 2008). Each region is explicitly defined under the Wine Australia Corporation Act of 1980 (Attorney-General's Department, 2010). The climatic properties of a GI Region are summarised by Sustainable Winegrowing Australia (2021), where regions of similar climates are amalgamated together into superset regions. The climatic regions were utilised to illustrate similar trends and explain differences between sets of regions. The data used in this analysis comes from Sustainable Winegrowing Australia and covers the period 2015 to 2022. The dataset contained 3342 samples across 52 GI Regions and 1072 individual vineyards.

### 2.2. XGBoosted Trees

XGBoosted (eXtreme Gradient Boosting) trees were created using the XGBoost library (Chen and Guestrin, 2016) in the Python Programming language (G. van Rossum, 1995). They were chosen for this analysis as they provide a both high predictive performance and ability to effectively capture complex relationships. Following Chen and Guestrin (Chen and Guestrin, 2016), XGboosted trees use a given set of data, to predict $y_i$ from the input $x_i$. The method of prediction is achieved through a tree ensemble model, using $K$ additive functions to predict the output.

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_K(x_i), f_K \in \mathcal{F} \tag{1}$$

Each function $f_K$ is a classification or regression tree, such that all functions are in defined in the set $\mathcal{F}$ of trees given by $f(x) = \omega_{q(x)}(q : \mathbb{R}^m \rightarrow T, \omega \in \mathbb{R}^T)$. Where, $f_K$ corresponds to an independent tree structure $q$ of $\omega$ weights. Each tree has $T$ leaves, which contain a continuous score, represented by $\omega_i$ for the i-th leaf. The final prediction is determined by the sum of the score of the corresponding leaves, given by $\omega$. The set of functions used by the tree is determined by minimising the regularised objective function, given by:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i^{t-1} + f_t(x_i)) + \sum_k \Omega(f_K) \tag{2}$$

The difference between the prediction and actual variable is a convex loss function $l$. To optimise $l$, the difference is calculated for the i-th instance at the t-th iteration. The function $f_t$ is selected according to which value minimises 1. The model complexity is penalised by the function $\Omega$, this acts to smooth weights in an attempt to prevent over fitting. As predictions are made using additive tree functions, it can be used for both classification and regression. For this analysis both classification and regression trees were used. The major difference between the types of trees created was the objective function. As variables were both continuous, binary and multi-class, three different objective functions were used, root mean squared error, binary:logistic and the soft max functions respectively.

Chen and Guestrin (Chen and Guestrin, 2016) further illustrate, using Taylor expansions how, for a fixed structure $q(x)$ the the optimal weight

$\omega_j^*$ for a leaf $j$ can be derived. Furthermore they show how to successfully enumarate tree structures using left $I_L$ and right $I_R$ instance sets of nodes and letting $I = I_L \cup I_R$. The loss reduction after the split is given by the function:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \qquad (3)$$

This means we greedily add the ft that most improves our model according to Eq. (2). Second-order approximation can be used to quickly optimize the objective in the general setting [12].

A common example is a linear model, where the prediction is given as $y_i = \Sigma_j \theta_j x_{ij}$, a linear combination of weighted input features. The prediction value can have different interpretations, depending on the task, i.e., regression or classification. For example, it can be logistic transformed to get the probability of positive class in logistic regression, and it can also be used as a ranking score when we want to rank the outputs.

The parameters are the undetermined part that we need to learn from data. In linear regression problems, the parameters are the coefficients $\theta$. Usually we will use $\theta$ to denote the parameters (there are many parameters in a model, our definition here is sloppy).

With judicious choices for $y_i$, we may express a variety of tasks, such as regression, classification, and ranking. The task of training the model amounts to finding the best parameters that best fit the training data and labels

. In order to train the model, we need to define the objective function to measure how well the model fit the training data.

A salient characteristic of objective functions is that they consist of two parts: training loss and regularization term: $obj(\theta) = L(\theta) + \Omega(\theta)$

where $L$ is the training loss function, and $\Omega$ is the regularization term. The training loss measures how predictive our model is with respect to the training data. A common choice of $L$ is the mean squared error, which is given by

$L(\theta) = \sum_i (y_i - \hat{y}_i)^2$

Another commonly used loss function is logistic loss, to be used for logistic regression:

The regularization term is what people usually forget to add. The regularization term controls the complexity of the model, which helps us to avoid overfitting. This sounds a bit abstract, so let us consider the following problem in the following picture. You are asked to fit visually a step function given the input data points on the upper left corner of the image. Which solution among the three do you think is the best fit?

XGBoosted Regression trees were used to predict continuous variables. With data being split into 80% training data and 20% testing data.

XGBoosted classification trees were used to classify the binary and multiclass variables. Data was split into 80% training, 10% testing and 10% validation data.

The modelled relationships are able to be scrutinised by using techniques such as feature importance analysis. The use of the XGBoost library also incorporates regularisation techniques built into the software to mitigate overfitting and enhance model generalisation. The further use of cross validated grid search functions allowed for the selection of better performing hyperparameters when selecting the final model.

### 2.3. Classification Trees

Classification Trees were developed to discern the different practices within regions and climates, comparing these relationships to those linked to grape quality. This was done using the rparts and caret packages (Kuhn, 2008; Terry Therneau and Beth Atkinson, 2022) in the R statistical programming language (R Core Team, 2021).
Three classifications were undertaken for region, climate and grape quality. Climate was further classified into two subcategories of rainfall and temperature, resulting in a total of 5 classification trees being created. Classification trees were validated using K-fold cross validation.Each model was validated using 10 folds, utilising a random selection of different samples ten separate times to validate each of the classification trees. A summary confusion matrix was then constructed to show the class bias and overall accuracy of each tree.

## 3. Results

### 3.1. Model 1 GI Regions

The first Model was used to classify GI regions and resulted in an accuracy of 36.48% across 52 classes. The most prominent features used to classify regions were the types of water resources available (see Figure 1). Two regions, the Riverland and Coonawarra, were the most accurate classes being 92.74% and 96.97% respectively. These regions differ greatly in practice and geophysical properties, with the Riverland being a dry warm inland region and Coonawarra being a cooler, wet coastal region. However, they are both similar in operational scales, with vineyards being relatively large compared with other regions. The differences in resources and practices between these regions are also significant, such as the Riverland utilising the river Murray as a water source. Many of the regions had significantly lower reporting rates, resulting much poorer classification performance. The regions with the most samples performed the best (see Table 1). Notably bordering regions were routinely grouped together and misclassified as the same region, for example the two closest regions to Coonawarra, Padthaway and Wrattonbulley, were misclassified as Coonawarra even though they had 147 and 137 samples respectively. The same case was found for the Murray Darling, with 143 samples, it was misclassified as the Riverland. These misclassifications are likely due to the incredibly similar regional properties and close proximity these regions have with one another. Other misclassifications were most likely due to lower reporting rates with many regions being under represented.

### *3.2. Climate*

Classifying the SWA climatic categorisation of the given regions had better performance than the GI Regions, with 41.66% being classified correctly. These categories were divided into 12 climatic classifications with 3 and 4 separate subsets for rainfall and temperature respectively. The decision tree behaved similarly and over classified climates with higher response rates. The results posed an interesting similarity with grape quality classifier, being influenced predominantly by water and area. The use of fungicide to separate regions that were 'Very dry' and 'Damp' can be considered as indicative of the different practices required due to climatic pressure; fungicides being

Table 1: Classification accuracy of the most prominent GI Regions.

| | **Accuracy** | **Predicted** | **Actual** |
|---|---|---|---|
| **Adelaide Hills** | 30.45% | 95 | 312 |
| **Barossa Valley** | 51.00% | 205 | 402 |
| **Coonawarra** | 96.97% | 192 | 198 |
| **Langhorne Creek** | 22.84% | 53 | 232 |
| **Margaret River** | 78.82% | 201 | 255 |
| **McLaren Vale** | 52.89% | 128 | 242 |
| **Riverland** | 92.74% | 345 | |

more prominent in cooler regions with greater rainfall due to the higher risk of disease pressure (Reynolds, 2010). This could also potentially explain the use of contractor tractor use to discern differences in grape quality, where the lack of contractor use to prevent disease could have led to lowered quality of grapes.

*3.2.1. Rainfall*

The rainfall decision tree showed a greater use of fungicides sprays to discern between damp and very Dry as shown in Figure 4; with the accuracy improving to 62% but was unable to effectively discern between dry and very dry regions (see Table 3).

#### *3.2.2. Temperature*

The classification of GI Regions by their temperatures (see Figure 5) showed similarities to the other trees, with a heavy reliance on the types of water resources used as dominant predictors. The use of contractors was again used to differentiate between warm and cool regions, likely being due to disease pressure. The temperature classification tree was only a minor improvement over the regional classification tree, with an accuracy of 49.26% as shown in the confusion matrix (see Table 4).

### *3.3. Model 3 Grape Quality*

The classification of grape quality through its grade had an accuracy of 55.72% across 5 separate grades. There was a notable issue with the classification of B grade grapes when compared to A and C (see Table 2). The classification tree itself shows similarities to that of classifying regions in Model 1, with the type of water resource used being a prominent determiner. Although not surprising the number of contractor tractor passes is new deciding factor due disease and pests reducing the potential quality of a crop. The prevalence of contractor use is greater in regions such as the Barossa Valley and the McLaren Vale, this could be due to the difference in operational scales, with larger sites being more likely to have ownership of their own equipment for weeding and spraying due to the cost benefit.

## 4. Discussion

The difference between grape quality is most notable between warm inland regions and coastal regions such as the Riverland and Coonawarra, respectively. Grape quality is only described by a singular variable within this study, however in reality it is driven by market demand and subject to complex forces such as international market pressure, fire, pests and disease (Wine Australia, 2019, 2020, 2021, 2022; Winemakers' Federation of Australia, 2015, 2016, 2017, 2018) The decision trees were able to offer some insights into the factors that influence grape quality and regional contrasts that contribute to different qualities. The most prominent being what readily available resources of each region were, particular the types of water available. Heavy water consumption is often linked to the mass production of grapes, where lower quality grapes are targeted in a quantity over quality strategy. These types of business decisions are unfortunately obfuscated by lack of in-depth data regarding vineyard business plans. Notably the literature shows that there are many complex decisions to be made on the ground depending on many compounding factors that influence both quality and yield (Abad et al., 2021; Cortez et al., 2009; Hall et al., 2011; I. Goodwin, et al., 2009; Kasimati et al., 2022; Oliver et al., 2013; Srivastava and Sadistap, 2018)

. There are also further differences when comparing winegrowers to other agricultural industries as they are vertically integrated within the wine industry, tying them to secondary and tertiary industries, such as wine production, packaging, transport and sales. This results in unique issues, where on-the-ground choices are influenced by other wine industry's decisions, such as the use of sustainable practices in vineyards to sell in overseas markets; notably these interactions are further complicated by some winegrowers being totally integrated into wine companies, while others are not (Knight et al., 2019). It is incredibly difficult to attribute external business decisions to produced grape quality but it is important to acknowledge that some growers are contracted to produce grapes of a particular grade; it is difficult to know whether another consumer may have graded the grape quality differently paying more or less for the same grapes given the opportunity to purchase them. It is difficult to untangle the contributing factors to the success of winegrowers and the quality of grapes produced without further specifics of choices made through out a season (Leilei He et al., 2022).

## 5. Conclusion

The type and availability of water resources were a major contributing factor when classifying grape quality and region. This was seen in the two most accurately classified regions, Coonawarra and the Riverland, with the Riverland predominantly utilising river water. Furthermore, the study highlighted the influence of water use, fungicide application, and contractor use in differentiating grape quality, climate and region respectively. These models provide insight into the complex dynamics between regional characteristics, sustainable practices, and grape quality in the Australian winegrowing industry. It is important to acknowledge that grape quality is subject to external influences such as market demands and prior established business arrangements. Further in-depth data and understanding are necessary to fully grasp the nuances of decision-making and the interplay of factors impacting grape quality.

## References


Abad, J., Hermoso de Mendoza, I., Marín, D., Orcaray, L., Santesteban, L.G., 2021. Cover crops in viticulture. A systematic review (1): <br>Implications on soil characteristics and biodiversity in vineyard. OENO One 55, 295–312. doi:10.20870/oeno-one.2021.55.1.3599.

Abbal, P., Sablayrolles, J.M., Matzner-Lober, É., Boursiquot, J.M., Baudrit, C., Carbonneau, A., 2016. Decision Support System for Vine Growers Based on a Bayesian Network. Journal of agricultural, biological, and environmental statistics 21, 131–151. doi:10.1007/s13253-015-0233-2.

Agosta, E., Canziani, P., Cavagnaro, M., 2012. Regional climate variability impacts on the annual grape yield in Mendoza, Argentina. Journal of Applied Meteorology and Climatology 51, 993–1009.

Attorney-General's Department, 2010. Wine Australia Corporation Act 1980.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 785–794. doi:10.1145/2939672.2939785.

Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Using data mining for wine quality assessment, in: Discovery Science: 12th International Conference, DS 2009, Porto, Portugal, October 3-5, 2009 12, Springer. pp. 66–79.

Fraga, H., Costa, R., Santos, J.A., 2017. Multivariate clustering of viticultural terroirs in the Douro winemaking region. Ciência Téc. Vitiv. 32, 142–153.

G. van Rossum, 1995. Python tutorial, Technical Report CS-R9526. Centrum voor Wiskunde en Informatica (CWI),.

Hall, A., Lamb, D.W., Holzapfel, B.P., Louis, J.P., 2011. Within-season temporal variation in correlations between vineyard canopy and winegrape composition and yield. Precision Agriculture 12, 103–117.

Halliday, J.C.J.C., 2009. Australian Wine Encyclopedia. Hardie Grant Books, VIC.

I. Goodwin,, L. McClymont,, D. Lanyon, A. Zerihun, J. Hornbuckle, M. Gibberd, D. Mowat, D. Smith, M. Barnes, R. Correll, 2009. Managing soil and water to target quality and reduce environmental impact.

Kasimati, A., Espejo-García, B., Darra, N., Fountas, S., 2022. Predicting Grape Sugar Content under Quality Attributes Using Normalized Difference Vegetation Index Data and Automated Machine Learning. Sensors 22. doi:10.3390/s22093249.

Keith Jones, 2002. Australian Wine Industry Environment Strategy.

Knight, H., Megicks, P., Agarwal, S., Leenders, M., 2019. Firm resources and the development of environmental sustainability among small and medium-sized enterprises: Evidence from the Australian wine industry. Business Strategy and the Environment 28, 25–39. doi:10.1002/bse.2178.

Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software, Articles 28, 1–26. doi:10.18637/jss.v028.i05.

Leilei He, Wentai Fang, Guanao Zhao, Zhenchao Wu, Longsheng Fu, Rui Li, Yaqoob Majeed, Jaspreet Dhupia, 2022. Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods 195.

Oliver, D., Bramley, R., Riches, D., Porter, I., Edwards, J., 2013. Review: Soil physical and chemical properties as indicators of soil quality in Australian viticulture. Australian Journal of Grape and Wine Research 19, 129–139. doi:10.1111/ajgw.12016.

Reynolds, A.G., 2010. Managing Wine Quality : Viticulture and Wine Quality. Woodhead Publishing Series in Food Science, Technology and Nutrition ; v.1., Elsevier Science, Cambridge.

SOAR, C., SADRAS, V., PETRIE, P., 2008. Climate drivers of red wine quality in four contrasting Australian wine regions. Australian journal of grape and wine research 14, 78–90. doi:10.1111/j.1755-0238.2008.00011.x.

Srivastava, S., Sadistap, S., 2018. Non-destructive sensing methods for quality assessment of on-tree fruits: A review. Journal of Food Measurement and Characterization 12, 497–526.

Terry Therneau, Beth Atkinson, 2022. Rpart: Recursive Partitioning and Regression Trees.

Wine Australia, 2019. National Vintage Report 2019 .

Wine Australia, 2020. National Vintage Report 2020 .

Wine Australia, 2021. National Vintage Report 2021 .

Wine Australia, 2022. National Vintage Report 2022 .

Winemakers' Federation of Australia, 2015. National Vintage Report 2015 .

Winemakers' Federation of Australia, 2016. National Vintage Report 2016 .

Winemakers' Federation of Australia, 2017. National Vintage Report 2017 .

Winemakers' Federation of Australia, 2018. National Vintage Report 2018 .