

1 Highlights

2 **???Grape Quality and its Link to Regional Differences in the Aus-**
3 **tralian Winegrowing Industry**

4 Author

5 • ???

6 • ???

7 • ????

8 • ????

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

11 Author^{1,1,1}

12 **Abstract**

13 **1. Introduction**

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

23
24 Through the use of classification trees this study aims to highlight the key
25 differences in sustainable practices at a regional level and how these practices
26 relate to the different grades of grape quality.

2. Methods

2.1. Data

Data used in this analysis were obtained from Sustainable Winegrowing Australia. Australia’s national wine industry sustainability program, which aims to facilitate grape-growers and winemakers in demonstrating and improving their sustainability (SWA, 2022). Data recorded by the SWA is entered manually by winegrowers using a web based interface tool. A total of 6091 observations were collected from 2012/2013 to 2021/2022 financial years. 23 variables were used for each observation reflecting a vineyards account for the given year (see Table 2.1). The profit variable was additionally transformed and included as another separate variable, profitable; depicting whether a vineyard was profitable or not.

The data originally contained only two multiclass variables: year and region. Variables that measured the same metric from different sources (such as water collected from rivers versus water from dams) were converted into multiclass variables representing the source. The total amount used from these variables was retained as a separate variable. Occurrences of multiple sources were defined as separate classes.

The variable region represented one of the 65 Geographical Indicator Regions (GI Region) used to describe different unique localised traits of vineyards across Australia (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).

Table 1: Summary of variables used in the analysis. The recorded column indicate values that were either greater than zero or that were not missing.

Variable	Units	Recorded	Number of Classes
Water Used	Mega Litres	5846	
Diesel	Litres	5585	
Biodiesel	Litres	25	
LPG	Litres	958	
Herbicide Spray	Times per year	2026	
Year	Class	6091	10
Disease	Class	6091	2
Region	Class	6091	58
Solar	Kilowatt Hours	622	
Irrigation Type	Class	6091	20
Petrol	Litres	4309	
Slashing	Times per year	2290	
Yield	Tonnes	5935	
Irrigation Energy	Class	6091	16
Area Harvested	Hectares	6091	
Electricity	Kilowatt Hours	1015	
Insecticide Spray	Times per year	1092	
Fertiliser	Kilograms of Nitrogen	795	
Fungicide Spray	Times per year	2260	
Cover Crop	Class	6091	32
Water Type	Class	6091	39
Profit	AUD	³ 853	
Operating Costs	AUD	853	

50 2.2. XGBoosted Trees

51 XGBoosted (eXtreme Gradient Boosting) trees were created using the
 52 XGBoost library (Chen and Guestrin, 2016) in the Python Programming
 53 language (G. van Rossum, 1995). They were chosen for this analysis as they
 54 provide both a high predictive performance and ability to effectively capture
 55 complex relationships. An XGBoosted tree was created for each variable to
 56 show how they interacted. Each tree included all but the economic vari-
 57 ables (profit and operating cost), which were only included once as response
 58 variables.

59 Following Chen and Guestrin (Chen and Guestrin, 2016), XGboosted
 60 trees predict a value y_i from the input x_i . The method of prediction is
 61 achieved through a tree ensemble model, using K additive functions to pre-
 62 dict the output.

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

63 where each function f_K is a classification or regression tree, such that all
 64 functions are in the set of all decision trees \mathcal{F} , defined by $f(x) = \omega_{q(x)}(q :$
 65 $\mathbb{R}^m \rightarrow T, \omega \in \mathbb{R}^T)$. Where, f_K corresponds to an independent tree structure
 66 q of ω weights. Each tree has T leaves, which contain a continuous score,
 67 represented by ω_i for the i -th leaf. The final prediction is determined by the
 68 sum of the score of the corresponding leaves, given by ω . The set of func-
 69 tions used by the tree is determined by minimising the regularised objective
 70 function, given by:

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

71 The difference between the prediction and actual variable is a convex loss
 72 function l . To optimise l , the difference is calculated for the i -th instance
 73 at the t -th iteration. The function f_t is selected according to which value
 74 minimises (2). The model complexity is penalised by the function Ω , this
 75 acts to smooth weights in an attempt to prevent over fitting.

76 As predictions are made using additive tree functions, XGboosted trees
 77 can be used for classification and regression. Due to the mixture of continu-
 78 ous, binary and multiclass variables in this analysis, both classification and
 79 regression trees were created. The difference between the trees created for
 80 this analysis was the objective function used. XGBoosted regression trees
 81 were created for continuous variables, using the root-mean-square as the ob-
 82 jective function. Binary class variables utilised the logistic loss function as
 83 the objective. And, Multiclass variable used the soft max function. All objec-
 84 tive functions are defined within the SKlearn library (Buitinck et al., 2013),
 85 linked via an API to the XGBoost library (Chen and Guestrin, 2016).

86 Chen and Guestrin (Chen and Guestrin, 2016) further illustrate, using
 87 Taylor expansions, that for a fixed structure $q(x)$ the optimal weight ω_j^* for
 88 a leaf j can be derived. Furthermore, they show the loss reduction after the
 89 split is given by the function:

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

90 with the tree structure defined using left I_L and right I_R instance sets of

91 nodes, with $I = I_L \cup I_R$. Instead of enumerating all possible tree structures,
 92 a greedy algorithm iteratively adds branches to the tree minimising \mathcal{L}_{split}
 93 in (3). The frequency of a variable’s occurrence within a tree is directly
 94 attributed to the minimisation of the objective function (or loss) through
 95 the minimisation of \mathcal{L}_{split} .

96 The frequency of a variable appearing as a node within the ensemble was
 97 used as a measure of importance. This measure was chosen as it connected
 98 a variable to the minimisation of its associated objective function, trans-
 99 lating the value into a simple count metric. Creating XGBoosted trees for
 100 each variable allowed the use of importance to show how strongly variables
 101 were associated with each other. The importance of predictor variables to
 102 economic variables was illustrated through the use of Sankey diagrams con-
 103 structed using the Holoviews python library (Rudiger et al., 2020). Other
 104 variable’s interconnectedness was demonstrated through the use of a chord
 105 diagram also created using Holoviews.

106 Each variable utilised 80% of the data to train the XGBoost ensemble,
 107 with 20% reserved for testing and validation. Testing was done through the
 108 iterative minimisation of the respective objective function for the variables
 109 type. For continuous variables 20% was used as testing data, minimising the
 110 root-mean-square function. The final model was validated using repeated
 111 k-fold cross validation for 10 folds, repeated 10 times. R^2 scores were used
 112 to determine the best regression models during validation. For binary and
 113 multiclass variables data was split into 80% training, 10% testing and 10%
 114 validation data. Due to class disparity in multiclass variables (most promi-
 115 nently in region) data was stratified into each subset at the same ratio of class

116 occurrence. Validation was summarised through the accuracy, the proportion
117 of true negatives and positives.

118 The use of the XGBoost library incorporates regularisation techniques
119 built into the software to mitigate over-fitting and enhance model gener-
120 alisation. The further use of cross validated grid search functions allowed
121 for the selection of better performing hyperparameters when selecting the
122 final model. The performance measure for model selection was root-mean-
123 square error for continuous variables. The receiver operator characteristic's
124 area under the curve was used for category variables (Hanley and McNeil,
125 1982). Multiclass variables utilised the one verse one approach to minimise
126 sensitivity to class disparity (Ferri et al., 2009; Hand and Till, 2001).

127 *2.3. Classification and Regression Trees*

128 Classification and Regression Trees were created for region, year, profit
129 and operating cost. These models describe the partitions that are useful
130 in predicting these variables; giving insight into the trees that make up the
131 ensembles created by XGBoost. These trees were created using the rparts
132 and caret packages (Kuhn, 2008; Terry Therneau and Beth Atkinson, 2022)
133 in the R statistical programming language (R Core Team, 2021).

134 Decision trees were validated using K-fold cross validation. Each model
135 was validated using 10 folds, utilising a random selection of different samples
136 ten separate times to validate each of the decision trees. The same measure
137 of accuracy as the XGBoosted trees was used for comparison.

138 3. Results

139 3.1. Region

140 Region classification performed at 32.34% (3.67% standard deviation) and
141 56.82% accuracy (50.58% validation accuracy), for the classification tree and
142 XGBoosted ensemble respectively. The most prominent feature used to clas-
143 sify regions with the classification tree was water source (see Figure 1). This
144 differed from the variables that illustrated the greatest importance for the
145 XGBoosted ensemble (see Figure (2), with predictor variables being highly
146 interrelated in importance. Area, water, fuel and yield were more deter-
147 mining factors when predicting region using XGBoost. Although water and
148 diesel were two of the three most frequently occurring variables in predicting
149 region, they were not as connected to the other predictor variables as Yield
150 and area harvested were.

151 It is reasonable that regions, being subjected to different rainfalls and
152 temperatures, would require different amounts of water, and would have
153 access to different water sources. The relation of area harvested and fuel
154 (particularly petrol) is prominent with other predictors. Due to the wide
155 variety of uses of petrol and diesel, it is likely that they are representative of
156 other activities within the vineyard, such as pruning and harvesting. With
157 predictors such as yield and area being highly interconnected as they likely
158 operate as proxy variables to other factors, possibly other present variables.

159 Many of the regions had significantly lower reporting rates, resulting in
160 much poorer classification performance. The regions with the most samples
161 performed the best. Notably bordering regions were routinely grouped to-
162 gether and misclassified as the same region. Two areas that suffered the

163 most from this, specifically with the classification tree were the Limestone
164 Coast (cool coastal areas in South Australia) and the warmer inland regions
165 along the Murray Darling. The classification tree likely had more difficulty
166 discerning vineyards closer to the river using only water sources due to the
167 greater access to river water in these areas.

168 3.2. *Year*

169 The classification tree and XGBoosted ensemble performed similarly for
170 classifying year with 35.20% (6.28% standard deviation) and 51.81% (42.20%
171 validation accuracy) respectively. Electricity and the type of irrigation were
172 highly influential within the classification tree. Similarly, electricity was the
173 most frequently occurring node in the XGBoost ensemble. However, other
174 variables such as slashing passes, and fungicide and herbicide spraying were
175 more prevalent than in the classification tree. Weed and disease outbreaks
176 are likely an influential factor when classifying different years, making the
177 decisions to spray and slash unique factors that differ year to year. Climatic
178 differences between years are likely tied to the influence of yield and water
179 use.

180 Over half of the interrelated importance of the predictor variables is domi-
181 nated by area harvested, yield and slashing passes. Although all the predictor
182 variables are highly connected, their relative importance is not as prominent
183 as the three major variables. It is of particular note of the relative impor-
184 tance of slashing to area, fuel and yield; as these are not directly related
185 activities. The connection between slashing and spraying is that those who
186 do a set number of spraying or slashing passes tended to do that many passes
187 for all slashing and spraying activities.

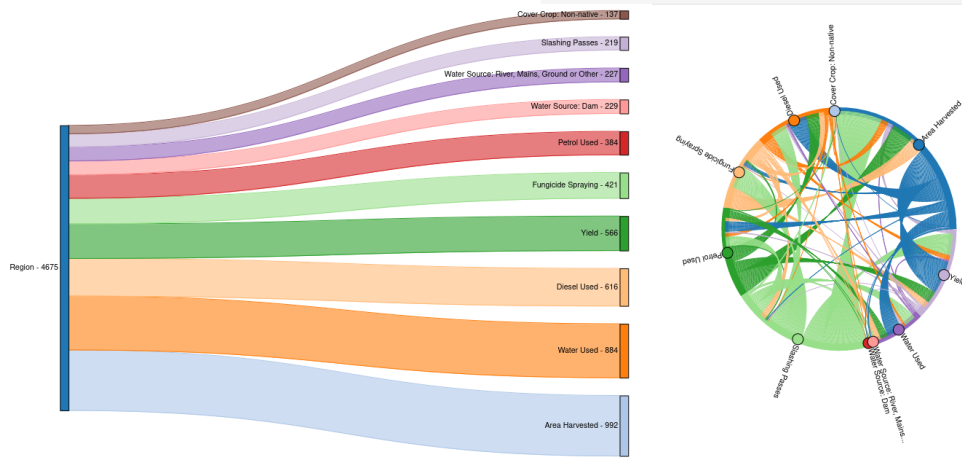


Figure 2: The left-hand side depicts the 10 most important variables in predicting Region using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

3.3. Operating Costs

There was a pronounced difference in accuracy between the regression tree and the XGBoost model when predicting Operating costs. With the regression tree achieving an R^2 of 0.0931 (with a standard deviation of 0.0197) in its cross validation. The XGBoosted regression ensemble achieved an R^2 of 0.8025 (with a standard deviation of 0.1033).

Within the XGBoost ensemble's nodes for operating costs (see figure 5) fuel, water, area and yield occurred the most, similarly to region. Both diesel and petrol were of more relative importance (being ranked higher) in operating costs than water was compared with region. It is surprising that electricity, slashing and spraying was not more prominent in operating costs. However, Figure 4 shows that electricity, slashing and spraying are

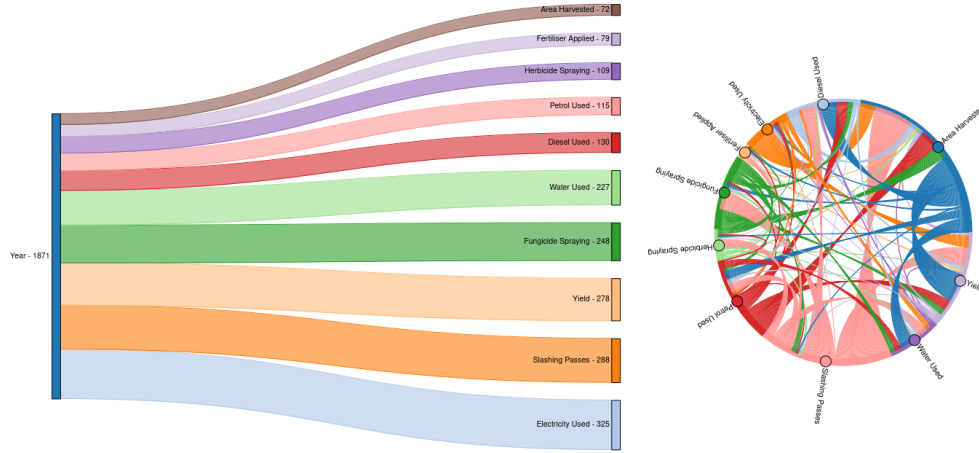


Figure 4: The left-hand side depicts the 10 most important variables in predicting Year using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

important variables in determining area and yield. Electricity in particular is used predominantly for irrigation and so is related largely to the size of a vineyard. However, slashing and spraying are measured in discrete tractor passes and show a surprising connection to the overall size of a vineyard, despite not being scaled to any measure of size. This would mean that, although measured as the same increment, a slashing or spraying pass in a larger vineyard would consume more fuel and wages than in a smaller vineyard.

3.4. Profit

Predictions of profit performed poorly compared to operating costs with the regression tree having an R^2 of 0.1873 (with a standard deviation of 0.0522) and the XGBoosted ensemble achieving an R^2 of 0.2535 (with a stan-

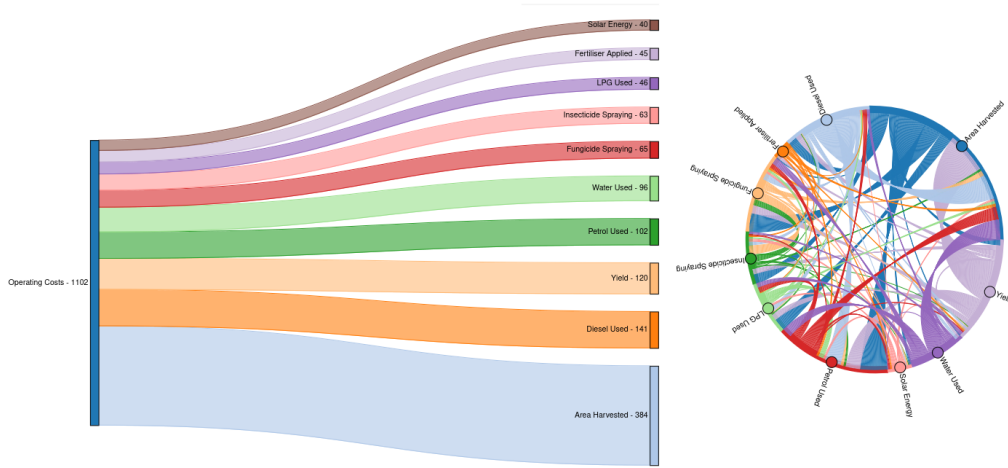


Figure 5: The left-hand side depicts the 10 most important variables in predicting Operating Costs using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

212 dard deviation of 0.3126). The high standard deviation in the XGBoosted
 213 tree was a bias in more accurately predicting vineyards that made profit compared to those that lost money. With much higher R^2 values being achieved
 214 in k-folds containing only those that made profits (recording a maximum of
 215 0.7634).
 216

217 There was a disparity of 66.63% of vineyards recording a profit than those
 218 that did not. When predicting if a vineyard would be profitable or not the
 219 classification tree and XGBoosted ensemble did not perform considerably differently from this proportion. With the regression tree achieving an accuracy
 220 of 68.66% (and a standard deviation of 0.01%) and the XGBoost ensemble
 221 achieving 71.97% accuracy (with a validation accuracy of 70.59%).
 222

223 It was surprising that operating costs performed substantially better in R^2

224 compared to profit. Interestingly the important variables when attempting
 225 to determine profit were similar to those used to classify region (see Figure
 226 7), with the exception of water used. Both the regression tree and the XG-
 227 Boosted ensemble used region, specifically the Hunter Valley. The regression
 228 tree also used Tasmania when determining profit. Both the Hunter valley
 229 and Tasmania are known for the production of high quality grapes used in
 230 export wines (Wine Australia, 2022). A major difference between region and
 231 profit was the importance given to water use, with water use being a more
 232 important variable in predicting region than profit.

233 **4. Discussion**

234 Several physical parameters such as climate, geology and soil are prede-
 235 termined by a vineyard’s location; making it a widely considered key deter-
 236 minant of grape yield and quality (Abbal et al., 2016; Agosta et al., 2012;
 237 Fraga et al., 2017). The association between yield and region is demonstrated
 238 by its position as fourth most occurring variable within the nodes of the XG-
 239 Boosted ensemble which determined region (see Figure 2). The association
 240 with area and region is likely a connection to the change in land costs, with
 241 inland Australian areas (particularly of lower rainfall) being substantially
 242 cheaper to buy than coastal regions, allowing larger areas to be purchased
 243 (Will Chancellor et al., 2019).

244 Regions with lower land costs are also warmer (Will Chancellor et al.,
 245 2019), which is known to be beneficial in hastening the ripening process of
 246 winegrapes (WEBB et al., 2011). Warmer regions are also associated with
 247 lower quality grapes, caused largely due to this hastened ripening (Botting

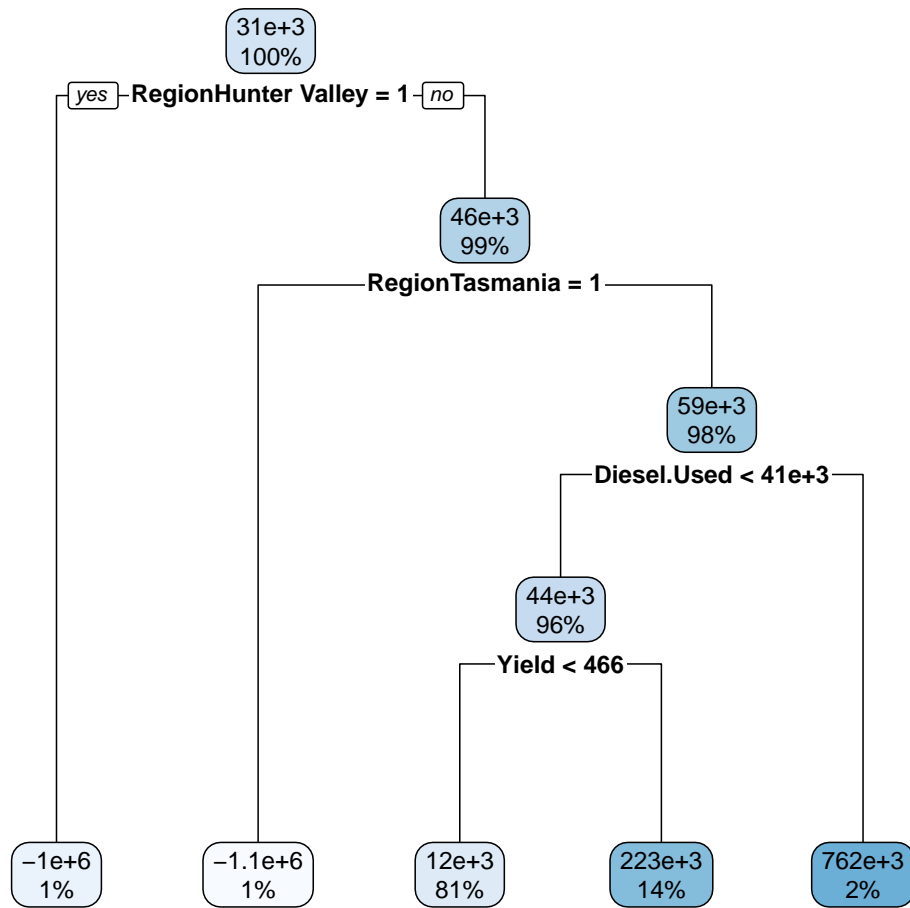


Figure 6: Decision tree predicting Profit. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

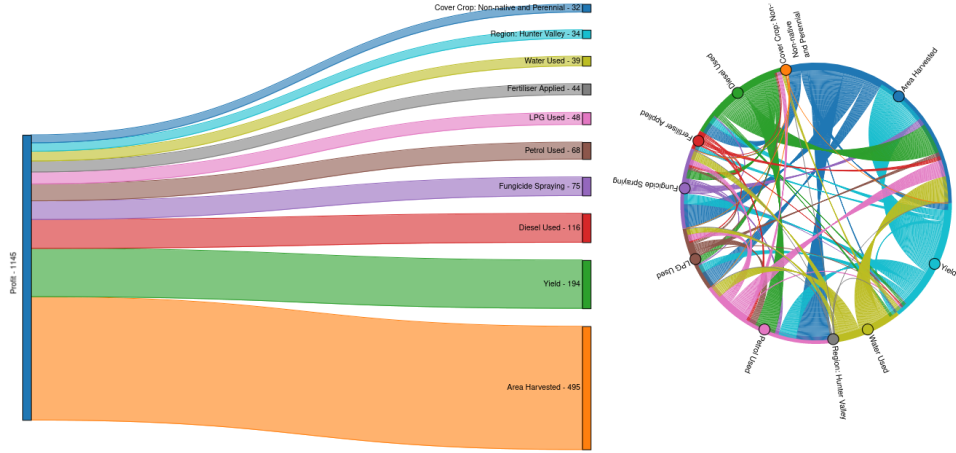


Figure 7: The left-hand side depicts the 10 most important variables in predicting Profit using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

et al., 1996). In general warmer regions have been associated with lower yields due to their generally lower rainfall, which can be mitigated through applying excess water (Camps and Ramos, 2012). It is likely that the combination of larger vineyards with higher water use is a determining factor in classifying regions which favour larger production of lower quality grapes; reflected through the variables' importance of water use in the XGBoost ensemble. The practice of utilising larger quantities of water for inland Australian wine crops is partly reflected in the prior use of flood style irrigation to saturate soil (BG Coombe and P Iland, 2004). This classification can be contrasted with other warmer regions of higher rainfall that use the warmer climate to concentrate their grapes, increasing the flavour profile (and thus quality) (Goodwin I, Jerie P, 1992; MG McCarthy et al., 1986). This is possibly the

260 connection between the presence of the Hunter Valley within the XGBoost
261 ensemble that determined profit (see Figure 7). With this connection reflect-
262 ing the restriction of possible strategies employable by winegrowers between
263 different regions.

264 In part some winegrowing strategies are restricted simply through access
265 to water resources, being reflected through the region classification tree (see
266 Figure 1). Regions are likely to have varying access to different water sources,
267 such as those along the River Murray being able to utilise river water for
268 crops unlike most coastal regions. Similarly, the connection between region
269 and fuel use is likely an indicator of the level of infrastructure within the
270 region. Where, the need to pressurise irrigation systems from river water or
271 to generate power would require larger amounts of diesel and petrol.

272 Operational costs showed similar importance across fuel, water and trac-
273 tor use. The dominating factor of area likely played a large part in deter-
274 mining how costly a tractor pass would be, or in defining the ratio of water
275 applied to the amount of vines. The node frequency was high for area but
276 much lower in general across the other variables, which could indicate the
277 need to be more circumstantial in determining operational costs. Although
278 it was attempted to capture the complexity between how variables interacted
279 when determining operational costs (see Figure 5), it is likely yet more com-
280 plicated still. An example of how interrelated operational costs can be, is the
281 optimisation of tractor passes being shown to reduce energy use in vineyards,
282 decreasing running costs, as well as reducing soil compaction (Capello et al.,
283 2019).

284 Decisions made on the ground have far-reaching effects and are difficult

285 to completely capture. Greater tractor use as a preventative measure for
286 occurrences such as disease, may incur higher operational costs but could be
287 critical in preventing long term losses. With factors such as erosion and soil
288 health being difficult to capture but also influenced by tractor use (Capello
289 et al., 2019, 2020). Although, performing well in R^2 , the ability to predict
290 operational costs is limited by the variables incorporated. Reductions in fuel,
291 water and tractor use are obvious methods to reduce operational costs but
292 not necessarily achievable decisions. Without fully capturing more granular
293 activities such as the specifics of what fuel was used for, it is hard to determine
294 what decisions specifically influence the operational costs.

295 Although less important in the XGBoost ensembles for profit, the vari-
296 ables: cover crops, fungicide spraying and slashing are likely linked to broad
297 environmental properties of regions (see Figure 2 and 7). Rainfall being re-
298 lated to fungal growth and disease, as well as weeds. With cover crops being
299 an effective and sustainable method to alleviate these issues (Delpuech and
300 Metay, 2018). It is difficult to extrapolate findings to these methods and the
301 reason for their use due to the broad and varying definition of the regions.
302 Utilising the Geographical Indicator regions defined by Wine Australia (Aus-
303 tralia, 2021) is a limitation, as it is too broad to fully capture a vineyards
304 location and its influence on more granular variables. The reasoning for us-
305 ing approaches such as cover crops can be widely varying. Where, a cover
306 crop may be employed to help increase soil water retention, reduce erosion,
307 increase biodiversity and reduce weeds (Capello et al., 2019, 2020; Delpuech
308 and Metay, 2018). However, cover crops can introduce competition with
309 grapevines and may reduce yield depending upon the plants used and the

density of the cover crop (Gosling and Shepherd, 2005; Monteiro and Lopes, 2007). A more granular definition of region may help to better discern the differences in practices, and the reason for employing them. More sophisticated models, specifically those that utilise expert opinion, may also help to capture and address the decision making process. An example is the optimisation of fungicide sprays using Bayesian models that forecast disease risk (Lu et al., 2020).

The disparity in accuracy between profit and operational costs is reflective of the complexity in trying to address challenges such as climate change, disease and changing market demands (Wine Australia, 2020, 2021, 2022). The difference between turning a profit or loss is dependent on decisions made and chance. The difference between vineyards that make profit and those that do not could be a multitude of factors including differences in farming practices not captured within this study. Some decisions leading to latent effects such as large scale soil deposition in extreme rain events can be caused by soil compaction due to overworking a vineyard (Capello et al., 2020).

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

334 provide insight into the complex dynamics between regional characteristics,
335 sustainable practices, and grape quality in the Australian winegrowing indus-
336 try. It is important to acknowledge that grape quality is subject to external
337 influences such as market demands and prior established business arrange-
338 ments. Further in-depth data and understanding are necessary to fully grasp
339 the nuances of decision-making and the interplay of factors impacting grape
340 quality.

341 **References**

- 342 Abbal, P., Sablayrolles, J.M., Matzner-Lober, É., Boursiquot, J.M., Baudrit,
343 C., Carbonneau, A., 2016. Decision Support System for Vine Growers
344 Based on a Bayesian Network. *Journal of agricultural, biological, and*
345 *environmental statistics* 21, 131–151. doi:10.1007/s13253-015-0233-2.
- 346 Agosta, E., Canziani, P., Cavagnaro, M., 2012. Regional climate variability
347 impacts on the annual grape yield in Mendoza, Argentina. *Journal of*
348 *Applied Meteorology and Climatology* 51, 993–1009.
- 349 Attorney-General’s Department, 2010. *Wine Australia Corporation Act*
350 1980.
- 351 Australia, W., 2021. *Wine Australia-Open Data*.
- 352 BG Coombe, P Iland, 2004. Grape Berry Development and Winegrape Qual-
353 ity. In: *Viticulture – Resources..* volume 1. 2 ed., Winetitles, Adelaide,
354 South Australia.

355 Botting, D., Dry, P., Iland, P., 1996. Canopy architecture-implications for
356 Shiraz grown in a hot, arid climate .

357 Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel,
358 O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R.,
359 VanderPlas, J., Joly, A., Holt, B., Varoquaux, G., 2013. API design for
360 machine learning software: Experiences from the scikit-learn project, in:
361 ECML PKDD Workshop: Languages for Data Mining and Machine Learn-
362 ing, pp. 108–122.

363 Camps, J.O., Ramos, M.C., 2012. Grape harvest and yield responses to inter-
364 annual changes in temperature and precipitation in an area of north-east
365 Spain with a Mediterranean climate. *International Journal of Biometeo-*
366 *rology* 56, 853–64. doi:10.1007/s00484-011-0489-3.

367 Capello, G., Biddoccu, M., Cavallo, E., 2020. Permanent cover for soil and
368 water conservation in mechanized vineyards: A study case in Piedmont,
369 NW Italy 15.

370 Capello, G., Biddoccu, M., Ferraris, S., Cavallo, E., 2019. Effects of Tractor
371 Passes on Hydrological and Soil Erosion Processes in Tilled and Grassed
372 Vineyards. *Water* 11. doi:10.3390/w11102118.

373 Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System,
374 in: *Proceedings of the 22nd ACM SIGKDD International Conference on*
375 *Knowledge Discovery and Data Mining*, ACM, New York, NY, USA. pp.
376 785–794. doi:10.1145/2939672.2939785.

- 377 Delpuech, X., Metay, A., 2018. Adapting cover crop soil coverage to soil depth
378 to limit competition for water in a Mediterranean vineyard. *European*
379 *Journal of Agronomy* 97, 60–69. doi:10.1016/j.eja.2018.04.013.
- 380 Ferri, C., Hernández-Orallo, J., Modroiu, R., 2009. An experimental com-
381 parison of performance measures for classification. *Pattern Recognition*
382 *Letters* 30, 27–38. doi:10.1016/j.patrec.2008.08.010.
- 383 Fraga, H., Costa, R., Santos, J.A., 2017. Multivariate clustering of viticul-
384 tural terroirs in the Douro winemaking region. *Ciência Téc. Vitiv.* 32,
385 142–153.
- 386 G. van Rossum, 1995. Python tutorial, Technical Report CS-R9526. Centrum
387 voor Wiskunde en Informatica (CWI),.
- 388 Goodwin I, Jerie P, 1992. Regulated deficit irrigation: Concept to prac-
389 tice. *Advances in vineyard irrigation. Australian and New Zealand Wine*
390 *Industry Journal* 7.
- 391 Gosling, P., Shepherd, M., 2005. Long-term changes in soil fertility in organic
392 arable farming systems in England, with particular reference to phosphorus
393 and potassium. *Agriculture, Ecosystems & Environment* 105, 425–432.
394 doi:10.1016/j.agee.2004.03.007.
- 395 Halliday, J.C.J.C., 2009. *Australian Wine Encyclopedia.* Hardie Grant
396 Books, VIC.
- 397 Hand, D.J., Till, R.J., 2001. A Simple Generalisation of the Area Under the
398 ROC Curve for Multiple Class Classification Problems. *Machine Learning*
399 45, 171–186. doi:10.1023/A:1010920819831.

400 Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a
 401 receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36.

402 Keith Jones, 2002. Australian Wine Industry Environment Strategy.

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

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

410 Lu, W., Newlands, N.K., Carisse, O., Atkinson, D.E., Cannon, A.J., 2020.
 411 Disease Risk Forecasting with Bayesian Learning Networks: Application
 412 to Grape Powdery Mildew (*Erysiphe necator*) in Vineyards. *Agronomy*
 413 (Basel) 10, 622. doi:10.3390/agronomy10050622.

414 MG McCarthy, RM Cirami, DG Furkaliev, 1986. The effect of crop load and
 415 vegetative growth control on wine quality. .

416 Monteiro, A., Lopes, C.M., 2007. Influence of cover crop on water use and
 417 performance of vineyard in Mediterranean Portugal. *Agriculture, Ecosys-*
 418 *tems & Environment* 121, 336–342. doi:10.1016/j.agee.2006.11.016.

419 Oliver, D., Bramley, R., Riches, D., Porter, I., Edwards, J., 2013. Review:
 420 Soil physical and chemical properties as indicators of soil quality in Aus-
 421 tralian viticulture. *Australian Journal of Grape and Wine Research* 19,
 422 129–139. doi:10.1111/ajgw.12016.

423 R Core Team, 2021. R: A Language and Environment for Statistical Com-
 424 puting. R Foundation for Statistical Computing.

425 Rudiger, P., Stevens, J.L., Bednar, J.A., Nijholt, B., Andrew, B, C., Randel-
 426 hoff, A., Mease, J., Tenner, V., maxalbert, Kaiser, M., ea42gh, Samuels, J.,
 427 stonebig, LB, F., Tolmie, A., Stephan, D., Lowe, S., Bampton, J., henri-
 428 queribeiro, Lustig, I., Signell, J., Bois, J., Talirz, L., Barth, L., Liquet, M.,
 429 Rachum, R., Langer, Y., arabidopsis, kbowen, 2020. Holoviz/holoviews:
 430 Version 1.13.3. Zenodo. doi:10.5281/zenodo.3904606.

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

434 SWA, S.W.A., 2022. Sustainable Wingrowing Australia.
 435 <https://sustainablewinegrowing.com.au/case-studies/>.

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

438 WEBB, L.B., WHETTON, P.H., BARLOW, E.W.R., 2011. Observed trends
 439 in winegrape maturity in Australia. Global change biology 17, 2707–2719.
 440 doi:10.1111/j.1365-2486.2011.02434.x.

441 Will Chancellor, Shiji Zhao, Lucy Randall, Kenton Lawson, Khanh Hoang,
 442 2019. Measuring Australian broadacre farmland value: Phase 1 - Statistical
 443 infrastructure.

444 Wine Australia, 2020. National Vintage Report 2020 .

⁴⁴⁵ Wine Australia, 2021. National Vintage Report 2021 .

⁴⁴⁶ Wine Australia, 2022. National Vintage Report 2022 .