

¹ Highlights

² **An analysis of interrelations between economic and environmental**
³ **variables in Australian Winegrowing.**

⁴ Author

⁵ • Highlight 1

⁶ • Highlight 2

⁷ • Highlight 3

⁸ • Highlight 4

31 through the incorporation of operating costs and grape revenue from grape
32 sales within the data. We use this data to study these economic outcomes
33 and their statistical relationships to vineyards’ utilisation of the resources.
34 Answering what the driving factors are behind vineyard economic outcomes,
35 and linking these outcomes to predictor importance. This is done through
36 analysing a new comprehensive nationwide data set using XGBoosted mod-
37 els. We further compare the relationships between different resources to
38 address the extensive collinearity found within the data (Chen and Guestrin,
39 2016). XGBoosted models were used because they are able to overcome
40 multicollinearity as well as highlight the level of importance that predictor
41 variables have on response variables; with importance being able to be sta-
42 tistically defined through multiple methods.

43 **2. Methods**

44 *2.1. Data*

45 Data used in this analysis were obtained from Sustainable Winegrowing
46 Australia. Australia’s national wine industry sustainability program. The
47 program aims to facilitate grape-growers and winemakers in demonstrating
48 and improving their sustainability (SWA, 2022). Data recorded by SWA is
49 entered manually by winegrowers using a web based interface tool. A total
50 of 6091 observations were collected from 2012/2013 to 2021/2022 financial
51 years. 23 variables were used for each observation reflecting a vineyards state
52 for the given year (see Table 2.1).

53 The data originally contained only two multiclass variables: year and re-
54 gion. Variables that measured the same metric from different sources (such as

Table 1: Summary of variables used in the analysis. The recorded column indicate the number of 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
Grape Revenue	AUD	³ 853	
Operating Costs	AUD	853	

55 water collected from rivers versus water from dams) were converted into mul-
56 ticlass variables representing the source through one-hot-encoding. Changing
57 each variable class into a binary value, with one indicating the presence of
58 the class and zero indicating its absence. Occurrences of multiple sources
59 were defined as their own separate classes. Where a class variable had a
60 recorded amount the total amount used from these variables was retained
61 as a separate variable; for example water used (in Mega Litres) was also
62 included alongside water source.

63 The variable region represented one of the 65 Geographical Indicator Re-
64 gions (GI Region) used to describe different unique localised traits of vine-
65 yards across Australia (Halliday, 2009; Oliver et al., 2013; SOAR et al., 2008).
66 Each region is explicitly defined under the Wine Australia Corporation Act
67 of 1980 (Attorney-General’s Department, 2010).

68 *2.2. XGBoosted Trees*

69 XGBoosted (eXtreme Gradient Boosting) trees were created using the
70 XGBoost library (Chen and Guestrin, 2016) in the Python Programming lan-
71 guage (G. van Rossum, 1995). XGBoosted trees are a boosted tree ensemble
72 method that can be used to classify classes, or predict continuous response
73 variables. They were chosen for this analysis as the data contained a mixture
74 of class and continuous variables. And, XGBoosted trees are unaffected by
75 multicollinearity, as well as offer high predictive performance for a wide vari-
76 ety of purposes (Chen and Guestrin, 2016). An XGBoosted tree was created
77 for each variable to show how they interacted. Each tree included all but the
78 economic variables (operating cost and revenue from grape sales), which were
79 only included within their own trees as response variables. Separately profit

80 (the difference between revenue and operational costs) was looked at in prior
 81 analyses (see appendix) but the results were not included due to low average
 82 loss values and model stability. This meant that every variable would have
 83 a measure of its importance to other variables (see Section 2.4), which was
 84 used to show the highly interrelated nature of variables within vineyards.
 85 The complicated interaction between variables was illustrated using Sankey
 86 and Chord diagrams; with variable importance measures being used to show
 87 the strength of connection between any two variables (see section 2.4).

88 Following Chen and Guestrin (Chen and Guestrin, 2016), XGboosted
 89 trees predict a value y_i from the input x_i . The method of prediction is
 90 achieved through a tree ensemble model, using K additive functions to pre-
 91 dict the output. Each of f_k functions is a classification or regression tree, such
 92 that all functions are in the set of all decision trees, given by \mathcal{F} , is defined
 93 by $f(x) = \omega_{q(x)}(q : \mathbb{R}^m \rightarrow T, \omega \in \mathbb{R}^T)$. Where each function corresponds to
 94 an independent tree structure q of ω weights. Each tree has T leaves, which
 95 contain a continuous score, represented by ω_i for the i -th leaf. The final
 96 prediction is determined by the sum of the score of the corresponding leaves,
 97 given by:

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

98 The set of functions, \mathcal{F} , used by the tree is determined by minimising a
 99 regularised objective function, \mathcal{L} given by:

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

100 , where

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (3)$$

101 As predictions are made using additive tree functions, XGboosted trees
 102 can be used for classification or regression. The difference between a predic-
 103 tion, $\phi(x_i)$, and actual variable, $f_k(x_i)$, is a differentiable convex loss function
 104 l . These properties of l allow the function to be versatile in which objective
 105 we choose to optimise for, which is also important in being able to process
 106 both continuous and categorical variables. To optimise l , the difference is
 107 calculated for the i -th instance at the t -th iteration.

108 2.3. Loss functions

109 The functions included as parameters in equation 2 mean that traditional
 110 optimisation methods for Euclidean space cannot be used. Chen and Guestrin
 111 (Chen and Guestrin, 2016) illustrate, using Taylor expansions, that for a fixed
 112 structure $q(x)$ the optimal weight ω_j^* for a leaf j can be derived. Importantly a
 113 loss function can be used to fit a model iteratively to data. For this analysis
 114 several loss functions were used, as variables took the form of continuous,
 115 binary and multi-call data. The loss function for making a split within the
 116 tree structure is given by:

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma. \quad (4)$$

117 The tree structure being defined using left I_L and right I_R instance sets of
 118 nodes, with $I = I_L \cup I_R$. Instead of enumerating all possible tree structures,
 119 a greedy algorithm iteratively adds branches to the tree minimising \mathcal{L}_{split}

120 in (4). The frequency of a variable’s occurrence within a tree is directly
121 attributed to the minimisation of the loss function through the minimisation
122 of \mathcal{L}_{split} .

123 The loss functions used for this analysis were the root-mean-square func-
124 tion for continuous variables. The logistic loss function for binary class vari-
125 ables. And, the soft max function for Multiclass variables. All objective
126 functions are defined within the SKlearn library (Buitinck et al., 2013), which
127 was utilised via an API to the XGBoost library (Chen and Guestrin, 2016).

128 *2.4. Variable Importance*

129 Due to XGBoost creating a large amount of decision trees, the inter-
130 pretability of these models is obfuscated by the intricate relationships within
131 complicated ensembles. A measure of variable importance was the technique
132 used to highlight a variables influence within the ensemble. Variable impor-
133 tance can be measured in multiple ways; we used the frequency of a variable
134 appearing as a node within the ensemble as a measure of its importance.
135 This measure was chosen as it connected a variable to the minimisation of
136 its associated objective function. The measure of a variable’s importance
137 within this study can then be interpreted as how often a variable was the
138 optimal choice in reducing the loss function of the ensemble. Importantly,
139 multiclass variables being one-hot-encoded are given an importance score
140 for each individual class; for example, each specific region will have its own
141 importance score.

142 Creating XGBoosted trees for each variable allowed the use of importance
143 to show how strongly variables were associated with each other. The impor-
144 tance of variables to one another was illustrated through the use of Sankey

145 and Chord diagrams. These diagrams were constructed using the Holoviews
146 python library (Rudiger et al., 2020). Both Chord and Sankey diagrams
147 illustrated variable importance through the size of the bands between two
148 variables. The number at the end of a connection in a Sankey diagram indi-
149 cated a variable’s importance, or the number of times it appeared within the
150 ensemble. Sankey and Chord diagrams were presented together; with Sankey
151 diagrams showing the connection of a variable to its 10 most important pre-
152 dictor variables. Chord diagrams were used alongside a Sankey diagram to
153 show the interconnectedness of the ten most prominent variables within its
154 associated Sankey diagram. Chord diagrams formed circles, with variables
155 being connected through their relative importance. The importance values
156 for the Chord diagrams were taken from the models of those individual
157 variables, with the diagram being simplified to just the ten variables in the
158 associated Sankey diagram, for readability’s sake.

159 2.5. Validation

160 As there were multiple different loss functions, multiple different forms
161 of validation were used. In each case the data was split into training data,
162 which constituted 80% of the original data. The remaining 20% was used
163 in testing and validation. Data was stratified when splitting the data into
164 these subsets to conserve the same proportion of class occurrences between
165 training, testing and validation data. For continuous variables 20% was used
166 as testing data, minimising the root-mean-square function. The final model
167 was validated using repeated k-fold cross validation for 10 folds, repeated 10
168 times. R^2 scores were used to determine the best regression models during
169 validation. For binary and multiclass variables, data was split into 80%

170 training, 10% testing and 10% validation data. For class variables, validation
171 was summarised through the accuracy, the proportion of true negatives and
172 positives.

173 The XGBoost library incorporates regularisation techniques built into
174 the software to mitigate over-fitting and enhance model generalisation. This
175 allowed us to utilise cross validated grid search functions when selecting for
176 better performing hyperparameters. The performance measure for model
177 selection was root-mean-square error for continuous variables. The receiver
178 operator characteristic’s area under the curve was used for category variables
179 (Hanley and McNeil, 1982). Multiclass variables utilised the one verse one
180 approach to minimise sensitivity to class disparity (Ferri et al., 2009; Hand
181 and Till, 2001).

182 *2.6. Surrogate Models*

183 The creation of more interpretable models such as linear regression in
184 parallel to AI systems has been used to explain variable’s relationships within
185 black box algorithms (Molnar, 2022). As XGBoost create an ensemble of
186 decision trees, here we use classification and regression trees to gain insight
187 into intricacies of the ensembles derived through XGBoost. Decision Trees
188 were created for operating costs, revenue and region. These models describe
189 the partitions that are useful in predicting these variables; giving insight into
190 the trees that make up the ensembles created by XGBoost. These trees were
191 created using the rparts and caret packages (Kuhn, 2008; Terry Therneau
192 and Beth Atkinson, 2022) in the R statistical programming language (R
193 Core Team, 2021).

194 Decision 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 decision trees. The same measure of accuracy as the XGBoosted trees was used for comparison.

3. Results

3.1. Revenue

We investigated the link between revenue to other variables in the SWA data by predicting it, and then linking each variable to revenue through variable importance. The prediction of revenue performed similarly to operating cost achieving an R^2 of 0.7716 (with a standard deviation of 0.1525). A regression tree was used as a surrogate model to present an example of the typical type of decision tree present within the XGBoost Ensemble, however the surrogate model only achieved an R^2 of 0.0961 (with a standard deviation of 0.0181) and the XGBoosted ensemble.

The important variables when attempting to determine revenue were size, yield, fuel and water (see .10). Due to regions being recorded separately for importance none appeared as the most important variables, overall regions contributed to 234 nodes in the ensemble making them collectively the third most important variable. Although performing poorly, the surrogate model highlights the importance of size in determining revenue. Area also appearing as a variable of higher importance is shown to be highly interrelated with other variables. The relation to area is likely to primarily be the effect of economies of scale, shown through its strong relations to other variables in figure .10. Area harvested is likely also an indicator of other variables such as slashing passes its strongest connection presented.

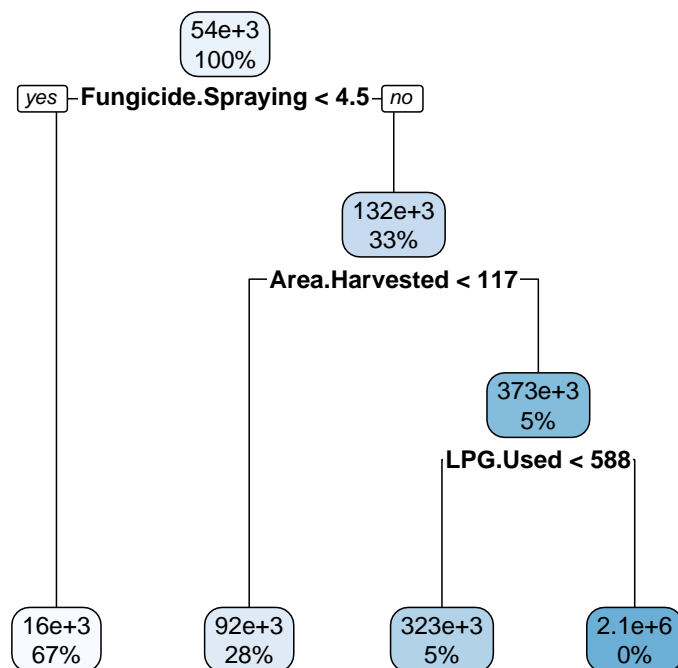


Figure 1: Decision tree predicting revenue. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

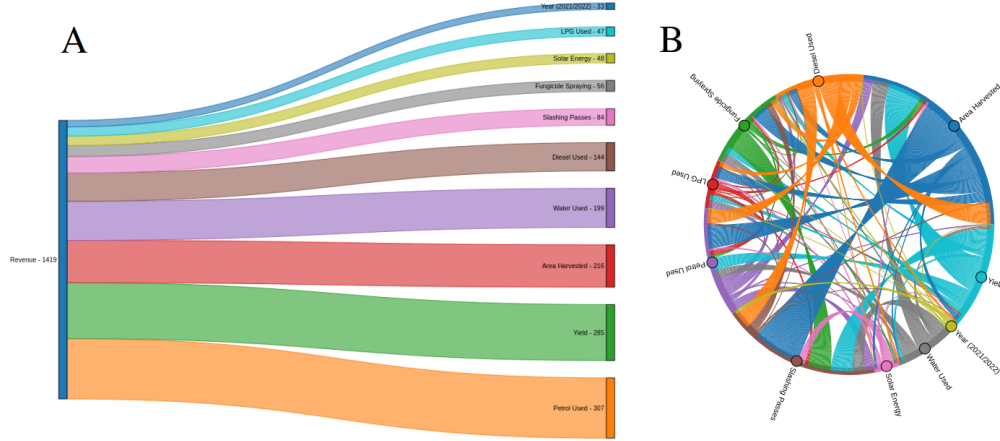


Figure 2: The left-hand side depicts the 10 most important variables in predicting revenue using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

219 3.2. Operating Costs

220 The relationships to operating cost through variable importance were
 221 found to be similar to that of revenue. With fuel, water, area and yield
 222 occurring the most (see figure 4). A surprising difference is that the most
 223 important operational consideration for operating cost is the use of fungicide,
 224 compared to revenue where slashing is the most important (comparing Figure
 225 6 and 4). The variables that feed into these decisions are also very differing
 226 with diesel being the most informative to slashing and area being the most
 227 informative to the need for fungicide.

228 Again, region played a determining factor overall but not as much indi-
 229 vidually with region contributing to 334 nodes within the ensemble making
 230 it the most important variable when considering all regions together. It

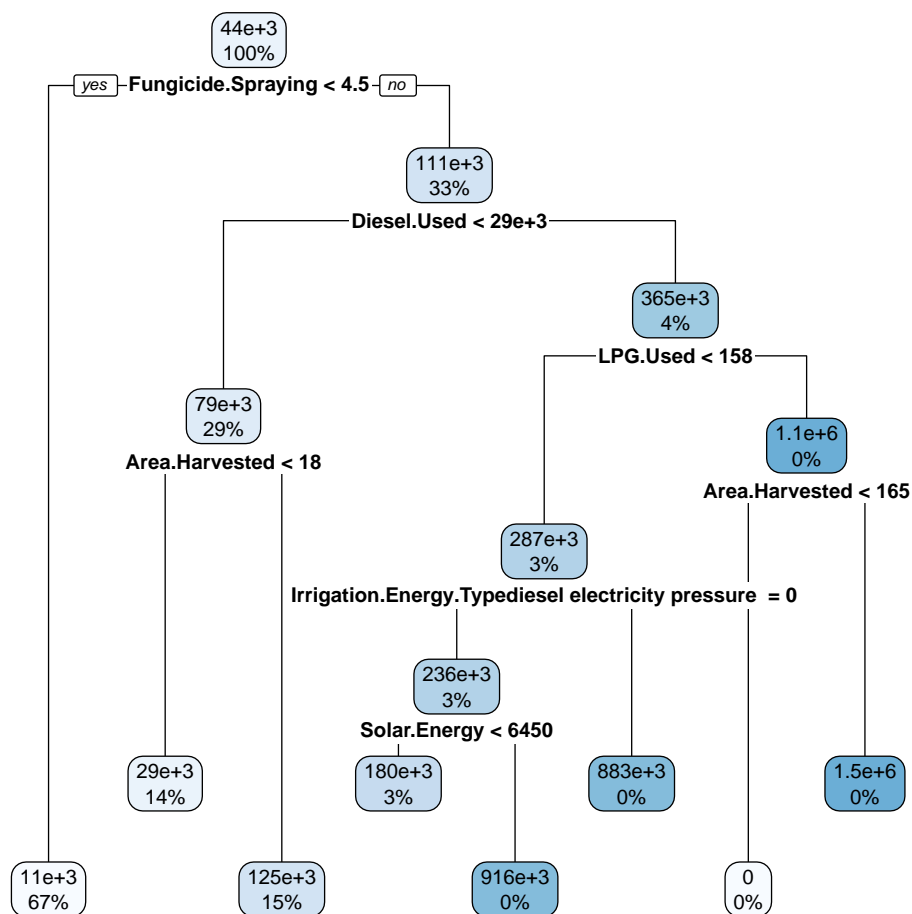


Figure 3: A surrogate model decision tree predicting operating costs. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

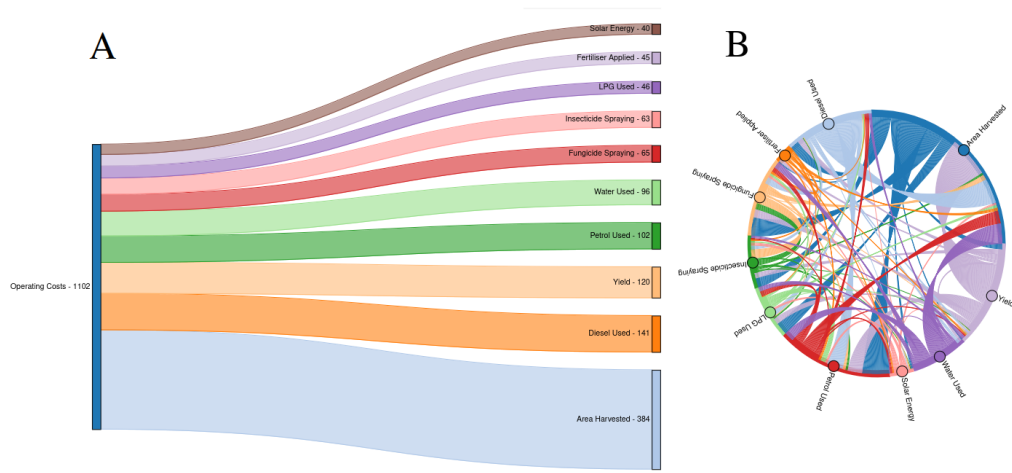


Figure 4: The left-hand side, A, depicts the 10 most important variables in predicting Operating Costs using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The number at the end of each band in the diagram is that variable's importance. The right-hand side, B, depicts the importance of the 10 variables in Sankey diagram relative to one another.

231 was surprising that electricity, slashing and spraying passes were not more
232 prominent in operating costs due to the intrinsic nature as an agricultural
233 expense.

234 Comparatively to revenue, operating cost performed better. The XG-
235 Boosted regression ensemble achieved an R^2 of 0.8025 (with a standard devi-
236 ation of 0.1033). Again the surrogate model did not perform well achieving
237 an R^2 of 0.0931 (with a standard deviation of 0.0197) but showed similarly
238 to revenue an importance placed on fungicide spraying and size (see figure
239 3).

240 3.3. Region

241 When considered overall, Region was a highly informative variable through
242 measure of importance to both operating cost and revenue. Notably the
243 overall importance of region to revenue was 234 (making it the third most
244 important variable when considering all regions together). The Barossa Val-
245 ley region and Tasmania were the two most important regions in relation to
246 revenue; these two regions are considered to be some of the highest revenue
247 per hectare regions in Australia (Wine Australia, 2022). These two regions
248 are also relative opposites in winegrowing climates with the Barossa being
249 warm and dry climate focussing on Shiraz grapes and Tasmania being a cool
250 wet climate that grows Pinot.

251 When considering all regions together, it had the most node contributions
252 to determining operating costs with an importance of 334. Of all the regions,
253 again Tasmania was the most important, followed by the Adelaide Hills. In
254 contrast to revenue, both climates are considered cool and wet, and warmer
255 drier regions such as the Barossa and Hunter Valley only contributed roughly

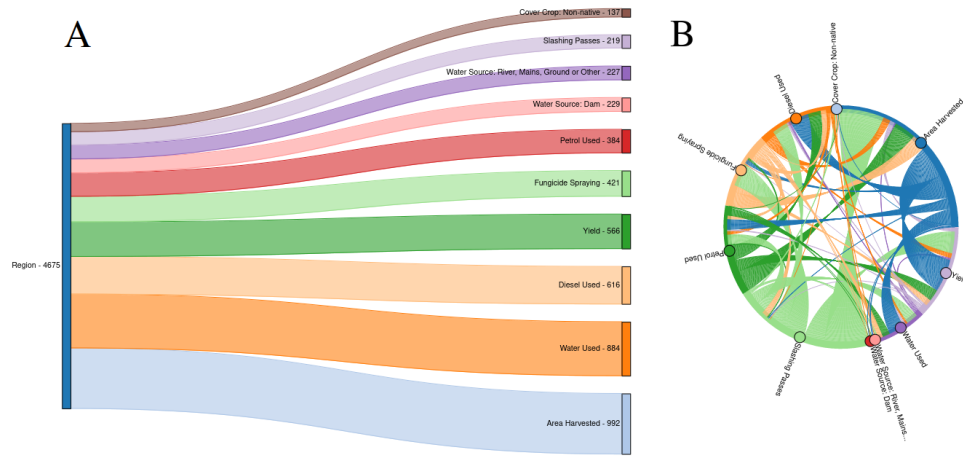


Figure 6: The left-hand side, A, depicts the 10 most important variables in predicting Region using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The number at the end of each band in the diagram is that variable's importance. The right-hand side, B, depicts the importance of the 10 variables in Sankey diagram relative to one another.

256 half the same number of nodes to the ensemble. When looking at 6 the
257 inclusion of slashing and fungicide spraying is the likely reason; with fungal
258 and weed pressure being greater in cooler wetter regions.

259 Both diesel and petrol were of more relative importance in operating costs
260 than water was compared with region. similar to those used to classify region
261 (see Figure 6, except water used. The surrogate model relied heavily on the
262 use of water source to classify regions, which is reflective of regional access
263 to resources being a deciding factor in vineyard management (see Figure 5).
264 A major difference between region and revenue was the importance given
265 to water use, with water use being a relatively more important variable in
266 predicting region than revenue (considering its rank in importance to other
267 variables).

268 The surrogate model for region performed better than other surrogate
269 models with 32.34% (3.67% standard deviation). The prominence of types
270 and use of water resources was in classifying region is reflective of difference
271 of availability of water resources is when comparing different regions (see
272 Figure 5). The XGBoost ensemble, did not perform as well as operating
273 costs or revenue with 56.82% accuracy (50.58% validation accuracy). The
274 difference in accuracy is in part due to the large number of classes (being
275 58). The ensemble did not differ greatly from the surrogate model, with a
276 continuing emphasis on Area, water, fuel and yield as determining factors
277 (see Figure (6).

278 Many of the regions had significantly lower reporting rates, resulting in
279 much poorer classification performance. The regions with the most samples
280 performed the best. Bordering regions were routinely grouped together and

281 misclassified as the same region. Two areas that suffered the most from
282 this, specifically with the classification tree were the Limestone Coast (cool
283 coastal areas in South Australia) and the warmer inland regions along the
284 Murray Darling. The classification tree likely had more difficulty discerning
285 vineyards closer to the river using only water sources due to the greater access
286 to river water in these areas.

287 **4. Discussion**

288 The explored relationships between vineyard resource use, operations and
289 geographical properties to revenue and operating costs highlight how deci-
290 sive regional influences can be determining a vineyard's economic outcomes.
291 Several physical parameters such as climate, geography and soil are predeter-
292 mined by a vineyard's location; making it a widely considered key determi-
293 nant of grape yield and quality (Abbal et al., 2016; Agosta et al., 2012; Fraga
294 et al., 2017). The association between yield and region is demonstrated by
295 its rank of fourth-highest variable importance when determining region (see
296 Figure 6).

297 Warmer regions are known to be beneficial in hastening the ripening pro-
298 cess of winegrapes (WEBB et al., 2011). Warmer regions are also associated
299 with lower quality grapes, caused largely due to this hastened ripening (Bot-
300 ting et al., 1996). In general warmer regions are not associated with higher
301 yields, but if a vineyard in a warmer region is sufficiently irrigated much
302 higher yields can be achieved than in cooler regions (Camps and Ramos,
303 2012). It is likely that the combination of larger vineyards with higher water
304 use is a determining factor in classifying regions which favour larger produc-

tion of grapes; reflected through region using water use so prominently in the XGBoost ensemble. The link to water resources in defining regions is also an important consideration, as vineyards can leverage higher irrigation rates given more accessible water resources. A further consideration in the link between revenue and region is that grape prices are set at a regional level by buyers (Wine Australia, 2022). It is also important to consider that some regions carry particular fame regarding the quality of their produce such as Tasmania, the Hunter Valley and Barossa Valley (Halliday, 2009). 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).

In part some winegrowing strategies are restricted simply through access to water resources, being reflected through the region classification tree (see Figure 5). Regions are likely to have varying access to different water sources, such as those along the River Murray being able to utilise river water for crops, unlike most coastal regions which may be drawing from surface or underground water sources. Similarly, the connection between region and fuel use is likely an indicator of the level of infrastructure within the region. Where, the need to pressurise irrigation systems from river water or to generate power would require larger amounts of diesel and petrol.

Operational costs showed similar importance across fuel, water and tractor use. The dominating factor of area likely played a large part in determining how costly a tractor pass would be, or in defining the ratio of water applied to the amount of vines. The node frequency was high for area but

330 much lower in general across the other variables, which could indicate the
331 need to be more circumstantial in determining operational costs. Although
332 it was attempted to capture the complexity between how variables inter-
333 acted when determining operational costs (see Figure 4), it is likely yet more
334 complicated. An example of how interrelated operational costs can be, is
335 the optimisation of tractor passes to achieve multiple goals in a pass, being
336 shown to reduce energy use in vineyards, decreasing running costs, as well
337 as reducing soil compaction (Capello et al., 2019).

338 When determining revenue, similar variables were used to operational
339 cost; with region also being of high variable importance relative to other vari-
340 ables (when considering all regions together in importance). It is difficult to
341 extrapolate the specific influence of location on a vineyard’s outcomes due to
342 the broad and varying definition of a region. Utilising the Geographical Indi-
343 cator regions defined by Wine Australia (Australia, 2021b) is a limitation in
344 one way, as it is too broad to fully capture a vineyards location and how that
345 influences variables at a more granular level. However, as buyers set prices
346 at regional levels, it is still important to consider a vineyards Geographical
347 Indicator region.

348 Decisions made on the ground have far-reaching effects and are difficult
349 to completely capture. A higher number of tractor passes used as a preven-
350 tative measure for occurrences such as disease, may incur higher operational
351 costs but could be critical in preventing long term losses. With factors such
352 as erosion and soil health being difficult to capture but also influenced by
353 tractor use (Capello et al., 2019, 2020). Although, performing well in R^2 ,
354 the ability to predict operational costs is limited by the variables incorpo-

355 rated. Reductions in fuel, water and tractor use are obvious methods to
356 reduce operational costs but not necessarily achievable decisions. Without
357 fully capturing more granular activities such as the specifics of what fuel was
358 used for, it is hard to determine what decisions specifically influence the op-
359 erational costs. Electricity in particular is used predominantly for irrigation.
360 Size is also a further consideration where slashing and spraying are measured
361 in discrete tractor passes and show a surprising connection to the overall size
362 of a vineyard, despite not being scaled to any measure of size. This would
363 mean that, although measured as the same increment, a slashing or spray-
364 ing pass in a larger vineyard would consume more fuel and wages than in a
365 smaller vineyard.

366 The reasoning for any particular decision can be widely varying. A more
367 granular definition of region may help to better discern the differences in
368 practices, and the reason for employing them. More sophisticated mod-
369 els, specifically those that utilise expert opinion, may also help to capture
370 and address the decision-making process. An example is the optimisation of
371 fungicide sprays using Bayesian models that forecast disease risk (Lu et al.,
372 2020).

373 Separately revenue and operating cost did have a greater predictability
374 than their counterpart profit. The disparity in accuracy between profit and
375 other economic outcomes is reflective of the complexity in trying to address
376 challenges such as climate change, disease and changing market demands
377 (Wine Australia, 2020, 2021, 2022). The difference between turning a profit
378 or loss is dependent on decisions made and chance. The difference between
379 vineyards that make profit and those that do not could be a multitude of fac-

380 tors including differences in farming practices not captured within this study.
381 Some decisions leading to latent effects such as large scale soil deposition in
382 extreme rain events can be caused by soil compaction due to overworking a
383 vineyard (Capello et al., 2020).

384 5. Conclusion

385 This study has provided valuable insights into the multifaceted dynam-
386 ics governing operational costs and revenue. The impact of different regions
387 highlighted the complex interrelatedness of variables within a vineyard. We
388 relate how factors such as water and fuel intersect to impact operational costs
389 and how different seasonal events affect these operations; as well as the signif-
390 icance of context-specific decision-making. While this investigation utilised
391 a broad regional classification, the potential benefits of adopting a more nu-
392 anced approach and incorporating expert knowledge have been highlighted.
393 Further work could pursue causal models and the creation of decision sup-
394 port systems. It is difficult to untangle the predictive and correlative nature
395 of a variable compared to the causal reasons. By delving deeper into the
396 complex interplay of variables, further advancements can be made in opti-
397 mising vineyard management strategies for lowering operational costs and
398 enhancing sustainability.

399 References

400 Abbal, P., Sablayrolles, J.M., Matzner-Lober, É., Boursiquot, J.M., Baudrit,
401 C., Carbonneau, A., 2016. Decision Support System for Vine Growers

402 Based on a Bayesian Network. *Journal of agricultural, biological, and*
 403 *environmental statistics* 21, 131–151. doi:10.1007/s13253-015-0233-2.

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

407 Attorney-General’s Department, 2010. *Wine Australia Corporation Act*
 408 1980.

409 Australia, W., 2021a. *Australian Wine: Production, Sales and Inventory*
 410 2019–20.

411 Australia, W., 2021b. *Wine Australia-Open Data*.

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

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

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

- 424 Capello, G., Biddoccu, M., Cavallo, E., 2020. Permanent cover for soil and
425 water conservation in mechanized vineyards: A study case in Piedmont,
426 NW Italy 15.
- 427 Capello, G., Biddoccu, M., Ferraris, S., Cavallo, E., 2019. Effects of Tractor
428 Passes on Hydrological and Soil Erosion Processes in Tilled and Grassed
429 Vineyards. *Water* 11. doi:10.3390/w11102118.
- 430 Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System,
431 in: *Proceedings of the 22nd ACM SIGKDD International Conference on*
432 *Knowledge Discovery and Data Mining*, ACM, New York, NY, USA. pp.
433 785–794. doi:10.1145/2939672.2939785.
- 434 Ferri, C., Hernández-Orallo, J., Modroiu, R., 2009. An experimental com-
435 parison of performance measures for classification. *Pattern Recognition*
436 *Letters* 30, 27–38. doi:10.1016/j.patrec.2008.08.010.
- 437 Fraga, H., Costa, R., Santos, J.A., 2017. Multivariate clustering of viticul-
438 tural terroirs in the Douro winemaking region. *Ciência Téc. Vitiv.* 32,
439 142–153.
- 440 G. van Rossum, 1995. Python tutorial, Technical Report CS-R9526. Centrum
441 voor Wiskunde en Informatica (CWI),.
- 442 Goodwin I, Jerie P, 1992. Regulated deficit irrigation: Concept to prac-
443 tice. *Advances in vineyard irrigation. Australian and New Zealand Wine*
444 *Industry Journal* 7.
- 445 Halliday, J.C.J.C., 2009. *Australian Wine Encyclopedia*. Hardie Grant
446 Books, VIC.

- 447 Hand, D.J., Till, R.J., 2001. A Simple Generalisation of the Area Under the
448 ROC Curve for Multiple Class Classification Problems. *Machine Learning*
449 45, 171–186. doi:10.1023/A:1010920819831.
- 450 Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a
451 receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36.
- 452 Kuhn, M., 2008. Building Predictive Models in R Using the
453 caret Package. *Journal of Statistical Software, Articles* 28, 1–26.
454 doi:10.18637/jss.v028.i05.
- 455 Lu, W., Newlands, N.K., Carisse, O., Atkinson, D.E., Cannon, A.J., 2020.
456 Disease Risk Forecasting with Bayesian Learning Networks: Application
457 to Grape Powdery Mildew (*Erysiphe necator*) in Vineyards. *Agronomy*
458 (Basel) 10, 622. doi:10.3390/agronomy10050622.
- 459 Luke Mancini, 2020. Understanding the Australian Wine Industry: A growers
460 guide to the background and participants of the wine grape industry.
- 461 MG McCarthy, RM Ciriaco, DG Furkaliev, 1986. The effect of crop load and
462 vegetative growth control on wine quality. .
- 463 Molnar, C., 2022. Interpretable Machine Learning: A Guide for Making
464 Black Box Models Explainable. 2 ed.
- 465 OECD, 2019. Innovation, Productivity and Sustainability in Food and Agri-
466 culture.
- 467 Oliver, D., Bramley, R., Riches, D., Porter, I., Edwards, J., 2013. Review:
468 Soil physical and chemical properties as indicators of soil quality in Aus-

469 tralian viticulture. Australian Journal of Grape and Wine Research 19,
470 129–139. doi:10.1111/ajgw.12016.

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

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

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

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

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

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

489 Wine Australia, 2020. National Vintage Report 2020 .

490 Wine Australia, 2021. National Vintage Report 2021 .

491 Wine Australia, 2022. National Vintage Report 2022 .

492 *Appendix .1. Year*

493 The classification tree and XGBoosted ensemble performed similarly for
494 classifying year with 35.20% (6.28% standard deviation) and 51.81% (42.20%
495 validation accuracy) respectively. Electricity and the type of irrigation were
496 highly influential within the classification tree. Similarly, electricity was the
497 most frequently occurring node in the XGBoost ensemble. Other variables
498 such as slashing passes, and fungicide and herbicide spraying were more
499 prevalent than in the classification tree. Weed and disease outbreaks are
500 likely an influential factor when classifying different years, making the de-
501 cisions to spray and slash unique factors that differ year to year. Climatic
502 differences between years are likely tied to the influence of yield and water
503 use.

504 Over half of the interrelated importance of the predictor variables is domi-
505 nated by area harvested, yield and slashing passes. Although all the predictor
506 variables are highly connected, their relative importance is not as prominent
507 as the three major variables. It is of particular note of the relative importance
508 of slashing passes to area, fuel and yield; as these are not directly related ac-
509 tivities. The connection between the number of slashing and spraying passes
510 is that those who do a set number of spraying or slashing passes tended to
511 do that many passes for all slashing and spraying activities.

512 *Appendix .2. Profit*

513 Predictions of profit performed poorly compared to operating cost and
514 revenue with an average R^2 of 0.2535 and standard deviation of 0.3126. With

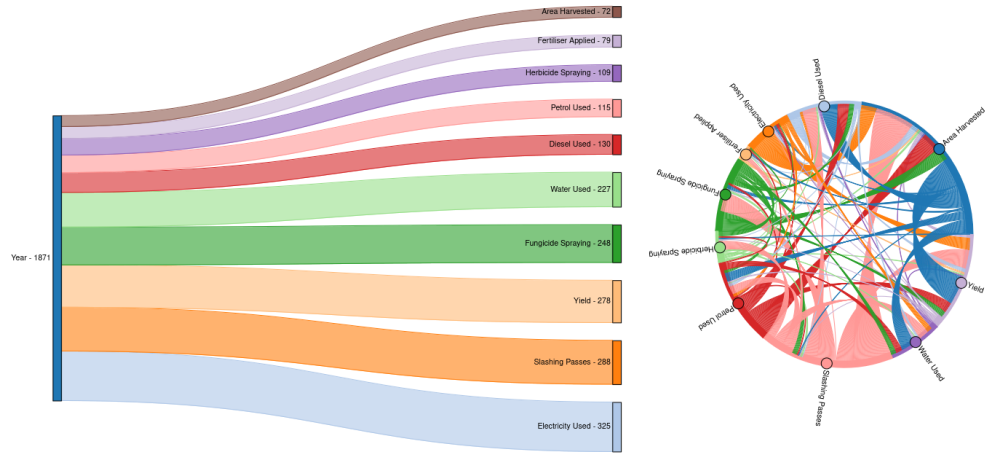


Figure .8: The left-hand side depicts the 10 most important variables in predicting Year using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

515 the large standard deviation being indicative of how unstable the models
 516 created were.

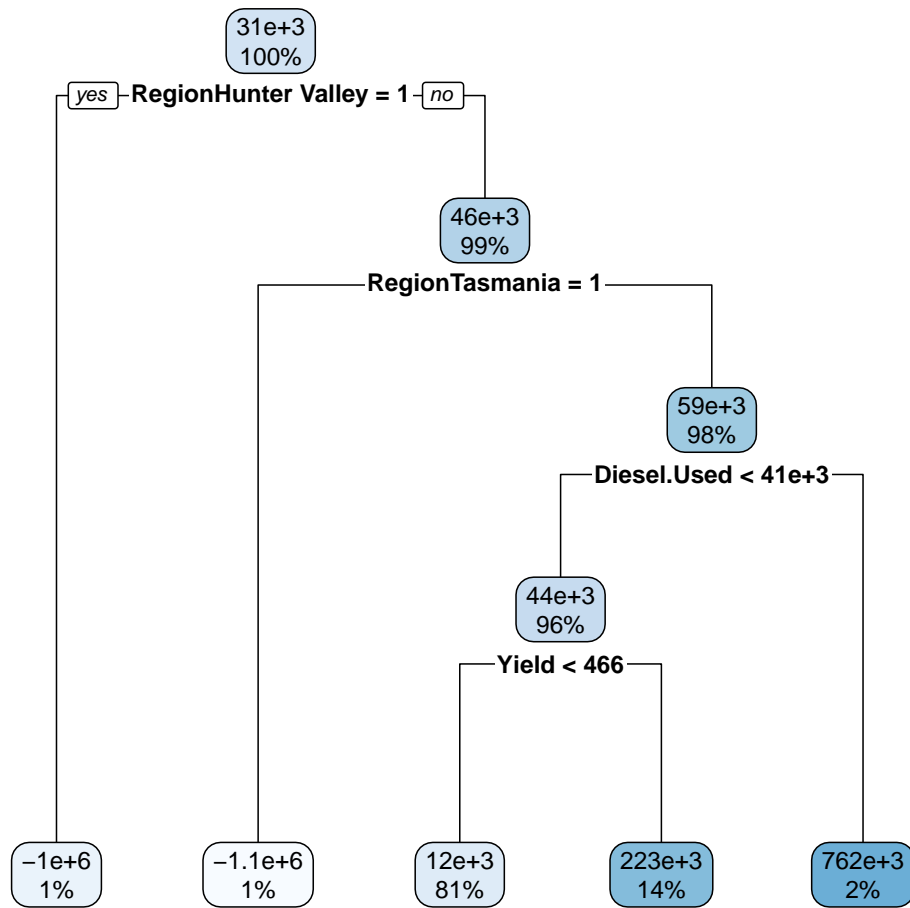


Figure .9: Decision tree predicting revenue. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

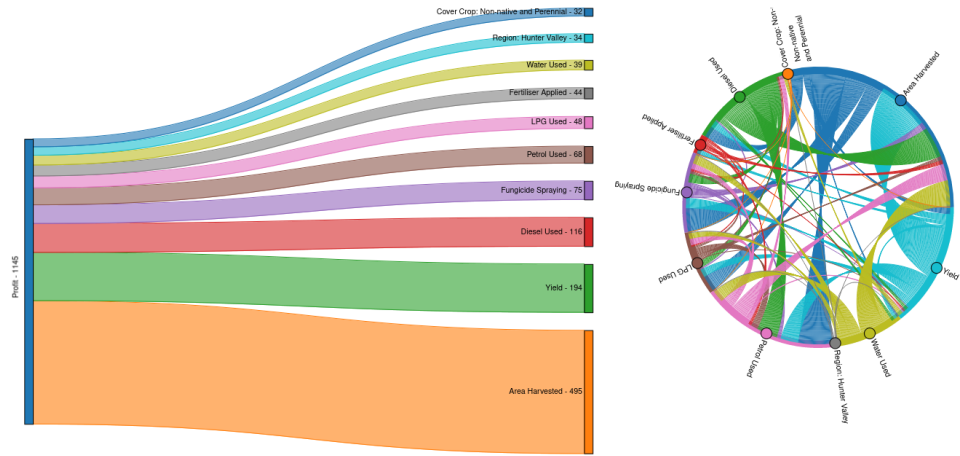


Figure .10: The left-hand side depicts the 10 most important variables in predicting revenue using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.