Assessing the Impact of Smoke-Exposure on Wine Grapes: Applying Data Reduction and Data Management Tools

DATA 501 Graduate Student Project
Saturday, April 8, 2017

Background

The exposure of wine grapes (*Vitis vinifera* [*V. vinifera*]) to smoke from wildland fire or prescribed burns changes the sensory profile of the berry (*i.e.*, the grape). More specifically, wine made from smoke-exposed berries shows an increased incidence of 'smoky', 'ashy', 'burnt meat' and 'Band-Aid' sensory attributes, all of which are undesirable in a quality product.[1–4] Chemically these negative sensory descriptors are associated with a specific class of compounds called volatile phenols (VP). This phenomenon is particularly problematic for the wine industry in the Okanagan Valley given the frequent occurrence of wildland fires during the growing season. However, it is also important for the global wine industry, as many key growing regions are also located near fire-prone regions. For instance, recent reports suggest that the economic impact of wildland fire on the Australian wine industry during the 2009 growing season was $299 million.[5] It is expected that this issue will increase in relevancy, as climate change models are suggesting an increase in the frequency of wildland fires in key wine growing regions (*e.g.*, California, British Columbia, Australia).[5]

Lignin, which accounts for 20-30% of the dry weight of wood, leads to the formation of a variety of VP during combustion. Many of these combustion products are known to correlate with the negative sensory descriptors associated with smoke-exposed berries (Figure 1). However, a subset of VP may also be present endogenously in the berry, where they are found in free (aglycone) and sugar-bound forms (glycosides), with the concentration of the glycosides typically much higher than the aglycones. Adding to the complexity of this problem is the fact that phenolic glycosides (VP-glycosides) may be enzymatically or chemically hydrolyzed during fermentation and aging. As such, despite possessing no sensory properties, VP-glycosides represent a 'sensory potential' that can influence the sensory profile of wine, even years after bottling. Existing methods (using VP and their glycosides) to quantify the risk associated with using smoke-exposed berries are only 50 – 80% predictive of negative sensory attributes in wine, leaving vineyards and wine producers at considerable financial risk.[6] My research aims to obtain a detailed assessment of the chemical composition of smoke-exposed berries (including and beyond VP and their glycosides), which will facilitate the development of a more accurate model for predicting wine quality issues, as well as inform remedial and preventative strategies.

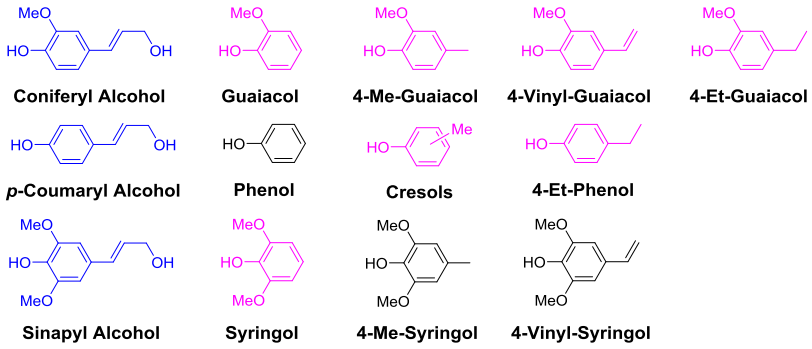


Figure 1: The core units of lignin (blue) and examples of combustion products with demonstrated relevance to the negative sensory properties of smoke-exposed *V. vinifera* berries. Nine of these VP (purple), including three cresols, were assessed quantitatively (*vide infra*).

Experimental Design and Sample Analysis

To assess the impact of smoke on the chemical composition of *V. vinifera* berries, a series of controlled field experiments were conducted. Using a custom-built enclosure that housed nine vines, four commercial varietals (Merlot, Pinot Noir, Cabernet Sauvignon and Cabernet Franc) were exposed to simulated wildland fire smoke (Figure 2). To ensure equal exposure for all vines, only the middle five vines were sampled as 'smoked' berries. A separate block of five vines per varietal were used as a control condition (*i.e.*, no smoke exposure). For each condition (smoked *versus* control) and each vine (5/condition) a series of time points were collected from immediately preceding smoke-exposure through until commercial maturity (Figure 2). Each sample was processed as whole berry homogenate (HMG) and free-run juice (FRJ) to mimic the raw materials for red and white wine production, respectively. Finally, a subset of time-points for some varietals were split into two fractions that were either washed or unwashed before processing as HMG and FRJ.

Cultivar	# Samples	# Time Points	Total Samples
Merlot	60	6	180
Cabernet Sauvignon	60	6	120
Pinot Noir	50	5	100
Cabernet Franc	50	5	160

Figure 2: The enclosure used to expose vines to simulated wildland fire smoke, showing the outside (top left), inside (top middle) and inside during smoke exposure (top right). The sample collection and processing scheme resulted in 560 total samples collected from the 2016 growing season. This includes HMG, FRJ, washed and unwashed samples.

In addition to the samples outlined above, a second set was collected by sampling 50-60 vines per varietal over an area of 1-2 acres. This was done to quantify endogenous levels of key VP in these varietals, with the goal of facilitating a rigorous statistical comparison between control and smoke-exposed berries. These samples were collected at commercial maturity, which corresponded to the last time-point for each varietal. Geographical information system (GIS) coordinates for each sample were collected to enable replicate analyses of the same vines across multiple years and to assess the presence of trends as a function of location.

To quantify the concentration of VP known to contribute to the negative sensory attributes of wine made from smoke-exposed berries (Figure 1), targeted analysis of nine VP was conducted. This quantitative analysis was performed on berry extracts using gas chromatography-mass spectrometry (GC-MS), which separates VP in the time domain (GC), then uses a sensitive and specific detector (MS) to ensure the correct signal corresponding to the desired VP are accurately quantified (Figure 3).

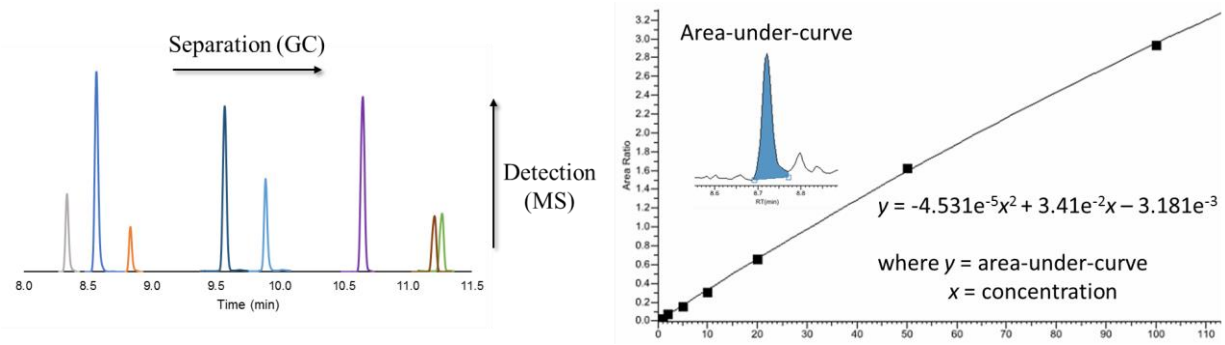


Figure 3: Gas chromatography-mass spectrometry (GC-MS) involves separation in the time domain (GC), with each peak in the above data color-coded to a single compound (left). This is followed by sensitive detection (MS) that is specific to each time-response pair (left). Integration of the area-under-the-curve for each compound and subsequent comparison to a calibration function results in accurate quantitative analyses (right).

To date there has been a myopic focus on VP and their glycosides to predict quality issues in wine made using smoke-exposed berries. While successful and informative, this approach has obvious limitations given that VP and their glycosides are only 50-80% predictive of wine quality issues. Improving this predictive accuracy requires a broad comparison of the chemical composition of smoke-exposed and control berries. The use of mass-spectrometry-based non-target screening workflows (*i.e.*, metabolomics) will facilitate this characterization. My approach uses ultra-high pressure liquid chromatography (uHPLC) to separate compounds prior to detection using MS. Conceptually this is similar to the GC-MS approach described above. However, in this workflow MS detection is non-targeted, often producing in excess of 10,000 unique masses per sample that need to be mined for significance. Moreover, rather than producing a simple-to-interpret VP concentration the output from non-targeted analysis is qualitative, producing a mass that the analyst must assign significance to. This is most often achieved by assigning an empirical chemical formula, or by performing a statistical comparison using a given accurate mass (Figure 4). As a final layer of complexity, the MS used in this study produces masses accurate to the fourth decimal place, but each measurement also has an uncertainty associated with it (± 2-5 parts-per-million). This creates a mass binning problem that must be addressed before significance can be assigned.

Data Analysis

The desired output from quantitative GC-MS analysis is a table of VP concentrations that need to be correlated back to specific experimental conditions (*vide supra*). Based on the tested sample matrix for control *versus* smoked-exposed wine grapes, GC-MS analysis will generate a total of 5,220 (580 samples x 9 VP) unique VP concentrations that need to processed into meaningful results summaries. As well, the samples associated with specific GIS coordinates (1,980 unique VP values) need to be visualized to assess spatial trends and determine summary statistics to establish baseline levels of VPs in control vines. To manage these large data sets and enable data from future studies to be easily integrated into an efficient data management system, I will build a series of relational databases in Access that will contain the quantitative GC-MS results and the associated metadata. Using SQL queries for data reduction and analysis, I then propose to take queried data sets and export them into R for statistical analyses and to Tableau and R for data visualization. As well, I intend to use R to model left-censored data (where some samples quantitate below a defined threshold(s)) when calculating summary statistics. This process will require an evaluation of the degree of censoring per VP and varietal, determining the nature of each VP

distribution (*e.g.*, normal, log-normal, *etc*.) and finally, applying the appropriate statistical tests to obtain values for the censored data.

To facilitate the generation of maps in Tableau, I will create a conversion algorithm in Excel to change GIS coordinates in degrees:minutes:seconds to decimal degrees, which is the format required by Tableau. After generating the appropriate background map in Tableau, I will use the converted GIS coordinates to map the concentration of VPs over the areas surveyed.

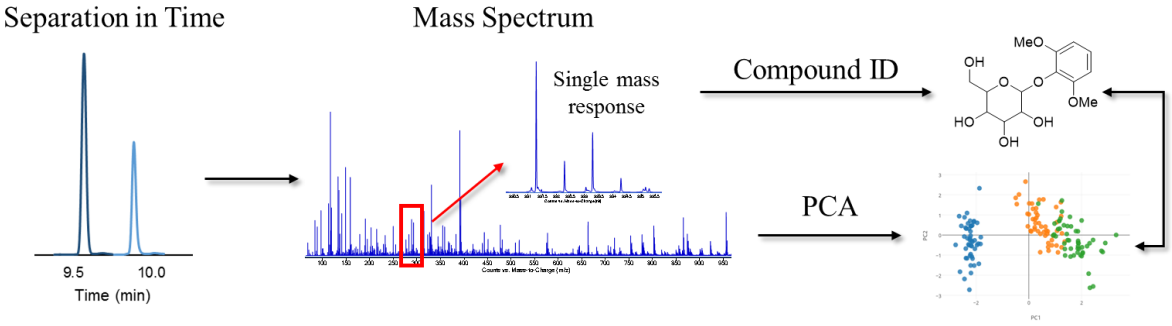


Figure 4: A simplified representation of a potential non-targeted (metabolomics) screening workflow. Following separation in the time domain a series of mass spectra are generated. The spectra can be mined for significance (*e.g.*, principle component analysis [PCA]), after which key mass responses can be targeted for compound identification, or *vice versa*.

A non-targeted analytical workflow produces data that requires a much different data analysis stream than quantitative analysis (Figure 4). Given the complexity associated with mining uHPLC-MS data, which involves identifying and binning relevant MS responses that correlate to chromatographic peaks (as per Figure 4), instrument vendor software will be used (MassHunter, Agilent Technologies) in tandem with custom-built Excel tools to facilitate data reduction.

On the front end of data reduction, Excel will be used to provide a list of masses that MassHunter should search for. This type of workflow is referred to as 'known-unknown' screening. Providing this list requires the calculation of exact masses for a series of chemical formulae. Most MS vendor software packages enable the calculation of the exact mass for a single chemical formula. However, when a series of formula conversions are required the user is left to do this one-by-one. To improve the efficiency of this process I will build an exact mass calculator in Excel. This tool will take a chemical formula (with a defined set of elements), parse out the quantity of each element and return the exact mass. As well, I will build an Excel template that will perform simple *in silico* chemical reactions to aid in the generation of a combinatorial database for VP-glycosides.

Results and Discussion

Non-Targeted Data Analysis

To support 'known-unknown' screening a parsing tool was created to do batch calculations of exact masses given a list of chemical formulae as input (Figure A1). This tool used the FIND() and ISERROR() functions to index the constituent elements from each formula string. The ISERROR() function was used to return '0' if no elements of given type were found, which was required to enable correct referencing in subsequent steps. After indexing, an iterative IF() statement was generated to pull out the number of atoms of each element. The IF() iterations locate the index for each element in the previous step using the MID() function, then define the length of each number sequence using the ISNUMBER() function. The number of IF() iterations used limited the maximum number of each element to 999, which is an acceptable range for the type of analyses conducted as part of my PhD research. The final piece of this tool was the calculation of the monoisotopic mass, which involved the sum of absolute references to accurate masses for each element multiplied by the quantity of each element parsed from the input formulae.

A combinatorial database of VPs and sugars known to be involved in glycoside formation in wine grapes was constructed using the exact masses of the constituent components. These masses were batch-calculated using the Excel tool described above. The glycoside chemical formulae were calculated by referencing the parsed chemical formulae of each component of the glycoside and adding up the constituent elements via a series of VLOOKUP() functions in tandem with concatenation of the resulting sums (Figure A2). The exact mass of the glycosides were calculated using VLOOKUP() to find the exact mass of the constituents and sum them. For the calculation of chemical formulae and exact masses, the loss of H_2O during the formation of a new glycosidic bond was accounted for.

The combinatorial database was imported directly into the MassHunter software to interrogate uHPLC-MS data. For this analysis, a series of control and smoke-exposed Cabernet Franc wine grapes were used. As an example of the output from MassHunter for this analysis, a schematic of the final workflow is shown in Figure A3. The first step required input in the form of exact masses from the combinatorial glycoside database. The MassHunter software then utilizes a proprietary algorithm (with exact masses as input) to identify potential matches in uHPLC-MS data. From the Cabernet Franc data analyzed, a variety of putative VP-glycosides were identified at elevated levels in the smoke-exposed sample set when contrasted to the control wine grapes (data not shown). Notable, was the strong match for syringyl-glucuronide, as the glucuronic acid family of sugars has not been reported in the literature. If this finding can be confirmed, it would mean that current methods for assessing the impact of smoke-exposure neglect an entire class of glycosides. It remains to be determined if such glycosides influence the sensory perception of wine made using smoke-exposed berries.

Targeted Data Analysis

<u>Endogenous VP Concentrations in Four *V. Vinifera* Varietals</u>

Excel was used for unit conversion and parsing to take GIS coordinates in degrees:minutes:seconds (dd° mm.sss') to decimal degrees, which is the format required by Tableau (Figure A4). In this formula the LEFT() and MID() functions were used to parse out the degrees, minutes and seconds from the coordinates obtained from a GIS device (*e.g.*, 49° 50.423' from a Garmin GPS unit). The parsed values were converted to decimal degrees as part of the same function, with a scaling factor used to adjust for the

datum output by the GPS device (referenced to WGS 84). The final step involved an IF() statement that checked the direction of each coordinate and assigned a negative value if the direction was West or South.

Using the GIS coordinates (in decimal degrees), the ability of Tableau to efficiently visualize the spatial distribution of VPs was evaluated for four varietals from two vineyards in the Okanagan Valley. To add interest to the final visualization, the addition of satellite images in lieu of the stock Tableau base maps was explored. Adding a satellite image in Tableau currently requires the use of Mapbox Studio as the source of satellite imagery. Going through this process yielded a more interesting and informative map, as the rows at each vineyard could be seen underlying the VP data (Figure A4). After importing the quantitative results from GC-MS analysis into Tableau, a dashboard was created to enable simultaneous review of all varietals (Figure A5). Each map in the dashboard was generated with the varietal coded by color and the amount of VP coded by size. To enable visual assessment of data trends, VP concentrations were binned to improve size discrimination. From figure A5 it is apparent that pinot noir has a higher guaiacol concentration than the other three varietals. It remains to be seen if this elevated level of guaiacol results in an increased susceptibility to sensory issues in wine following pinot noir grapevine smoke-exposure. Moreover, it could be argued that summary statistics for each VP in each varietal (*e.g.*, box-whisker plot) would be a more efficient data comparison. This argument seems valid for peer-reviewed literature. However, for graphically representing this data in conferences and presentations the use of the Tableau map in tandem with summary statistics visually guides the audience as to how the data should be interpreted, which is a more effect means of conveying large amounts of data with complex relationships.

To organize the quantitative results, a series of relational databases were created in Access (Figure A6). These relational databases were interrogated using SQL queries to generate reduced data sets, which were migrated into R Studio to calculate appropriate summary statistics. As well, a combination of R Studio and Tableau were used to generate publication quality summary figures. As an example of this approach, I used an SQL query to left-censor the raw quantitative GIS data based on the calculated limits of detection (LOD) and quantitation (LOQ) for the GC-MS method (see 'Censoring Query' for query code). For simplicity, the derivation of the LOD/LOQ values are not shown here, but the absolute values are summarized in the PhysicalParameters database.

The percentage of left-censored data, the distribution of the data (*vide infra*) and the number of data points determine the statistical methods that can be used to model censored data. Such modelling is often necessary to calculate meaningful summary statistics for data sets where a significant portion could be left-censored. Choosing an accurate censorship model is critical, as it enables the calculation of summary statistics with minimal bias, as might be obtained from the more common practice of substituting a value (*e.g.*, zero, detection limit/2, *etc*.)[7].

Following data censoring, I used an SQL query to tabulate the percent of left-censored data for each varietal (see 'Percent Censored Data' query and 'percentCensored' table in smokeVolatiles.mdb). After saving the query results in Access, I linked Tableau to the Access database and generated a summary plot indicating the suggested course of action for each data set by color code (Figure A7). Displaying the data in this way was an efficient summary for an audience unfamiliar with my methodologies, as it clearly indicated how the GIS data were treated statistically to produce summary statistics. Based on the summary in Figure A7 I chose to proceed with further statistical calculations of the Pinot Noir data, as the degree of censoring in the other varietals was greater than 80%, making it is difficult to develop a model for the censored data, regardless of the algorithm employed[7].

As the next step in data processing, it was necessary to evaluate the probability distributions of the uncensored data to determine if they were normally distributed. To accomplish this, an SQL query ('pinotnoirProb' in smokeVolatiles.mdb) was written to isolate the Pinot Noir data from the 'censoredData' table. Once extracted, the Pinot Noir data was exported to Excel where the previously

censored values were replaced with either null or zero values. The use of different data fields for the censored data was necessary to accommodate subsequent statistical calculations (*vide infra*). The resulting table was imported into R Studio from a text file and probability distributions were generated using the built-in qqplot and qqline functions (Figure A8). For ease of presentation, the plots were displayed in a 2 x 2 matrix layout using the par() and mfrow() functions. The observed distributions suggested that guaiacol followed a normal distribution, while syringol fit a log-normal distribution. With this difference in probability distributions for known (uncensored) data sets it was deemed inappropriate to assume a normal distribution for data sets containing censored values. This meant that non-parametric methods should be used when performing hypothesis tests and calculating summary statistics.

In spite of the above conclusion regarding non-parametric tests, to better understand the differences between parametric and non-parametric approaches to modelling left-censored data, summary statistics were calculated for the eugenol (11% censored) and p-cresol (50% censored) results. Generation of the Kaplan-Meier, regression on order statistics (ROS) and maximum-likelihood estimation (MLE) models was performed in R Studio using the NADA and Survival packages, which include built-in functions to perform the required calculations. The mean and standard deviation for each model were transposed directly from R Studio into Excel for the creation of summary plots (Figure A9). For eugenol, which was only 11% censored, the choice of statistical model or substitution method did not have an apparent impact on the mean or standard deviation. However, for p-cresol, which was 50% censored, the method of calculating summary statistics had an obvious impact, with overt differences in means and standard deviations. For the p-cresol data, the ROS and MLE results were quite different, despite both methods being suggested as reasonable approaches to modelling left-censored data[7]. The source of this discrepancy is currently unclear. However, given the strong literature support for using MLE and the generation of a standard deviation in-line with the variance of the uncensored data, this model was used to generate summary statistics for all compounds with less than 80% left-censored values.

Impact of Smoke-Exposure in Four *V. Vinifera* Varietals

Originally, the impetus for rigorously calculating summary statistics from the GIS data set was to have a solid understanding of the endogenous levels of key VPs in *V. vinifera* berries at commercial maturity (*i.e.*, when the GIS data set was collected). Having this data would enable a strong statistical comparison to the levels observed in smoke-exposed berries at different points in berry maturation. Given the high degree of censoring across varietals for the endogenous VP levels (Figure A7), the application of VP censoring models to the smoke-exposure experiments was not feasible. As such, statistical evaluation of the smoke-exposed berries was performed by comparing control and smoke-exposed berries at each time-point.

Like the GIS data set, the smoke-exposure results were compiled in Access (‘quantData’ table). The raw data was left-censored using the same LOD and LOQ values that were applied to the GIS data set. In this instance, the results were censored in Excel prior to compiling the Access database. To obtain a measure of the degree of censoring an SQL query was used (‘smokeCensored’) to reduce the data, after which it was imported to Tableau to generate a meaningful data summary (Figure A10; see ‘smokeCensoredPlot.twb’). This summary figure efficiently illustrated that the control samples were heavily censored, such that the calculation of summary statistics would not be possible (similar to the GIS data). As well, it showed that the Merlot results were almost fully censored for control and smoke-exposed treatments. This result was anticipated, as Merlot was the first varietal to receive smoke treatment and the difficulties associated with field studies lead to inconsistent smoke exposure. Modifications made to the smoking procedure for subsequent field trials lead to more consistent smoke-exposure and a

corresponding increase in VPs the other varietals. Further support for this finding was obtained by generating a stacked bar plot in Tableau to illustrate the total VPs in each varietal as a function of time-point (Figure A11). This qualitative figure was generated by linking directly with Access from Tableau. After linking, the 'quantData' table was used to populate the stacked bar plot, with averages across five biological replicates calculated automatically by Tableau[1]. Clearly, the absolute concentration of VPs and the breadth of observed VPs in the Merlot were lower than in the other varietals. Based on these observations, the Merlot data was removed from further evaluation.

Significance of Findings and Conclusions

Data reduction and data management tools were utilized to build efficiency into the work-flows necessary to characterize the impact of smoke-exposure on *V. vinifera* berries. A custom-built Excel template was created to perform simple *in silico* chemical reactions (*e.g.,* formation of an O-glycosidic bond) and generate a combinatorial database of potential compounds that might be found in smoke-exposed berries. The output from this Excel tool was used to tentatively identify a guaiacyl-glucuronic acid species that has not been reported to date. If the identity of this compound can be confirmed, it means a whole class of sugar conjugates (including all the VPs discussed herein) has gone unreported, which may help explain why current analytical methods are only 50% predictive of quality issues when making wine using smoke-exposed berries.

Using SQL queries (for Access), Excel manipulations, R Studio and Tableau, the nature of the GIS data was characterized to determine the correct statistical tests to generate accurate summary statistics and hypothesis tests. Included in this initial analysis was the generation of a Tableau figure that summarized the GIS data set with respect to the degree of left-censoring, an assessment of the normality of the data in R Studio. The use of Tableau to create information-rich visual summaries will be invaluable when submitting work for peer-reviewed publication, as they efficiently convey a large amount of information. After establishing the need for non-parametric statistical methods, various approaches for calculating the mean and variance of left-censored data were compared. The outcome of these studies will inform data treatment through the remainder of my graduate studies.

Like the GIS data, SQL queries, Excel manipulations and Tableau were utilized for data analysis. The heat map generated in Tableau was an effective way of summarizing this large data set. From this summary, it was apparent that the Merlot data was heavily left-censored in control and smoked treatments, leading to the removal of this varietal from subsequent analyses. The stacked bar plot qualitatively suggested that there were differences in the VP ratios as a function of varietal, although these differences have not yet been assessed statistically.

Automation of hypothesis testing is an ongoing part of this project. The number of individual tests required to compare all time point, VP and varietal permutations makes single calculations time-consuming. Excel could handle standard (parametric) t-tests using built-in functions or the Data Analysis add-in. It could also handle non-parametric hypothesis testing, in the form of the Wilcoxon Rank Sum test, if a custom template were built. However, given the large number of hypothesis tests required for the current data set, as well as those acquired in subsequent harvests, R Studio will be the ideal venue to streamline these calculations. Once the correct R program is written, it can also be modified to automate other summary statistical functions (*e.g.,* mean).

[1] This data was censored and the averages are reported without accounting for the censored data. Since this graphic was intended to be qualitative, it was not deemed requisite to have the appropriate measures in place to calculate summary statistics on censored data like this, where n was very low.

Bibliography

[1] R. Ristic, A. L. Fudge, K. A. Pinchbeck, R. De Bei, S. Fuentes, Y. Hayasaka, S. D. Tyerman, K. L. Wilkinson. Impact of grapevine exposure to smoke on vine physiology and the composition and sensory properties of wine. *Theor. Exp. Plant Physiol.* **2016**, *28*, 67.

[2] R. Ristic, K. A. Pinchbeck, A. L. Fudge, Y. Hayasaka, K. L. Wilkinson. Effect of leaf removal and grapevine smoke exposure on colour, chemical composition and sensory properties of Chardonnay wines. *Aust. J. Grape Wine Res.* **2013**, *19*, 230.

[3] K. R. Kennison, M. R. Gibberd, A. P. Pollnitz, K. L. Wilkinson. Smoke-derived taint in wine: The release of smoke-derived volatile phenols during fermentation of Merlot juice following grapevine exposure to smoke. *J. Agric. Food Chem.* **2008**, *56*, 7379.

[4] K. R. Kennison, K. L. Wilkinson, H. G. Williams, J. H. Smith, M. R. Gibberd. Smoke-derived Taint in Wine : Effect of Postharvest Smoke Exposure of Grapes on the Chemical Composition and Sensory Characteristics of Wine Smoke-derived Taint in Wine : Effect of Postharvest Smoke Exposure of Grapes on the Chemical Composition and Senso. *J. Agric. Food Chem.* **2007**, *55*, 10897.

[5] M. P. Krstic, D. L. Johnson, M. J. Herderich. Review of smoke taint in wine: Smoke-derived volatile phenols and their glycosidic metabolites in grapes and vines as biomarkers for smoke exposure and their role in the sensory perception of smoke taint. *Aust. J. Grape Wine Res.* **2015**, 537.

[6] M. Parker, P. Osidacz, G. A. Baldock, Y. Hayasaka, C. A. Black, K. H. Pardon, D. W. Jeffery, J. P. Geue, M. J. Herderich, I. L. Francis. Contribution of several volatile phenols and their glycoconjugates to smoke-related sensory properties of red wine. *J. Agric. Food Chem.* **2012**, *60*, 2629.

[7] D. Helsel. *Nondetects and Data Analysis*. John Wiley & Sons, Inc., **2005**

Appendix

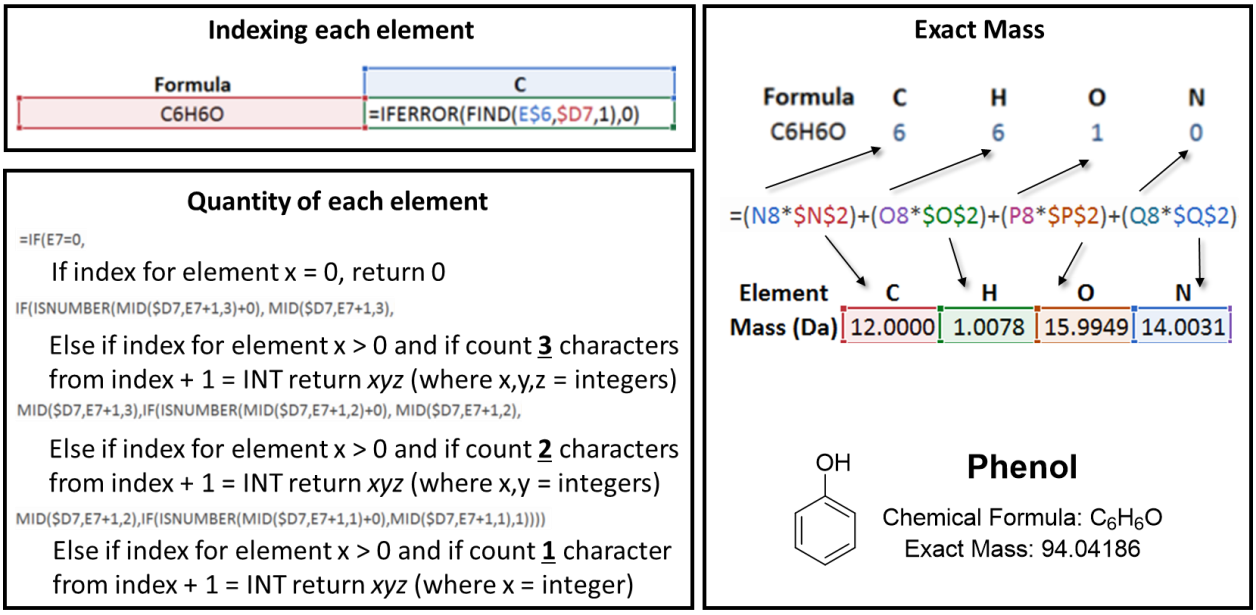
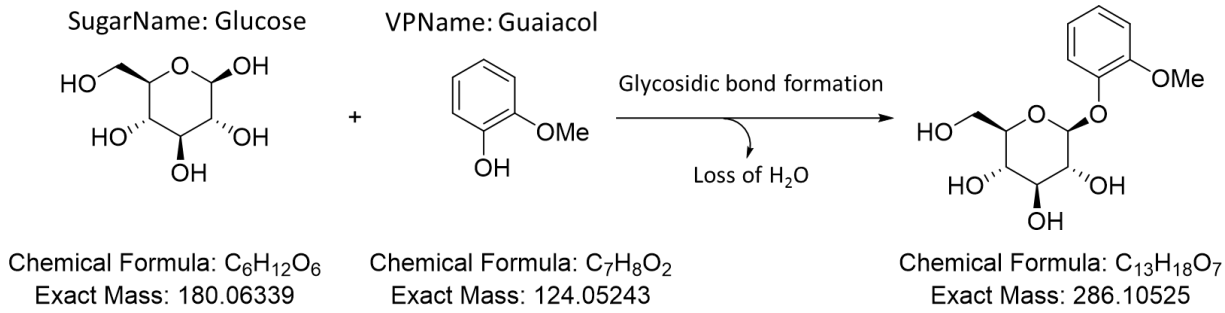


Figure A1: Explanation of the functions used to calculate exact mass using a list of formulae as input. After indexing (top left), the number of elements in a given formula are determined (bottom left) by iteratively evaluating the length of the integer following a given element and returning it as in integer (*e.g.*, $C_{21}H_{200}O_9$ returns exactly 21 carbon, 200 hydrogen and 9 oxygen atoms). Absolute referencing to exact elemental masses, the parsed quantity of each element and a summation formula yield the final output of exact mass (right).

Element	Excel Formula Component
C	=“C”&(VLOOKUP(SugarName,ParsingSheet,QuantityC,FALSE)+VLOOKUP(VPName,…))
H	&“H”&(VLOOKUP(SugarName,ParsingSheet,QuantityH,FALSE)+VLOOKUP(VPName,…)-2)
O	&“O”&(VLOOKUP(SugarName,ParsingSheet,QuantityO,FALSE)+VLOOKUP(VPName,…)-1)



Component	Excel Formula Component
Sugar	=VLOOKUP(SugarName,ParsingSheet,ExactMass,FALSE)
VP	+VLOOKUP(VPName,ParsingSheet,ExactMass,FALSE)
H_2O Loss	-ModificationsSheet(H2OExactMass)

Figure A2: A glycoside combinatorial database of known sugars and VPs was constructed. The exact mass of each possible component was calculated using the parsing tool discussed in Figure A1. A series of VLOOKUP() and concatenation functions, including correction for the loss of H_2O during formation of the new glycosidic bond (-2 and -1 for H and O, respectively), yielded the chemical formulae for the possible glycosides (top table). The exact mass of the glycosides were calculated with VLOOKUP() functions to pull from the original parsed formula information for each component, followed by a correction for the loss of H_2O (bottom table). An example of the separate components and their resulting glycoside is given (middle).

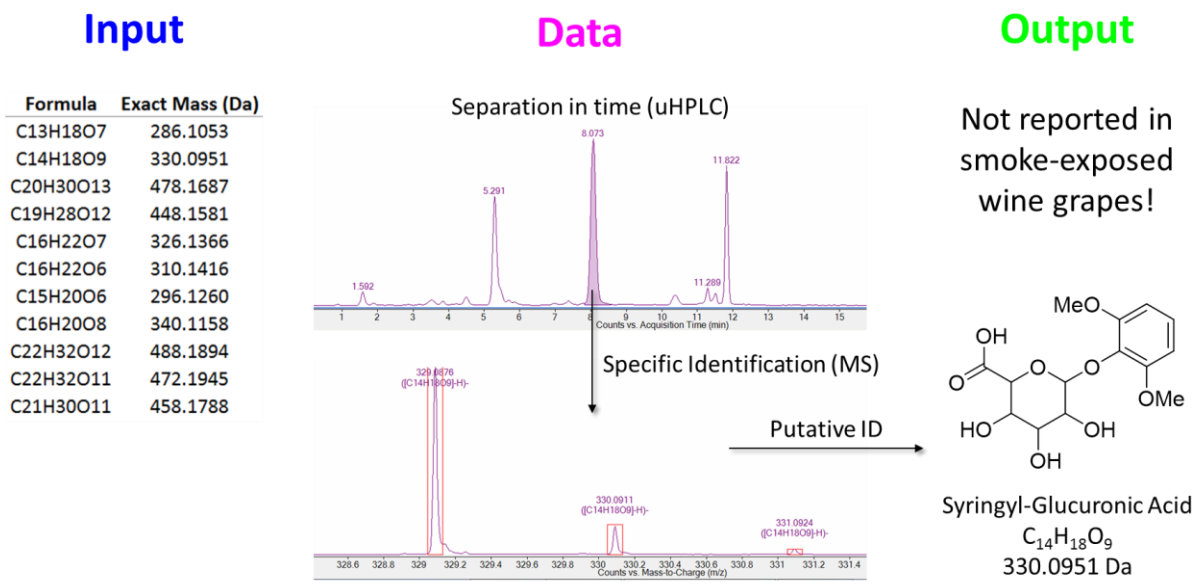


Figure A3: A sample of the output from the MassHunter software for 'known-unknown' screening. The first step required input in the form of exact masses from the combinatorial glycoside database. The MassHunter software uses a proprietary algorithm that utilizes exact masses to identify potential matches from uHPLC-MS data. From the Cabernet Franc data analyzed, a variety of putative VP-glycosides were identified (data not shown). Of particular interest was the strong match for syringyl-glucuronide, which was present at higher levels in the smoke-exposed wine grapes when contrasted against the control samples. The glucuronic acid family of sugars has not been reported in the literature.

Direction	GPS (dd:mm:sss)	GPS (decimal degrees)
N	49° 50.584'	49.8429
W	119° 34.226'	-119.5704

=(LEFT(K226,2)+(MID(K226,5,2)/60)+(MID(K226,8,3)/3600/16.92))*(IF(L226="S",-1,1))

Degree Minute Seconds Direction

=(LEFT(N226,3)+(MID(N226,6,2)/60)+(MID(N226,9,3)/3600/16.92))*(IF(O226="W",-1,1))

Tableau Base Map

Satellite Base Map

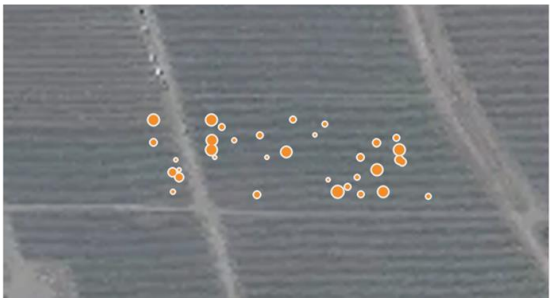


Figure A4: To convert the output of the handheld GPS device in degrees:minutes:seconds (dd° mm.sss') to the decimal degrees required by the Tableau software, an Excel template was constructed (top). The use of satellite base maps (bottom right) improved the visual appearance of the data overlay compared to the default Tableau base map (bottom left). The satellite image also added information content since the data was superimposed over the vineyard rows. The size of the data points represents the relative amount of guaiacol.

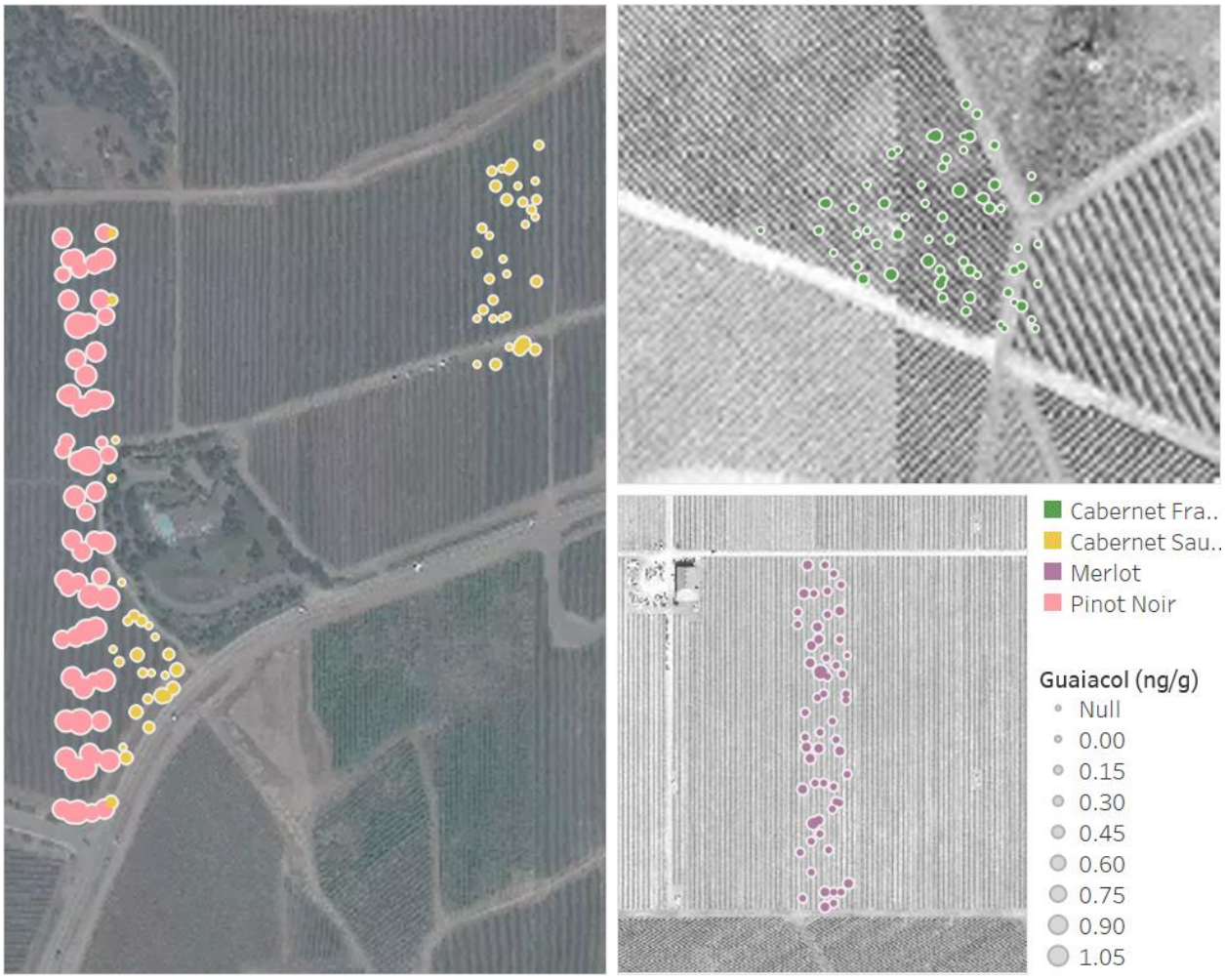


Figure A5: A Tableau dashboard created to review the relative levels of VPs in the four varietals examined. Guaiacol concentrations are shown, with the color matched to the varietal and the size of each data point correlated to guaiacol concentration (ng/g) that was binned in 0.15 ng/g steps for ease of presentation. This sample data clearly shows that pinot noir has a higher baseline level of guaiacol compared with the other three varietals.

harvestDetails	**PhysicalParameters**	**timePoints**	**varietalDetails**	**quantResults**	**gpsData**
ID	Compound	Varietal	ID	Varietal	dateCollected
Varietal	Type	timePoint	Vineyard	harvestYear	vineyardHarvestDate
timePoint	Formula	Date_Smoked	Varietal	timePoint	Vineyard
Condition	Mass	Date_Harvested	Clone	Condition	Varietal
Plant	pKa	timeDays	Rootstock	Plant	Bag #
Brix	logP	timeHours	yearPlanted	Washed	Sample ID
Brix_StdDev				Tissue	Latitude
BerryWeight				Index	Longitude
				sampleCode	Brix °
				vialCode	Ethylguaiacol
				amtWeight	Ethylphenol
				Guaiacol	Methylguaiacol
				d3Guaiacol	Eugenol
				Methylguaiacol	Guaiacol
				4Ethylguaiacol	oCresol
				d5Ethylguaiacol	pCresol
				oCresol	Syringol
				pCresol	Vanillin
				Eugenol	totalPhenolics
				Ethylphenol	
				d4Ethylphenol	
				Syringol	

Figure A6: Summary of the relational databases and the attributes of each database created in Access to manage quantitative GC-MS data. Primary keys are indicated with red type-face.

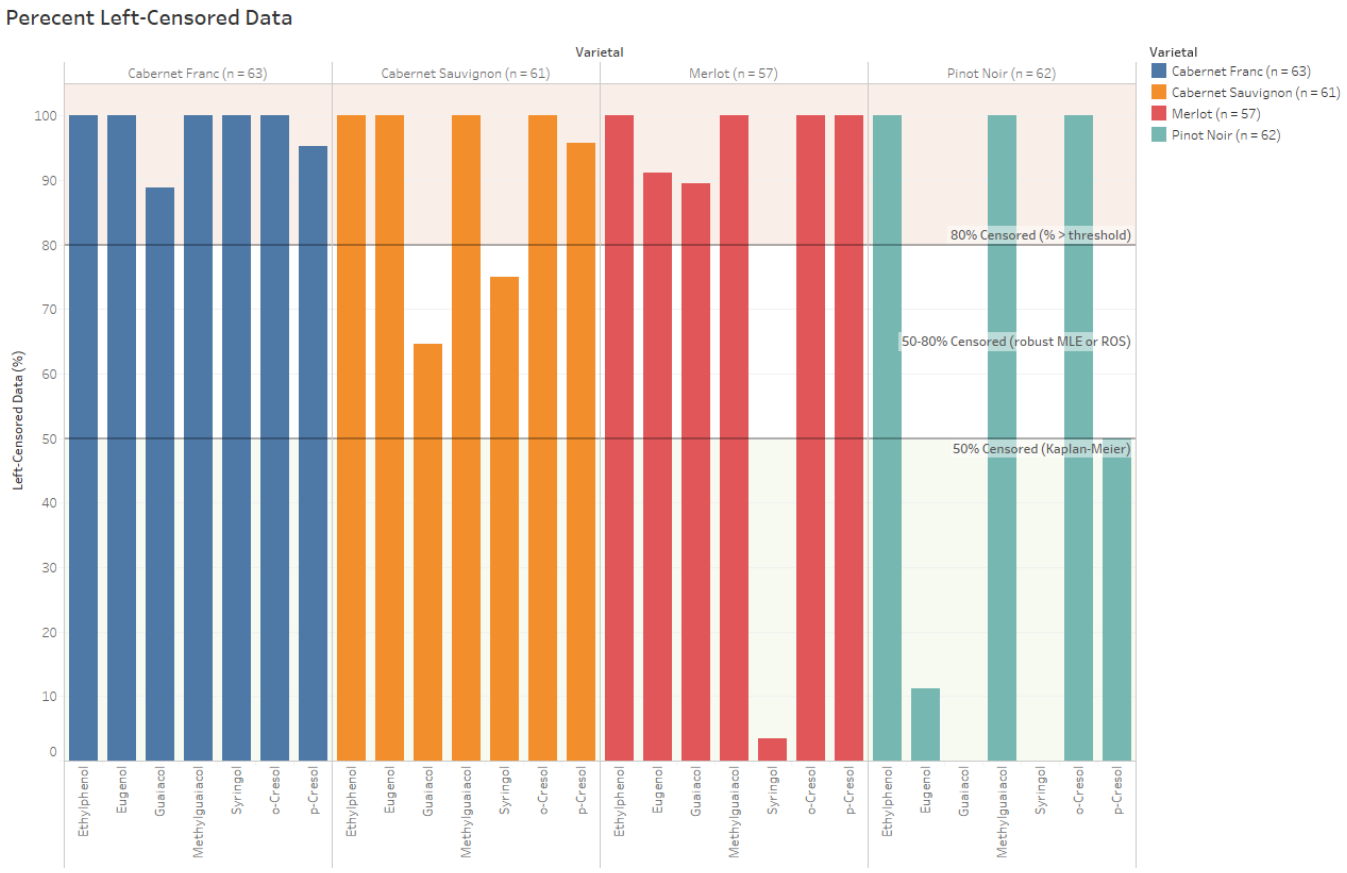


Figure A7: An SQL query was used to calculate the percent of each compound that was left-censored as a function of varietal. To quickly convey these findings and the actions following from them, a Tableau summary was created via a direct link to the SQL query in Access. The reference lines indicate the appropriate statistical treatment of the censored data given the degree of censoring. MLE = maximum likelihood estimation; ROS = regression on order statistics.

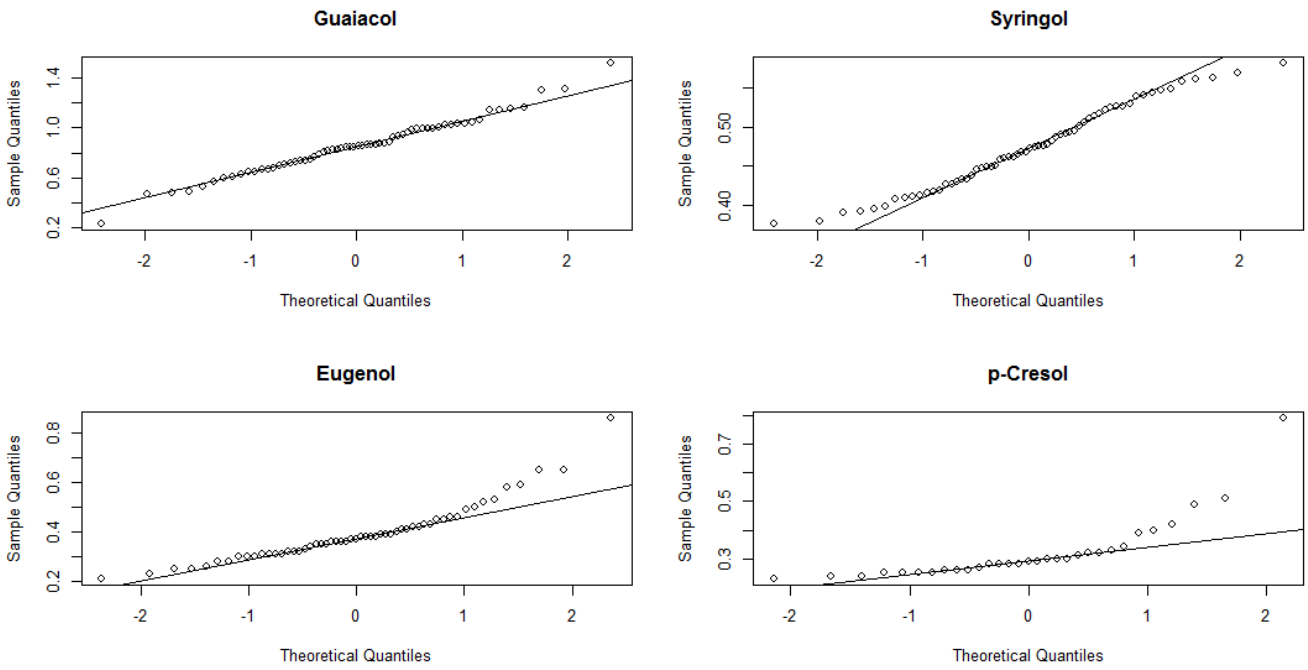


Figure A8: Probability distribution functions for the key compounds observed in Pinot Noir grapes. Guaiacol appeared to follow a normal distribution, while Syringol showed a generally normal distribution with outliers. Due to ambiguity for eugenol and p-cresol, parametric and non-parametric approaches to modelling left-censored data were compared for all compounds.

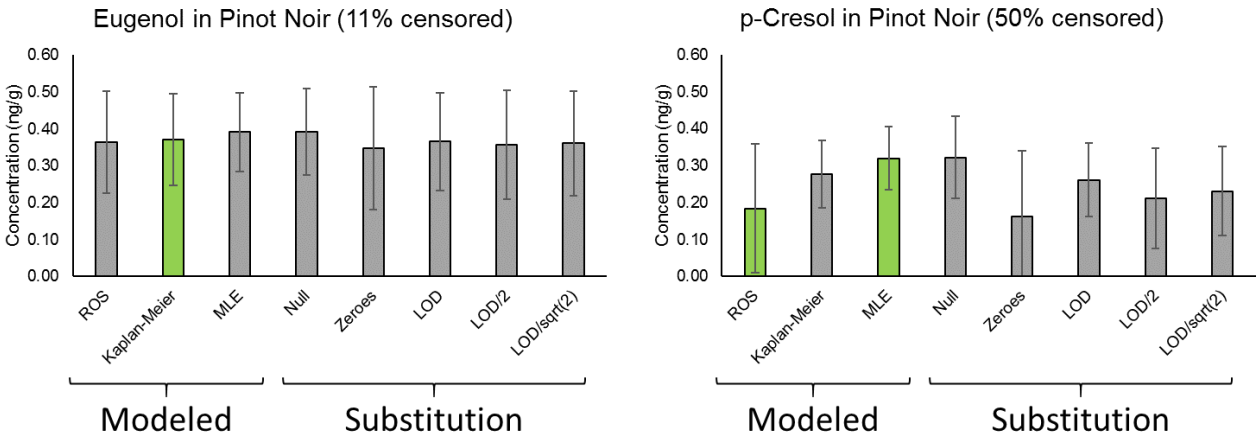


Figure A9: Comparison of statistical modelling and substitution-based approaches to calculating summary statistics for left-censored data sets. Results are shown for eugenol (left) and p-cresol (right) observed in Pinot Noir extracts. With low percent censoring, the choice of model or substitution method does not significantly impact the sample mean or standard deviation. When censoring is high, the mean and standard deviation change depending on the method employed. The green bars represent the suggested approach to modelling based on the degree of censoring. All data is shown ±1 standard deviation. ROS = regression on order statistics; LOD = limit of detection.

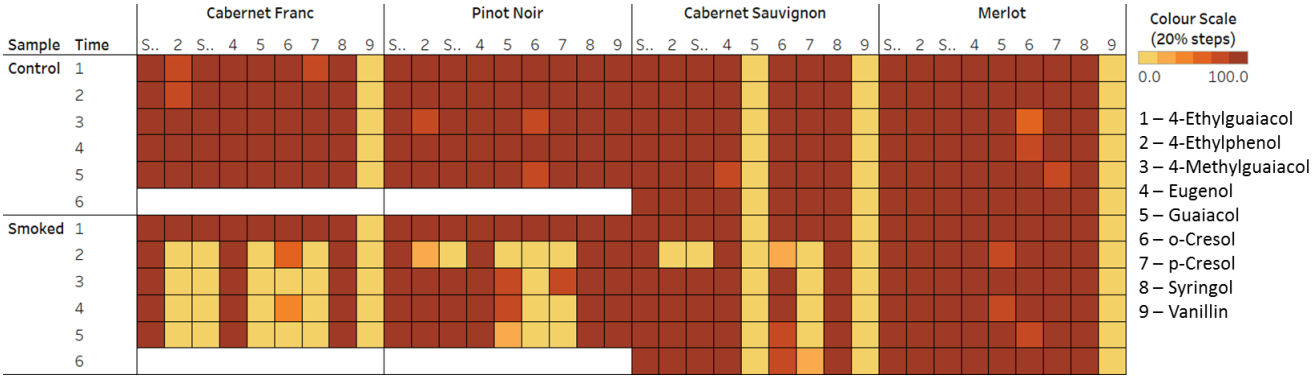


Figure A10: Degree of left-censoring in the smoke-exposure data. The control samples were heavily censored such that summary statistics were not calculable. As well, the Merlot data was heavily censored for the control and smoke-exposed sample groups, leading to the removal of this varietal as a comparison point.

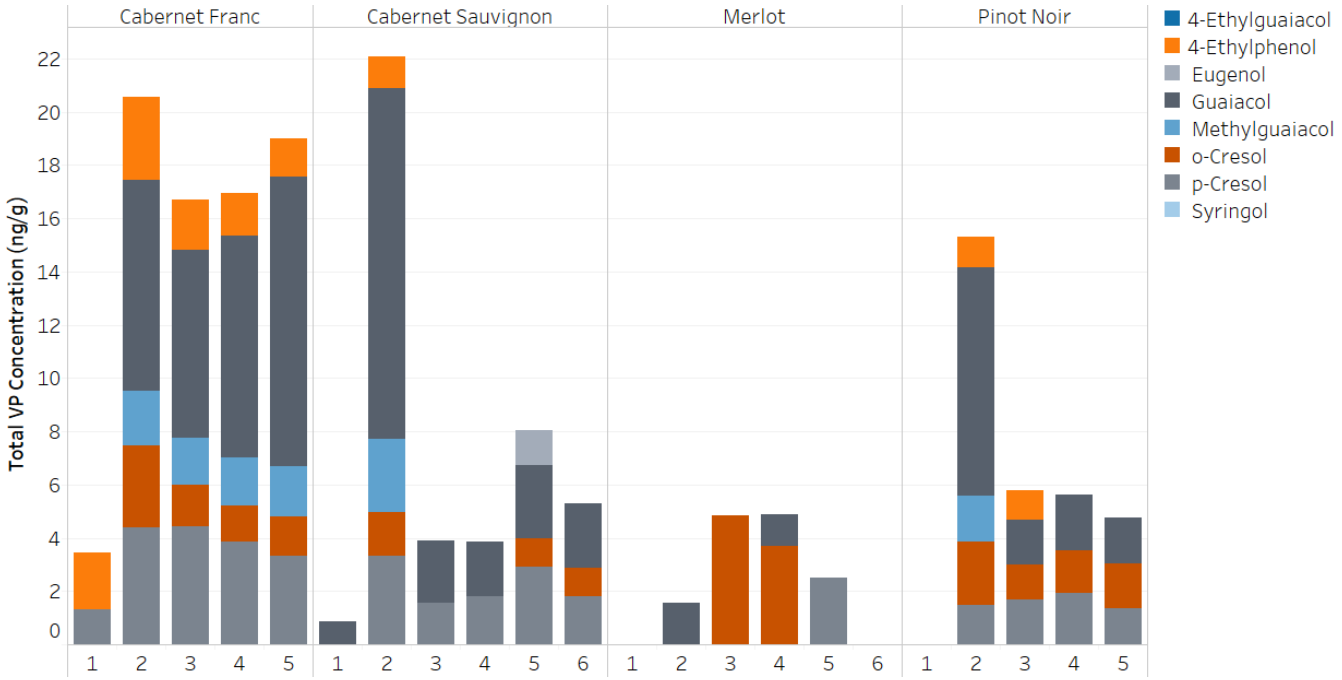


Figure A11: Comparison of total VPs as a function of *V. vinifera* varietal. The total amount and breadth of VPs present in smoke-exposed Merlot lead to the removal of this varietal from subsequent calculations.