

Data Mining Techniques for Modelling Seasonal Climate Effects on Grapevine Yield and Wine Quality

Subana Shanmuganathan and Philip Sallis

Geoinformatics Research Centre
School of Computing and Mathematical Sciences
Auckland University of Technology, New Zealand.
e-mail: subana.shanmuganathan@aut.ac.nz

Ajit Narayanan

School of Computing and Mathematical Sciences
Auckland University of Technology, New Zealand
e-mail: philip.sallis@aut.ac.nz,
ajit.narayanan@aut.ac.nz

Abstract—*The paper describes ongoing research in data mining techniques investigated for modelling seasonal climate effects on grapevine phenology that determines the ratio of grape berry composition that in turn determines the fineness of wine vintage in addition to winemaker experience and talent. A brief introduction to the literature in this problem domain is followed by a discussion on conventional statistical data analysis methods that looks at the problems in using these methods with only a decade old data, often considered as incomplete in sequence. Data relating to vineyard yield with its coincident seasonal climate change is used in this study to model seasonal climate effects at micro scales i.e., vineyard, using data mining techniques, decision trees and statistical methods. The initial results show potential for predicting future grapevine yield using vineyard data for more specific scenario building than is possible now, using macro climate data.*

Keywords; *self-organising maps, grapevine phenology*

I. INTRODUCTION

Gaining the best possible insights to vineyard micro climatic conditions helps viticulturists and enologists to refine the quality of vintage and wine style [1]. The fineness of a vintage relating to its wine appellation greatly depends on the winemaker but grapes with an appropriately balanced mix of sugar and pro-compounds of phenols responsible for aroma, colour and flavour are the essential ingredients for producing premium quality wine. These factors are influenced considerably by prevailing weather conditions. This paper reflects research focussed on modelling seasonal weather variability to quantify the anecdotal information on wine vintage and style quality dependencies that are specific to a site (i.e., vineyard). On the other hand, the environmental and basic climate conditions are the major factors considered when determining the wine style (grapevine variety) suitable for a site.

The vintage (year-to-year) variability in wine quality is considered to be determined mainly by the elements of weather (seasonal) that ripened the grapes based on the fact that the influences by other environmental and viticulture practices are seen as less significant in this time scale [2]. This is why better understanding is continuously sought on the associations between the two sets of data; wine vintage quality, the dependent and the weather, the independent

variables; interestingly the variability of the latter could be *dramatic* in some locations [3].

The literature section outlines major conventional but recently seen as *popular* approaches to analysing the associations between weather and vintage data (or wine ratings) covering over five decades at regional scales i.e., winegrowing regions. The paper then describes the major constraint when applying these approaches proven to be good at establishing the associations at regional scales to data sets at different spatial and temporal scales, such as to a vineyard data covering only a decade. This is followed by data sources, pre-processing operations adopted and the methodology being investigated in this research to overcome the constraint. The final section presents the initial results produced with an example of yield and weather data sets from a vineyard in north of Auckland covering a 12 year period, with future directions for continued research.

II. LITERATURE IN CLIMATE EFFECTS ON GRAPEVINE

The literature reviewed for this research reveals some recent successful approaches to modelling the climate effects on viticulture and wine quality with five or more decade old data and are outlined here.

Grapevines, being one of the oldest cultivated crops, combined with traditional winemaking processes (some of them even dating back to mediaeval times), means that the science of viticulture comprises a rich geographical and cultural history of development [4]. Present day wine regions located in considerably narrow geographical hence invariantly climate niches have eventually become famous for their wine style and appellation developed over significantly long periods of time, i.e., decades or centuries. The wine appellation produced from a region is an outcome of its base climate, but the quality of the vintage is the result of seasonal (year-to-year) variability from the average climatic conditions of that region/ vineyard. In view of this fact various temperature based matrices, such as growing degree-days, mean temperature of the warmest month, average growing season temperature and similar climate variables, have been developed to describe viticulture and wine production. The independent weather variables are used to model climate effects on wine vintage quality, the dependent variable, the latter generally represented by wine

ratings and comments by sommeliers as well as alcohol content or amount of phenols/ other compounds that are deemed to be responsible for releasing specific aroma (i.e., pine apple, peach or plum), colour and flavour unique to the wine style, also with a few specifics to the vintage, for example, Chardonnay produced from northern New Zealand with grapes of warmer ripening temperatures are described to be of containing more of tropical, pine apple and peach flavours whereas those of cooler temperatures with lime and lemon like flavours. The New Zealand wine regions and wine styles produced based on the warmest month temperature ranges (grape ripening for producing premium quality wine) are presented in figure 1 based on data compiled from [5].

A. Modelling climate change effects on wine quality

All recent key approaches consisting of statistical data analysis to modelling climate effects in grapevine phenology and wine quality reviewed for this study revealed the data required for such a study to be over five decades and it is also clear that researchers have made considerable efforts when integrating data from different sources for this purpose.

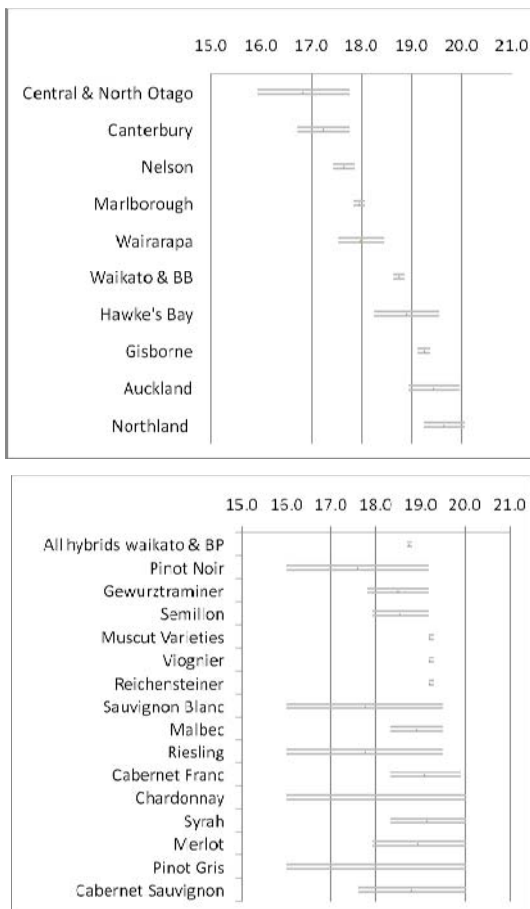


Figure 1. a & b: New Zealand wine regional and wine appellation based groupings based on the warmest month/ grapevine maturity (Oct-April) average temperature (°C). Pinot Noir, Sauvignon Blanc, Chardonnay and Pinot Gris varieties in New Zealand seem to be thriving well under unusually diverse temperature ranges. BB: Bay of Plenty.

For example, in [6] for compiling **grapevine phenology** of *floraison*, *veraison*, and harvest dates of Bordeaux (region) wines in France, the authors used reference vineyards for 1952 to 1997 time span and calculated the average dates (averaged between châteaux and variety). However, for establishing budburst dates used simple models based on an observation that in most viticulture regions, on average, budburst starts to occur when the mean daily temperature exceeds 10°C for five consecutive days and six during a cold spell.

Meanwhile, **vintage ratings** for the study were compiled from a wide variety of sources to cover the whole time period analysed stating that any qualitative assessment of a vintage to be a generalisation, with ratings commonly seen as serving as the industry-wide benchmark for comparing vintages. The overall vintage quality rating for the reference vineyards (1940-1995) used was scaled from 1 to 7 with 1 being a terrible year and 7 an exceptionally good year. In many recent studies, even though quality ratings are considered to be inherently subjective as the rating in general does not consider variations in quality among the individual châteaux, the relative measure, when tabulated in a consistent manner is presumed to give a reliable quality variable to assess general climatic influences.

Finally, for the **climate data**, the authors used meteorological recordings from a Bordeaux station, METEO-France weather for 1949 to 1997, the station was chosen simply because it had not been relocated over the period of the study. The data used consisted of daily observations of maximum temperature (T_{max}), minimum temperature (T_{min}), hours of insolation, and precipitation. Using these general climate parameters and other variables commonly used in viticulture studies for the region, such as The Sum of Average Temperatures ($SAT = (T_{max} + T_{min})/2$), Estimated Potential evapotranspiration ($PET = SAT - precipitation$) as well as the number of days with extreme cold were derived. The extreme cold was calculated by adding the numbers of days with minimum temperatures less than $-2.5^{\circ}C$ and less than $-10^{\circ}C$, with two variables for the assessment of both *moderate* and *extreme* cold events. Similarly, the numbers of days with maximum temperatures greater than $25^{\circ}C$ as well as $30^{\circ}C$ were derived for the assessment of both *moderate* and *extreme* warm events.

The associations between viticulture data (on phenology, yield, must composition, and vintage ratings), and climate data (summed by phenological intervals), both representing the dependent and independent variables respectively, were then analysed with multiple regression procedures.

In another study [7], authors modelled the response of yields to temperature and precipitation changes in some perennial crops with statistical models developed from 1980-2003 records of state-wide yield, monthly average temperature (minimum and maximum) and rainfall data.

Interestingly, in [8], the authors used an exploratory data analysis and then multiple regressions to develop a suitable model to predict the annual yield for different crops, namely,

Grapes (wine & table), Lettuce, Almonds, Strawberries, Hay, Oranges, Cotton, Tomatoes (processing), Walnuts, Avocados and Pistachios. The initial exploratory analysis with independent regressions was performed to select appropriate predictor climate variables, of these two most important climate averages (daily/ monthly) were later used to run multiple regressions to develop the best fit model with three climate variables alone for each of the crops analysed.

III. THE DATA, ISSUES AND METHODOLOGY

In this study, data on local weather conditions gathered and disseminated by NIWA [9] via its web portal is analysed with yield data from a vineyard in northern New Zealand.

A. Climate and wine quality data sets

Climate data for 1996 June - 2009 Dec, gathered from a nearby weather station run by NIWA was extracted from its web portal. Grapevine yield data provided by the winery covers 1998-2006 harvest. This 12 year yield was classified as *high*, *low* and *moderate* by the winemaker. The yield data was integrated with climate monthly average data for analysis with data mining and statistical methods to find associations between the dependent and independent variables; monthly weather variables, total rain (mm), growing degree days (base 10°C), frost day (occurrence), mean of daily, maximum and minimum temperatures, are the dependent and vineyard yield is the independent variable. A scatter plot of the vineyard data is given in Table 1.

B. The methodology

Data on grapevine yield and monthly variability in climatic conditions described above is collectively analysed using; 1) an unsupervised algorithmic artificial neural network (Kohonen self-organised map method) based data mining techniques, 2) decision trees and finally 3) with discriminant analysis to gain more understanding on the associations between the two sets of variables.

IV. THE RESULTS

A. Data mining (clustering) results

The SOM of yield properties and its three cluster profiles (figures 2 a-c) show the distinction in the properties of the yield classes *high*, *moderate* and *low*. In the profiles (figure 2 c), annual yield is slightly higher 9.83 tons/ha in *high* and Brix is slightly high in *low* class. However, in the correlation coefficient matrix, pH shows the highest correlation among the properties analysed with the yield class (Table 1).

TABLE I. CORRELATION COEFFICIENT

	Yieldt/ha	Brix	Acid	pH	rateNum
Yieldt/ha	1				
Brix	0.050256	1			
Acid	-0.45827	-0.3148	1		
pH	0.571357	0.44143	-0.59227	1	
rateNum	0.698016	0.012479	-0.03556	0.473365	1

Table showing the correlation coefficient between vineyard yield attributes and *high*, *low* and *moderate* yield years classified by the winemaker in northern New Zealand

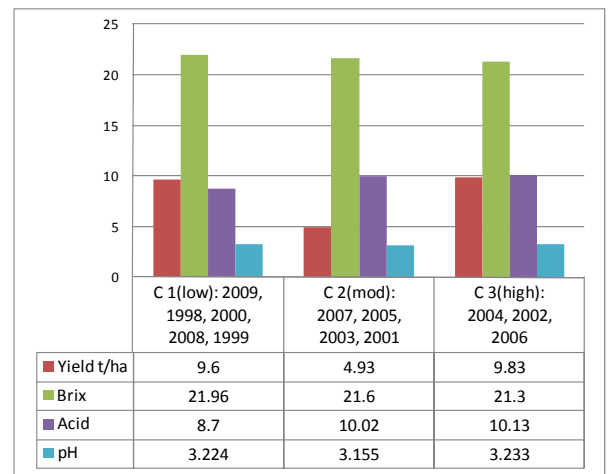
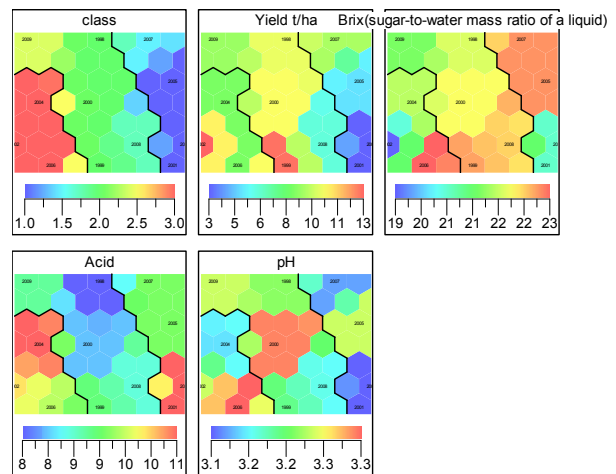
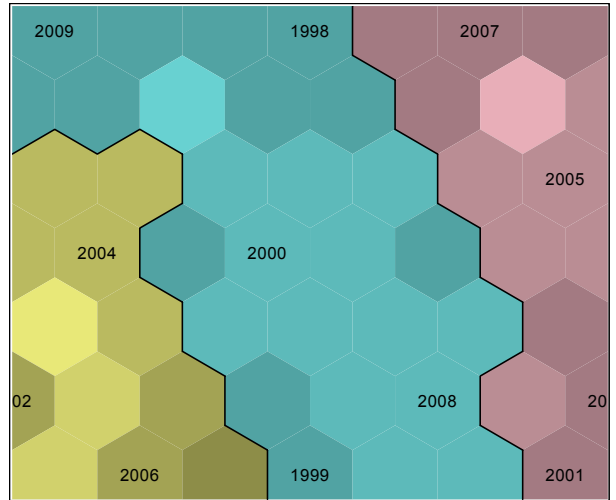


Figure 2. a: Three cluster SOM created with all yield and monthly climate data. The yield year class (*low-1*, *moderate-2* and *high-3*) was given a higher priority to group the yield and weather attributes based on the year classes. b: Yield related components of SOM (figure a). c: *low*, *moderate* (mod) and *high* yield year profiles. Of these, yield and acid of *high* years have higher values. However, variability in Brix and pH (sugar-to-water mass of a liquid) for the three classes are not so distinguishable.

Interestingly, SOM results suggest that some of the monthly weather variables analysed (rainfall, mean, minimum and maximum temperatures) have a greater impact on the grapevine growth, phenology and hence on wine vintage quality as corroborated by local winegrower knowledge relating to the particular vineyard. Of the monthly figures, July and March temperatures are found to be critical to grapevine annual yield. Low total monthly rainfall and temperatures (mean, max and min) in July and March favour *high* yield, the former affect the dormancy while the latter interfere with the berry ripening process. For example, the graph of the SOM cluster profiles relating to the three classes (figure 3a) shows that more rain i.e., 105 mm monthly rainfall in March (during harvest) and 226.7 mm in July (during dormancy) leads to *low* yield. Similarly, correlations between temperature and the three classes as well during dormancy and before harvest can be established (figures 4 a and b) for the vineyard.

Test carried out using the decision tree rules listed in table II, produced results 100% accurate in predicting the yield classes of this vineyard (table III).

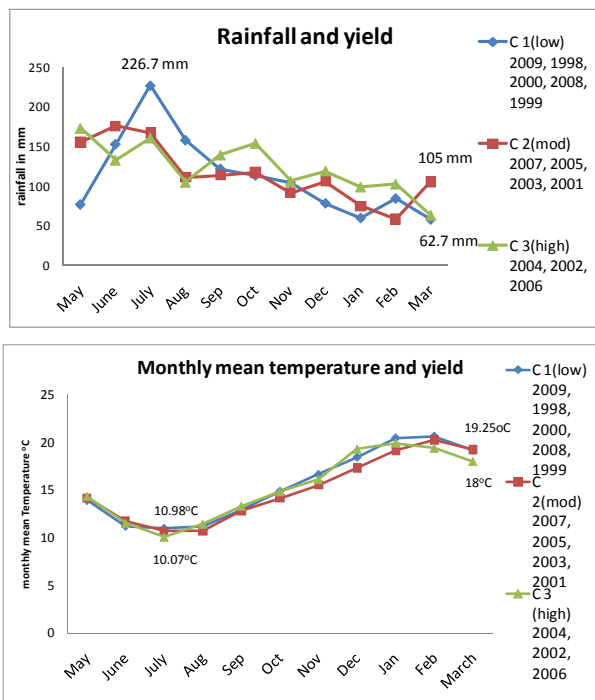


Figure 3. Graphs of SOM cluster profiles show associations between monthly weather conditions and annual yield at this vineyard in north of Auckland. a: Graph showing monthly total rainfall against three yield classes, 226.7 mm rainfall in July and 105 mm rainfall in March are associated with *low* yield years (C2). b: 18 °C March mean temperature is linked with *high* and 19.18 °C and 19.25 are linked to *low* and *moderate* yield years respectively.

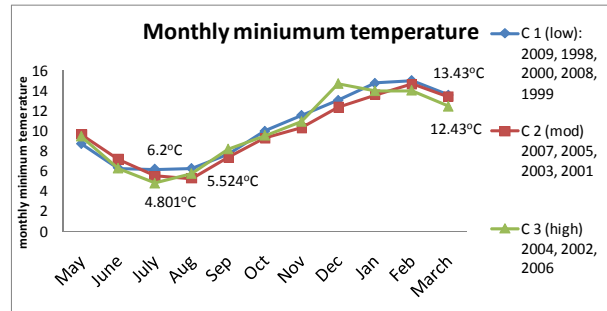
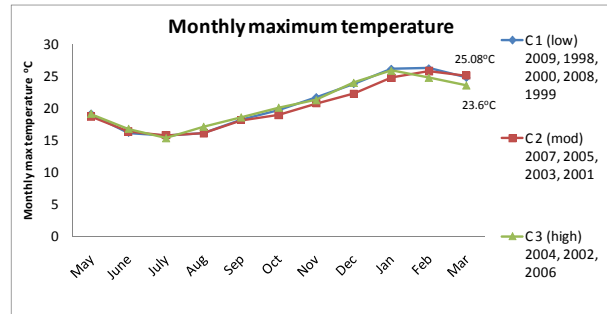


Figure 4. Graphs of SOM cluster profiles of monthly mean maximum (b: minimum) show the temperatures associated with *low*, *moderate* and *high* yield years for the vineyard.

B. Decision trees

The decision tree created with software ISEE (www.rulequest.com/see5-info.html) show that monthly temperatures of March maximum and November minimum to be associated with yield year classes. The following are the rules derived from this decision tree analysis:

- Rule 1: March max temperature > 24.5°C and November min temperature > 10.5 °C → class *moderate*.
- Rule 2: March max temperature > 24.5 and November min temperature ≤ 10.5 °C → class *low*.
- Rule 3: March max temperature ≤ 24 °C → class *high*

TABLE II. DECISION TREE RESULTS

```

Rules:

Rule 0/1: (3, lift 2.4)
  MarmaxT > 24.2
  NovminT > 10.5
  -> class moderate [0.800]

Rule 0/2: (3, lift 2.4)
  MarmaxT > 24.2
  NovminT <= 10.5
  -> class low [0.800]

Rule 0/3: (3, lift 2.4)
  MarmaxT <= 24.2
  -> class high [0.800]

```

TABLE III. DECISION TREE EVALUATION RESULTS

```

Evaluation on training data (9 cases):
Rules
-----
No      Errors
3      0 (0.0%)  <<
class   (a)    (b)    (c)    <-classified as
Mod     3
low     3
high    3    (c) :

Attribute usage:
100% MarmaxT
67%  NovminT
    
```

C. Statistical data analysis and results

Graphs of low, moderate, high yield year and monthly weather data show the months and their respective temperature ranges associated with the yield year classes. For instance, the graph (figure 5) of all three low yield year (2001, 2003 and 2005) monthly mean temperatures show the months and the ranges that are critical to grapevine yield in this vineyard. Even though November temperature does not harm, December mean temperate seems to have a major impact on the annual yield, anything over or under 20°C has an adverse impact on the yield as seen in figures 6 a and b.

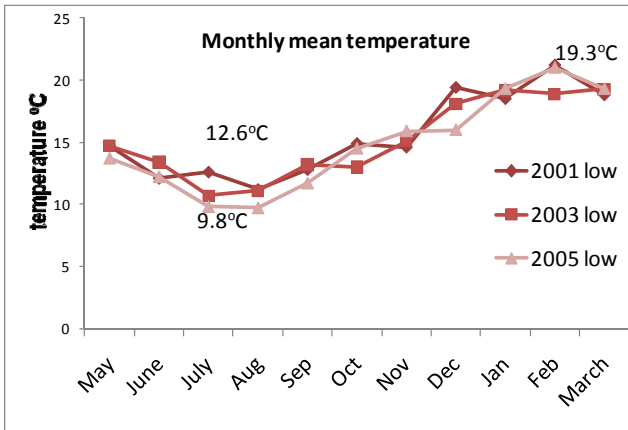


Figure 5. Monthly mean temperature (°C) of all three low yield years as described by the winemaker.

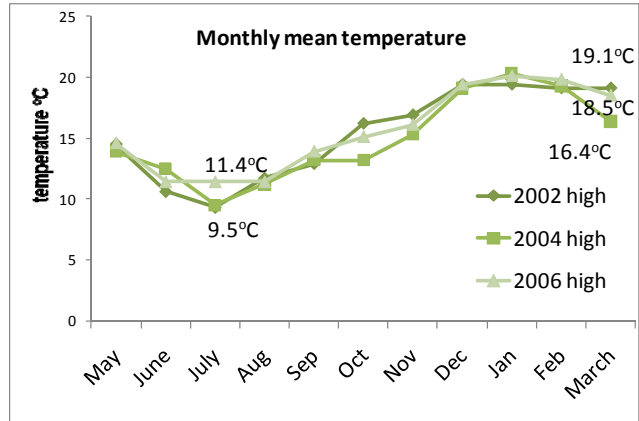
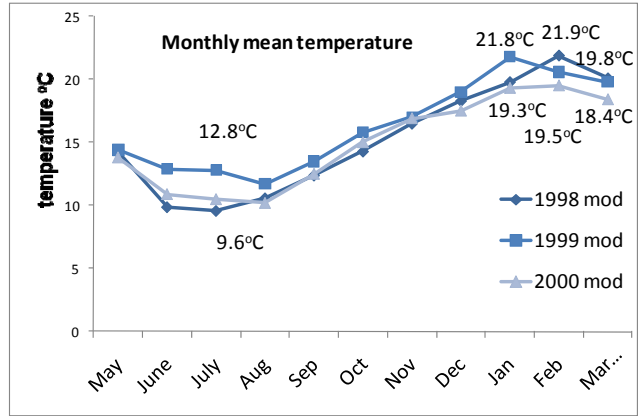


Figure 6. Graphs showing the real monthly mean temperature of a: moderate b: high yield years classified by the wine-maker of Kumeu River Wines in north of Auckland, New Zealand.

Finally, discriminant analysis that carried out to verify the findings of explorative data analysis (data mining by clustering) and results are discussed in this section

The discriminate model results point to December total rainfall being negatively correlated to yield (Table 1V). The reason for this could be that more rain during this period in general affects pollination. Meanwhile, January and March maximum temperatures show positive correlations as high temperatures during this time favour flowering and grape ripening. November minimum relates to positive shoot growth and October frost affects budburst significantly even leading total loss in yield.

Overall, it then appears that data mining techniques as well as statistical methods, such as discriminant analysis could be useful in determining the weather variable dependencies that are significant to a vineyard yield. The results of this research not only reemphasise the anecdotal evidence but gives the precise weather factors and their respective ranges that are critical to grapevine growth, phenology and vineyard yield.

TABLE IV. FIGURE 6: A DISCRIMINANT ANALYSIS MODEL RESULTS AND VALIDATION ACCURACY.

Classification Results ^{b,c}					Eigenvalues						
Original	ratecode	Count	Predicted Group Membership				Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
			1	2	3	Total					
Original	Count	1	3	0	0	3	1	6926.397 ^a	99.9	99.9	1.000
		2	0	3	0	3					
		3	0	0	3	3	2	5.813 ^a	.1	100.0	.924
		Ungrouped cases	0	2	1	3					
	%	1	100.0	.0	.0	100.0					
2	.0	100.0	.0	100.0							
3	.0	.0	100.0	100.0							
	Ungrouped cases	.0	66.7	33.3	100.0						
Cross-validated ^a	Count	1	3	0	0	3	1 through 2	.000	43.048	10	.000
		2	0	3	0	3					
		3	0	0	3	3	2	.147	7.675	4	.104
		Ungrouped cases	.0	.0	.0	100.0					
	%	1	100.0	.0	.0	100.0					
2	.0	100.0	.0	100.0							
3	.0	.0	100.0	100.0							

a. First 2 canonical discriminant functions were used in the analysis.
 b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
 c. 100.0% of original grouped cases correctly classified.
 d. 100.0% of cross-validated grouped cases correctly classified.

December total rainfall shows negative correlation as more rain during this period affects pollination. January and March maximum temperatures show positive correlations as high temperature favour grape ripening. November minimum relates to positive shoot growth and October frost affects budburst significantly even leading total loss in vineyard yield.

V. CONCLUSIONS

Literature reveals that conventional climate data analysis methods and models can only be applied to studying seasonal climate effects on grapevine phenology and wine vintage quality at regional scales with data covering over five decades, this is a major constraint especially in cases where there is no data covering such longer periods, for studying the climate effects at micro scales using macro climate data. The paper looked at data mining approaches (Kohonen SOM based clustering, ISEE sw based decision tree and statistical methods, i.e., canonical discriminant methods) to modelling wine vintage quality using vineyard yield and historical climate data, the latter 10 year old macro scale gathered at a national meteorological station, and the results show potential for use with live data from GRC wireless sensor network of nodes with weather monitoring devices installed in vineyards [10] for yield prediction purposes.

VI. FUTURE WORK

Further research with SOM based and other data mining as well as regression techniques for predicting vineyard yield using monthly climate variability is presented in [11].

ACKNOWLEDGMENT

The authors wish to acknowledge winemaker Michael Brakvich of Kumeu River Wines and Peter Sumich of Auckland University of Technology for their support.

REFERENCES

[1] Jones, G. V. 2005. "How Hot Is Too Hot?". Vol. Feb. 2005, Wine Business Monthly pp. 1-4 www.winebusiness.com

[2] van Leeuwen, Cornelis, Friant, Philippe, Chone, Xavier, Tregcoat, Olivier, Koundouras, Stephanos, Dubourdiou, Denis. Influence of Climate, Soil, and Cultivar on Terroir. 2004, Am. J. Enol. Vitic. 2004 55 pp. 207-217.

[3] Ashenfelter, O., Ashmore, D., and Lalonde, R.,. Bordeaux Wine Vintage Quality and the Weather. 1995, Chance vol 8 No. 4 1995 pp.7-14.

[4] Jones, G V. Climate change: observations, projections, and general implications for viticulture and wine production. 2007, Practical Winery & Vineyard. pp. 44-64.

[5] Cooper, M. *Wine Atlas of New Zealand (2nd Ed)*. Hodder Moa. ISBN: 1869710916 pp. 408, 2008.

[6] Jones, G V, and Davis, R E, Climate Influences on Grapevine Phenology, Grape Composition, and Wine Production and Quality for Bordeaux, France. Am. J. Enol. Vitic., Vol. 51, No.3, 2000 pp. 249-261.

[7] Lobell, David B, et al. Impacts of Future Climate Change on California Perennial Crop Yields. Model projections with climate and crop uncertainties. 2006, Agricultural and Forest Meteorology 141 (2006) pp. 208-218.

[8] Lobell D B, Cahill K N and Field C B. Historical effects of temperature and precipitation on California crop yields. Springer, 2007, Climatic Change (2007) 81 pp. 187-203.

[9] National Institute of Water and Atmosphere. [Online] <http://cliflo.niwa.co.nz>.

[10] Geoinformatics Centre. [Online] www.geoinformatics.org/publications.aspx

[11] Shanmuganathan S, Sallis P, Narayanan A, Modelling the seasonal climate effects on grapevine yield at different spatial and unconventional temporal scales. Proc. International Environmental Modelling and Software Society (iEMSS) 2010 International Congress on Environmental Modelling and Software Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.) www.iemss.org/iemss2010/index.php?n=Main.Proceedings (in press)

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
DecprainK	-23.115	-.670
JanmaxT	13.204	.349
MarmaxT	3.688	1.898
NowminT	25.073	-.160
OctpFrostK	-11.517	1.269