AgriRisk: Bayesian Network models. Data, analyses, and models

Timothy Lynam

Reflecting Society

March 15th, 2018

Risk Proofing Nova Scotia Agriculture: A Risk Assessment System Pilot (AgriRisk)
Nova Scotia Federation of Agriculture would like to recognize the collaborative relationships that exist among Agriculture and Agri-Food Canada and the Nova Scotia Departments of Agriculture and Environment.

Acknowledgements

The author gratefully acknowledges the many dimensions of support provided by Nova Scotia Federation of Agriculture (NSFA), Nova Scotia Environment (NSE), Nova Scotia Department of Agriculture (NSDA) and Agriculture and Agri-Food Canada (AAFC). Many people have contributed to the development of the materials that went in to the scenarios workshops and thence the workshops themselves. I am grateful to them all. In particular I am grateful to the many contributions made by Francisco Diez, Rachael Cheverie and Amy Sangster of Perennia, thank you. Joanne Moran (Grape Growers Association of Nova Scotia) and Jerry White (Winery Association of Nova Scotia) have been tremendously helpful and supportive, thank you both. The growers and wine makers of Nova Scotia have been extraordinarily generous with their time and knowledge. Thank you, it makes all the difference in the world. The support provided by Heather MacDougall, Stephen Ford and Ritu Gautam of NSLC has been unprecedented and is greatly appreciated, thank you. Mark MacPherson, Jen Graham and Meredith Flannery have been wonderfully supportive and constructive in guiding the project to fruition, thank you.

Table of Contents

Table of figures

Tables

Introduction

This document seeks to present a high level overview of the data and analytical procedures used to develop the Bayesian Network (BN) models for the AgriRisk project and then use that overview to explore in more detail the components of the model, the data that was used to build each component and the analyses and analytical procedures that were used to develop and parameterise each component. The report does not seek to describe BNs or how to use them. It is simply a description of the data and analytical procedures developed to build the AgriRisk BNs.

In the next section of the report an overview of the BN Models is presented to highlight the different components and their linkages. Also presented in this section are the results of a sensitivity analysis conducted with the integrated BN model which is then useful for highlighting important gaps in data or knowledge.

Following the overview section, the four core components of the BN (grape growing, wine making, sales / distribution and consumption) are described in more detail.

Finally, a summary of the key data and modelling gaps, opportunities and challenges is presented and discussed with a view to informing future activities designed to take over from where AgriRisk left off.

Model overview

At a very aggregated level the BN network model for the AgriRisk project comprises four linked components: the grape growing component; the wine making component; the sales and distribution component and the consumption component (see figure below). Each of these four components has a number of elements and relationships. The model was conceived as being developed using Norsys's Netica software[1]. The BN model was developed using Netica but with much of the data analyses conducted using R (R Core Team, 2017).

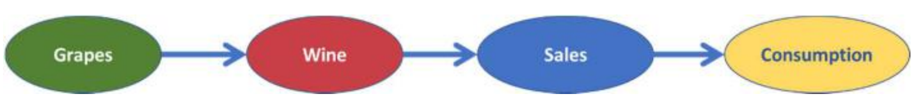

Figure 1. Overview of the four components of the integrated BN model of the grape and wine industry of Nova Scotia.

The model was designed to represent the value chain of the grape and wine industry in Nova Scotia. This meant that the model needed to represent the critical dependencies that exist between components: for example, without good grapes wine producers cannot produce good wine, sales are likely to decline, and consumers will be unhappy. These dependencies go both ways and the BN modelling approach is ideally suited to exploring backwards, as well as forwards

[1] http://www.norsys.com/netica.html

linkages. The current (March 2018) full model has demand components for Nova Scotia wine and wine bottled in Nova Scotia from local and imported juice; grape and wine production components and industry economics components based on the economics of grape growing (Figure 2). Sensitivity analyses of the network based on two variables (Industry gross revenue ($/year) and NS wine production (litres / year)) were undertaken and the results are shown in Figure 3 and Figure 4. Grape yield on a per hectare basis was a dominant source of uncertainty for total wine supply where the revenue for all growers was the dominant source of uncertainty for total industry gross revenue from wine growing.

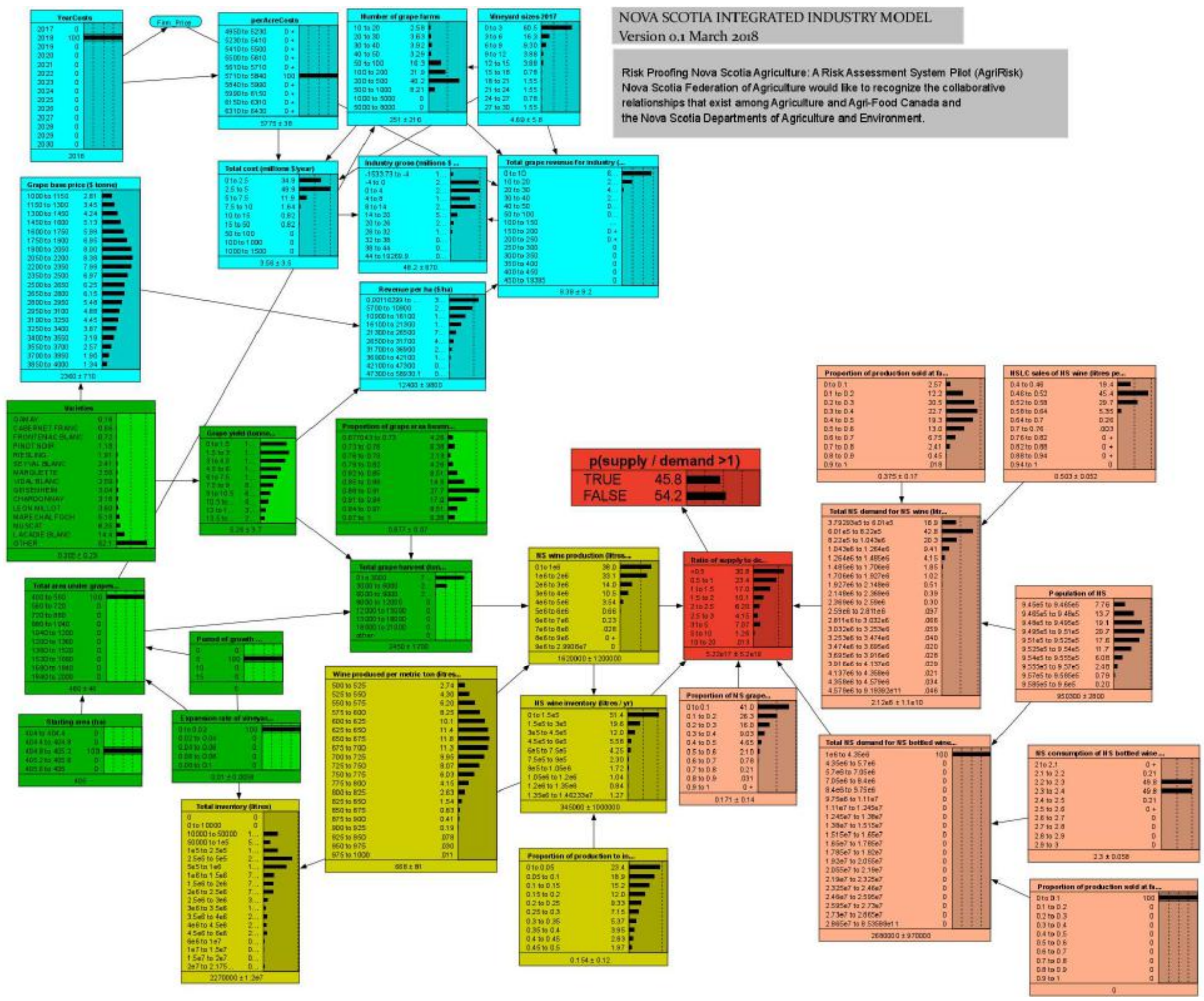

Figure 2. Overview of the full BN model as of March 2018. Blue nodes are the economic variables; green nodes the grape growing variables; yellow nodes the wine production variables; light orange the sales variables and red the industry supply / demand variables.

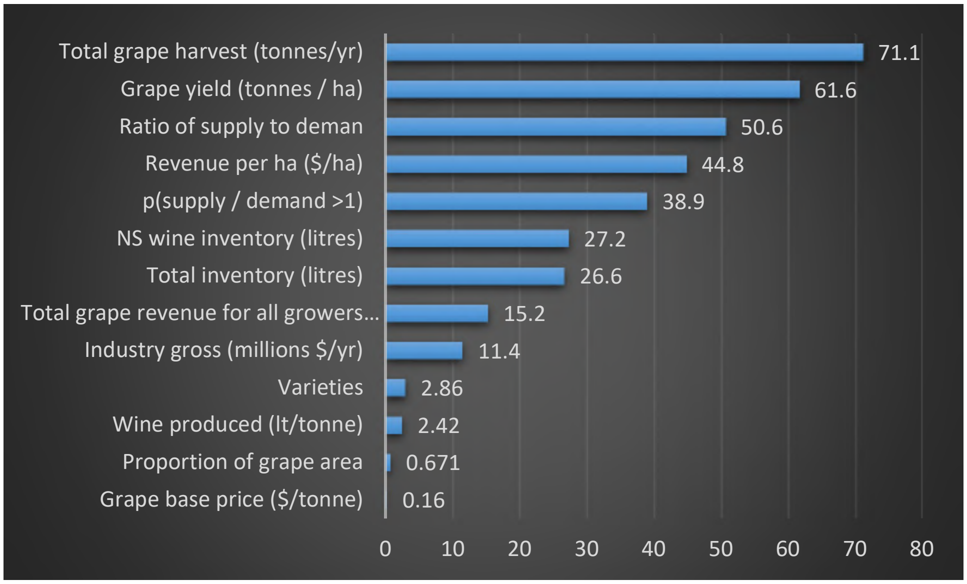

Figure 3. Sensitivity of total wine production to key variables in the network. Measure based on mutual information: how much would we learn about NS wine production from one more finding at each of the network nodes. Conditionally independent nodes (i.e. where the sensitivity analyses results were zero) are not shown.

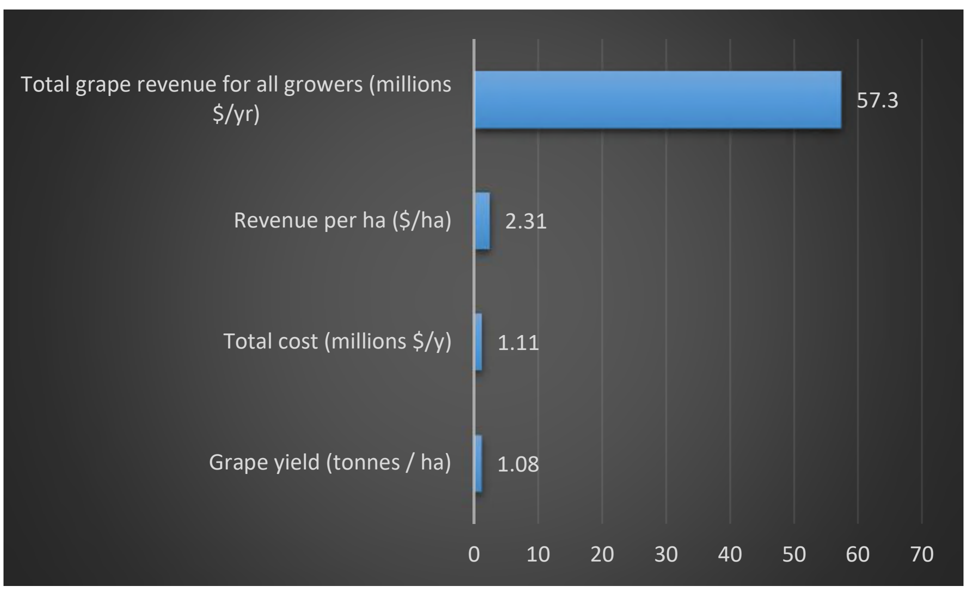

Figure 4. Sensitivity of industry gross revenue to additional findings key variables in the network. Measure based on mutual information: how much would we learn about NS wine production from one more finding at each of the network nodes. Conditionally independent nodes (i.e. where the sensitivity analyses results were zero) are not shown.

In the next section of the report each of these four sections of the overall model is expanded and the data, analyses and challenges of building and using those elements are described.

Component models and data

Grape growing

The original grape model was conceived as addressing environmental (e.g. climate, soil, pest and disease), management (pruning, leaf thinning, rootstock selection) and economic (prices, costs) to form an integrated grape growing model (Figure 5). Unfortunately, as the project began the process of assembling existing data it became apparent that the data to build this idealised model was not available.

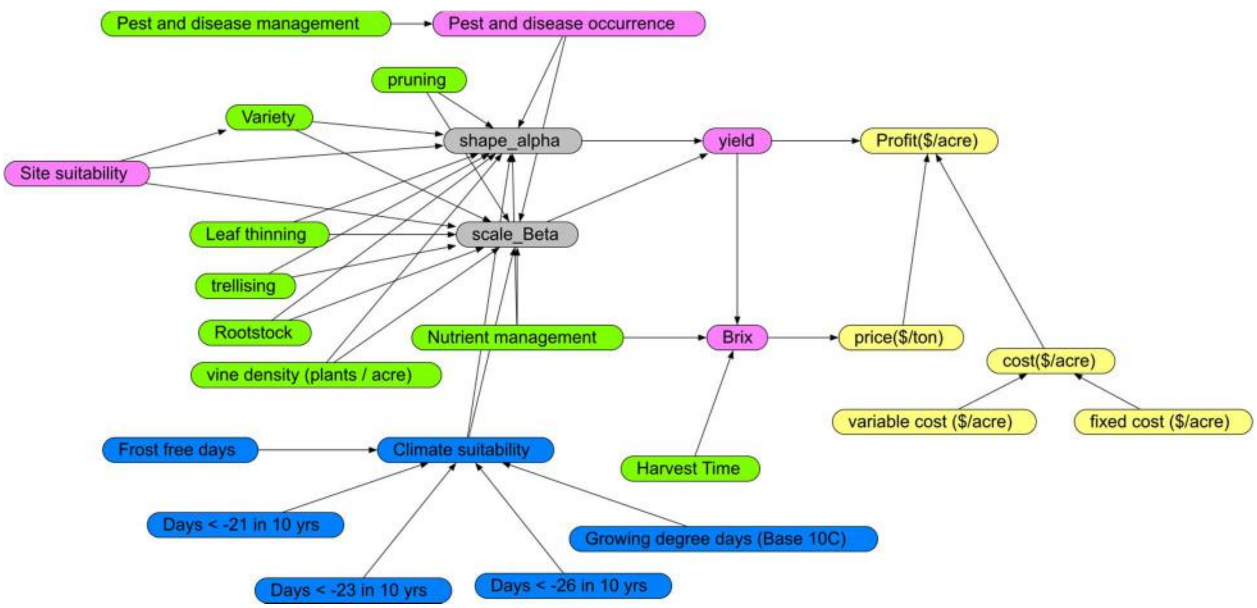

Figure 5. Original conception of the grape growing model, April 2017.

The following, largely data driven decisions were made:

1) The climate suitability components were moved to the grape suitability mapping and thus site suitability was taken as a given or input data source;
2) There was insufficient yield data that could be linked to climate or site suitability, so the model had to take yield as an aggregate probability distribution for a variety (14 varieties were selected to work with);
3) The data on the relationships between yield and brix was insufficient to derive reliable relationships
4) We had not data on pest occurrences and their consequences, only on three diseases but these data came from experiments conducted in Quebec;
5) We had insufficient data to derive relationships between management actions and grape yield or quality outcomes.

As such the primary grape growing model[2] evolved to a much simpler set of relationships as depicted in Figure 6 below. The blue nodes in the model represent economic relationships and the green nodes represent the grape growing components. Initially the model was developed using theoretical distributions fit to the price and yield data, but the data were not sufficient to support reliable distribution estimation and the resulting distributions were hugely uncertain.

Data on yields was derived from the grower and winery survey conducted by the AgriRisk project in conjunction with yield values reported to NSFA by wineries and provided to AgriRisk by NSDA (with all results anonymised). The grower data comprised 279 records of yield and that from NSDA comprised 475 records to provide 754 records in total of which 576 were for the 14 selected varieties. The yield data were unevenly distributed across years and varieties with an increasing number of values available through time (Figure 7).

Yield relationships for the select varieties were developed by fitting a Bayesian general linear model (GLM) to the yield data. The analyses was done using package brms (Burkner, 2017) in R (R Core Team, 2017). GLM's of the form:

fit1<-brm(yield_t_ha ~ variety -1, data = df4, family="XXX")

were fit to the data, where XXX included the following distributional forms: Weibull, gamma, gaussian, lognormal and exponential. Model performance was evaluated using leave one out validation (LOO) with the best model being the Weibull (but it was only marginally better than either the gamma or exponential models).

Probably as a consequence of the small number of samples for some varieties some of the model fits were highly uncertain (Table 1, Table 2).

Table 1. Summary of the GLM fit to yield data.

```
Family: weibull(log)
Formula: yield_t_ha ~ variety
- 1
   Data: df4 (Number of observations:
275)
Samples: 4 chains, each with iter = 2000; warmup =
1000; thin = 1;
         total post-warmup
samples = 4000
    ICs: LOO = NA; WAIC = NA;
R2 = NA

Population-Level
Effects:
```

	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
varietyCHARDONNAY	0.9	0.2	0.5	1.32	4000	1

[2] Several grape growing models have been developed but all were based on this primary model.

varietyFRONTENAC	2.31	0.32	1.72	2.95	4000	1
varietyGAMAY	1.71	0.79	0.41	3.55	4000	1
varietyGEISENHEIM	1.87	0.33	1.25	2.57	4000	1
varietyLACADIE_BLANC	1.6	0.18	1.26	1.96	2413	1
varietyLEON_MILLOT	1.54	0.21	1.13	1.97	2963	1
varietyMARECHAL_FOCH	2.12	0.36	1.47	2.88	4000	1
varietyMARQUETTE	1.88	0.41	1.15	2.76	4000	1
varietyMUSCAT	1.86	0.3	1.3	2.46	3266	1
varietyOTHER	2.01	0.16	1.71	2.32	2104	1
varietyPINOT_NOIR	1.32	0.24	0.88	1.82	4000	1
varietyRIESLING	1.91	0.32	1.31	2.56	4000	1
varietySEYVAL_BLANC	2.54	0.63	1.43	3.93	3676	1
varietyVIDAL	1.58	0.43	0.81	2.52	4000	1

Family Specific Parameters:

	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
shape	1.24	0.06	1.12	1.35	1570	1

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample is a crude measure of effective sample size, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).

Table 2. Parameter estimates for GLM of yield by variety.

Variable	Estimate	Est.Error	2.5%ile	97.5%ile
varietyCHARDONNAY	2.455	1.227	1.647	3.725
varietyFRONTENAC	10.099	1.374	5.604	19.202
varietyGAMAY	5.525	2.194	1.503	34.834
varietyGEISENHEIM	6.493	1.394	3.507	13.126
varietyLACADIE_BLANC	4.948	1.193	3.542	7.064
varietyLEON_MILLOT	4.649	1.239	3.096	7.161
varietyMARECHAL_FOCH	8.307	1.427	4.329	17.733
varietyMARQUETTE	6.565	1.508	3.153	15.849
varietyMUSCAT	6.410	1.344	3.671	11.675
varietyOTHER	7.464	1.169	5.519	10.194
varietyPINOT_NOIR	3.757	1.276	2.404	6.153
varietyRIESLING	6.724	1.377	3.721	12.956
varietySEYVAL_BLANC	12.698	1.884	4.175	51.046
varietyVIDAL	4.872	1.538	2.257	12.457

Three versions of the grower model were developed but only the first (the base model) is described in this report. The second model uses Netica's decision modelling algorithms to make a decision model that maximises the expected gross returns based on decisions as to what varieties to grow. The third model is like the second but maximises the minimum gross return, in essence identifying the variety that minimises possible loss.

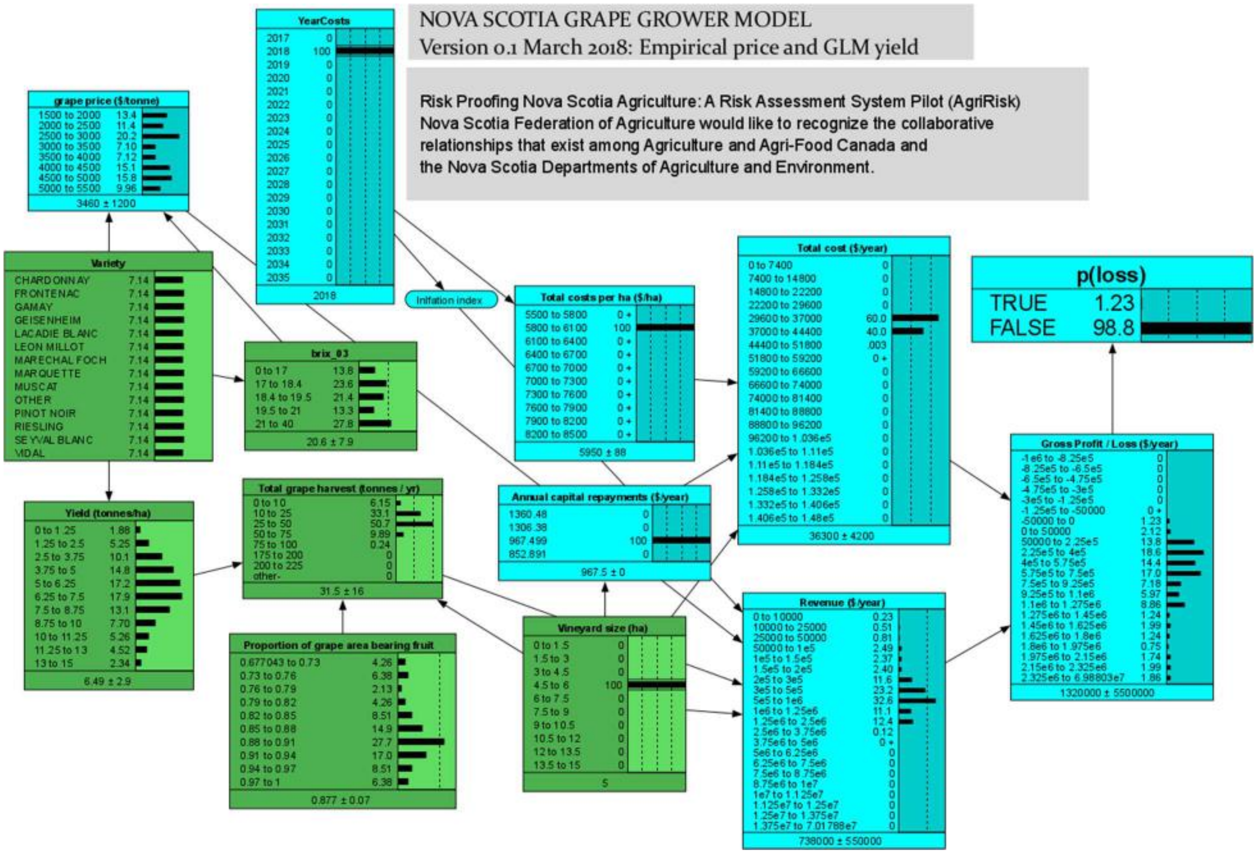

Figure 6. Overview image of March 2018 version 0.1 BN model of grape growing based on empirical data for price and yield distributions. Blue nodes are associated with the economic components of the model and green the grape growing components. See text for details.

Grape growth components

The grape growth component of the BN model was a simple set of conditional multiplications:

Total grape harvest = yield (metric tonnes per hectare) being a probability distribution for all varieties for which there was data, times vineyard area (a discretised continuous variable[3] of areas of vineyards) times the proportion of the area that was bearing fruit which was an empirical distribution developed from Statistics Canada data[4] that played the role of identifying what

[3] In Netica all continuous variables must be discretised.
[4] http://www5.statcan.gc.ca/cansim/pick-choisir?lang=eng&p2=33&id=0010009 accessed March 2018.

proportion of vineyards were either not yet producing (i.e. young vineyards) or were non-productive areas (e.g. headlands).

BN model users could select varieties or distributions around varieties (likelihoods for subsets of the varieties available). This resulted in a conditional yield distribution (conditional on variety) which was multiplied by vineyard size to yield total grape harvest. Grape price and yield nodes were derived using a tree augmented naïve Bayes (TAN) model (Friedman, Geiger, & Goldszmidt, 1997) with price as the target and using the combined grower survey and NSDA data described above. This model was then expanded with the addition of the economic components, the vineyard size multiplier and the proportion of grape bearing areas node (Figure 6).

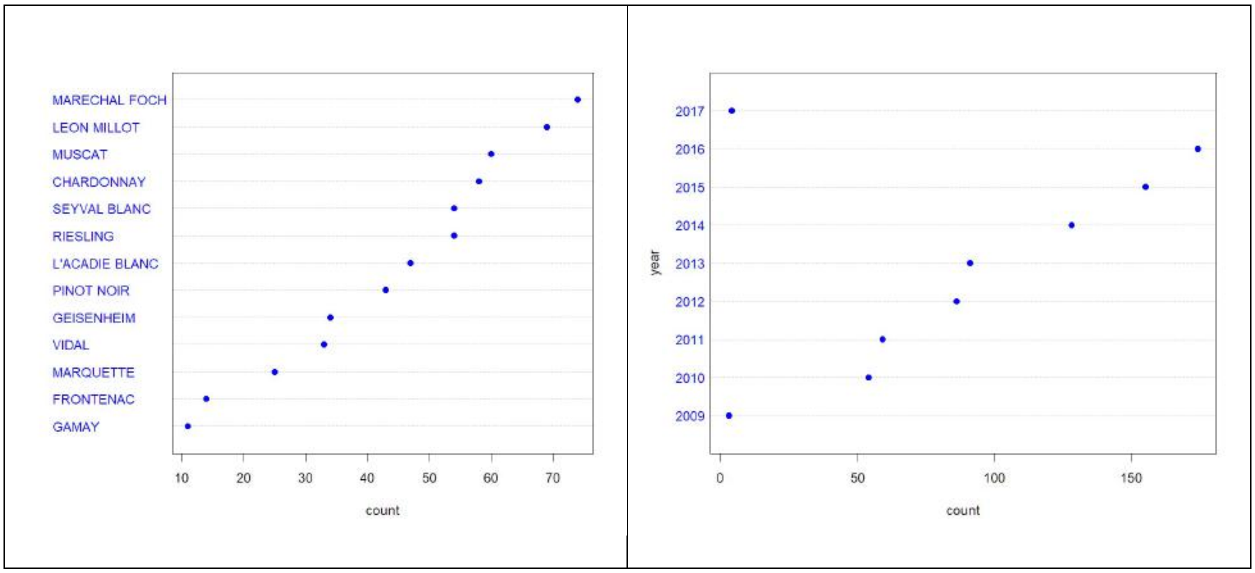

Figure 7. Counts of yield data per variety (left) and year (right) that was used for developing the BN grower models.

Economic components of grape growing

Data and functional relationships for the economics of grape growing were provided by Mike Foster of CanMac (CEL, 2018). The economic modelling took costs of production data from a NSDA report (NSDA, 2014) and used these in an econometric model to project growth in costs through time based on GDP deflator (inflation index). Revenue was modelled as yield times price for specific varieties with price taken from the requests to NSDA. Price was inflated using the same inflation index as was used for costs.

The economic analyses were based on several discrete vineyard sizes (5, 10, 20 and 30 acres) with deterministic costs derived from the original cost data, projected through to 2035 with an inflation multiplier. Although models were developed for the 14 varieties the costs were the same for all varieties. Varietal yield and prices paid to growers for grapes differed by variety. Yield was modelled as a random draw from a normal distribution (with the mean and standard deviation parameters estimated from the data provided by NSDA) and price was a constant for each variety (CEL, 2018). Revenue was estimated as yield times area with a correction for inflation.

In converting the grape economics model for use in the BN the following adjustments were made:

1) Yield as described above for the grape model was used rather than the yields provided in the economics model;
2) Costs and revenue were linearly related to acreage so the acreage component of the economic model (i.e. for 5, 10, 20 and 30 acres) was removed from the cost estimates. Within the BN area was a user adjustable metric (hectare) based value. All per area costs and revenues were converted to metric equivalents;
3) Capital costs, which varied by vineyard size, were removed from the total cost calculation and hence total costs were simply a linear function of area (hectares). Capital costs were non-linearly related to area and hence were added as a separate cost component in the BN model.
4) As noted in the description of the grower model prices were estimated using a TAN model and the data from the AgriRisk grower survey.

To estimate the prices the AgriRisk grower survey data was used. Altogether 379 values were available with which to estimate price per variety. Just as with the yield data described in the previous section the price data was unevenly distributed across time and varieties (Figure 8). As was observed for the yield data there was an increasing availability of price data through time. Brix and yield were both included in the TAN model of price, but price was not particularly sensitive to either. Price was most sensitive to variety (Figure 9).

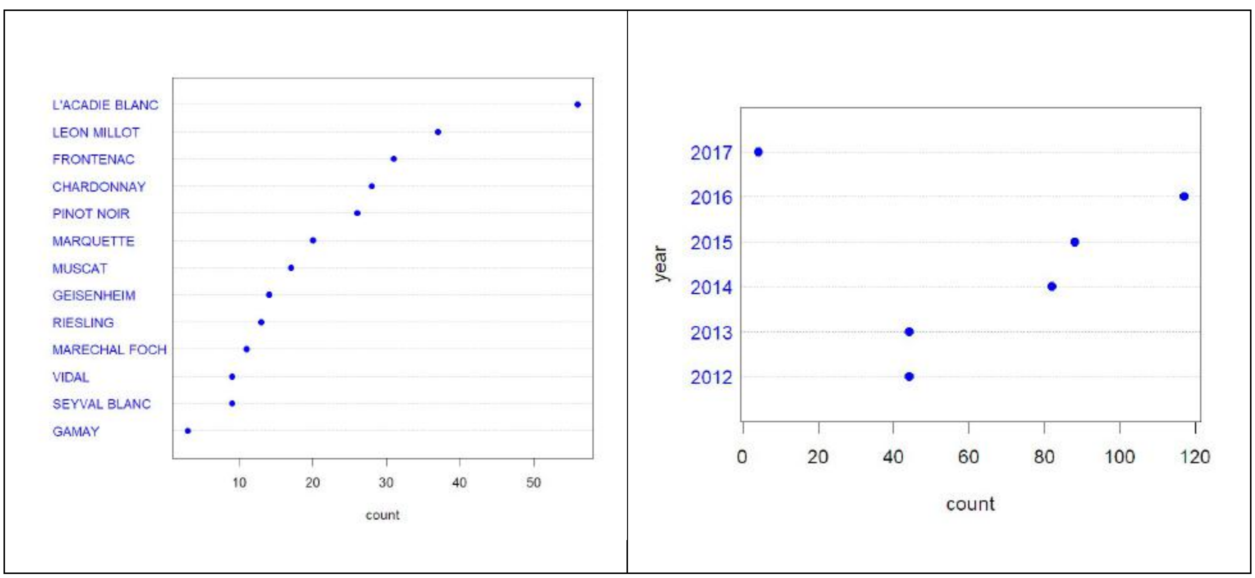

Figure 8. Counts of price data per variety (left) and year (right).

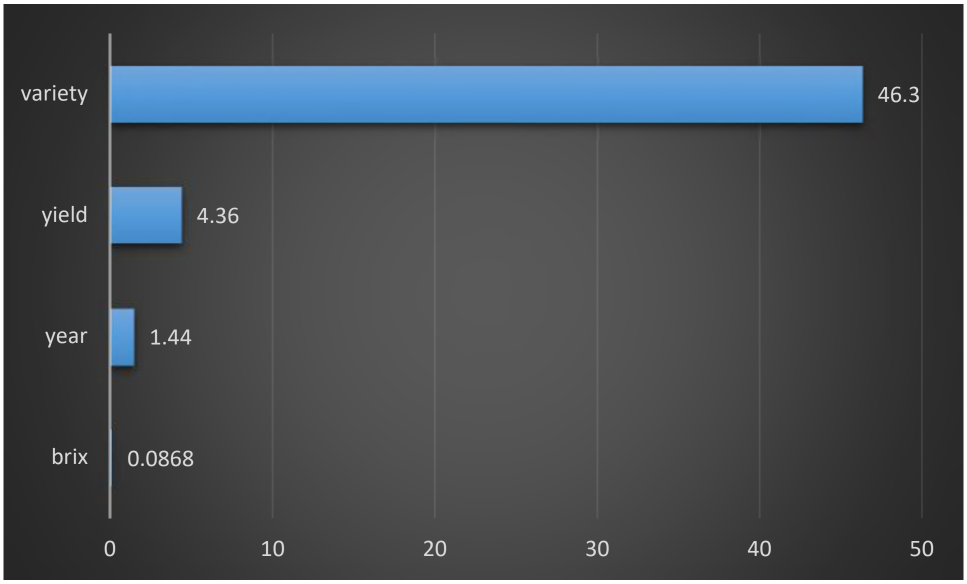

Figure 9. Sensitivity analysis (based on price) of TAN model for price. Estimates based on variance reduction: relative to the reduction in variance due to an additional finding at the target node (price) how much would variance be reduced by an additional finding (i.e. data point) at each of the other nodes in the network.

Sensitivities in the grape growing model

Sensitivity analyses were undertaken with two BN model nodes as targets: the first was the total grape harvest and the second was the gross profit or loss. When using Netica's sensitivity analysis function any evidence added to the BN is taken into consideration. Sensitivity analyses were therefore conducted with vineyard size set to 5 ha and the year for costs set to 2018, otherwise no evidence was added to the model.

The dominant source of uncertainty in total grape harvest was yield followed by variety (Figure 10). For Gross profit / loss the BN was most sensitive to additional findings as to revenue and, to a much lesser extent, variety (Figure 11).

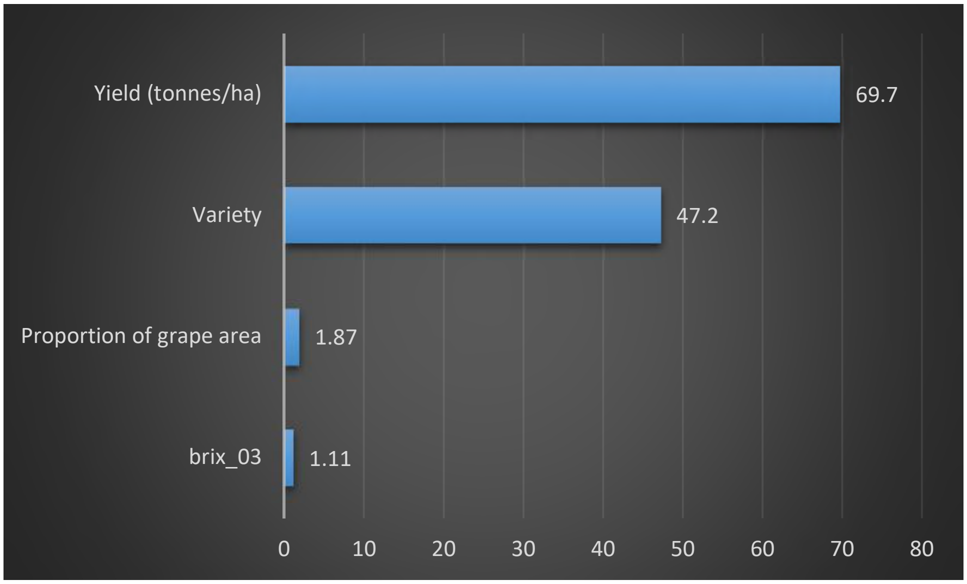

Figure 10. Sensitivity analysis (based on total grape harvest) of BN grape growing model. Estimates based on variance reduction: relative to the reduction in variance due to an additional finding at the target node (total grape harvest = 100%) how much (%) would variance be reduced by an additional finding (i.e. data point) at each of the other nodes in the network.

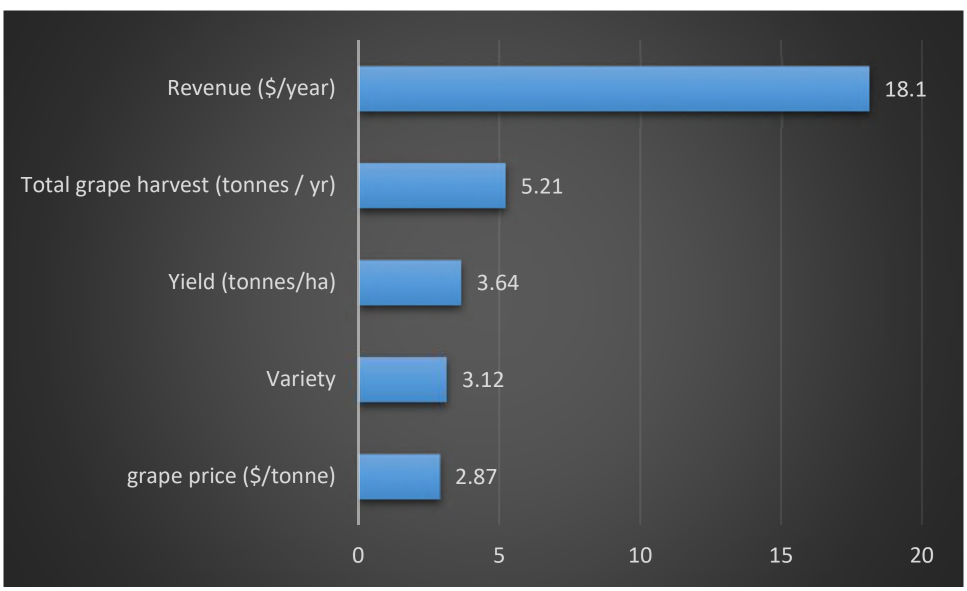

Figure 11. Sensitivity analysis (Gross Profit / Loss) of BN grape growing model. Estimates based on variance reduction: relative to the reduction in variance due to an additional finding at the target node (Gross Profit / Loss = 100%) how much (%) would variance be reduced by an additional finding (i.e. data point) at each of the other nodes in the network.

Wine making

The wine making model was developed to enable wineries to explore options for making wine from grapes they grew themselves and purchased from external growers. The basic building bloc of the winery model is the grape growing model. The winery model combines a number of growers (for illustration purposes with the current model just two) and produces select wines based on grapes from the two growers. One of the growers is assumed to be the winery itself (Figure 12). The winery represented in the current model produces L'Acadie Blanc, Pinot Noir and Chardonnay wine but this could be set to any number of wines and or blends.

The revenue generated through wine sales was estimated using retail prices from NSLC sales data for NS wines (i.e. those with 85% or more local grape content) adjusted for the NSLC mark-up (43%) to the price component in the model uses the values wineries would receive from NSLC. The price data from NSLC was fit to a simple Bayesian price estimator for each wine variety with the component learning using Netica's expectation maximisation algorithm.

The Nova Scotia Winery model estimates the total volume of wine produced per type, revenue per type and total revenue generated by the winery. Users can select grower areas for each of the two growers and identify the varieties that each grower will produce. Only the three varieties identified above are linked to winery production in the current pilot model.

The data and functions that underpin the winery model are, for the most part, the same as those described for the grower model. Ne in the winery model was the data supporting estimation of the conversion rate of tonnes of grapes to litres of wine. This data was derived through searching available online sources for estimates of wine conversion rate (i.e. litres of wine per tonne of grapes). The results were used to derive a range of values which were thereafter used to parameterise a truncated Beta Distribution which was multiplied by the range of values observed to be conversion rates (550 to 750 litres per tonne).

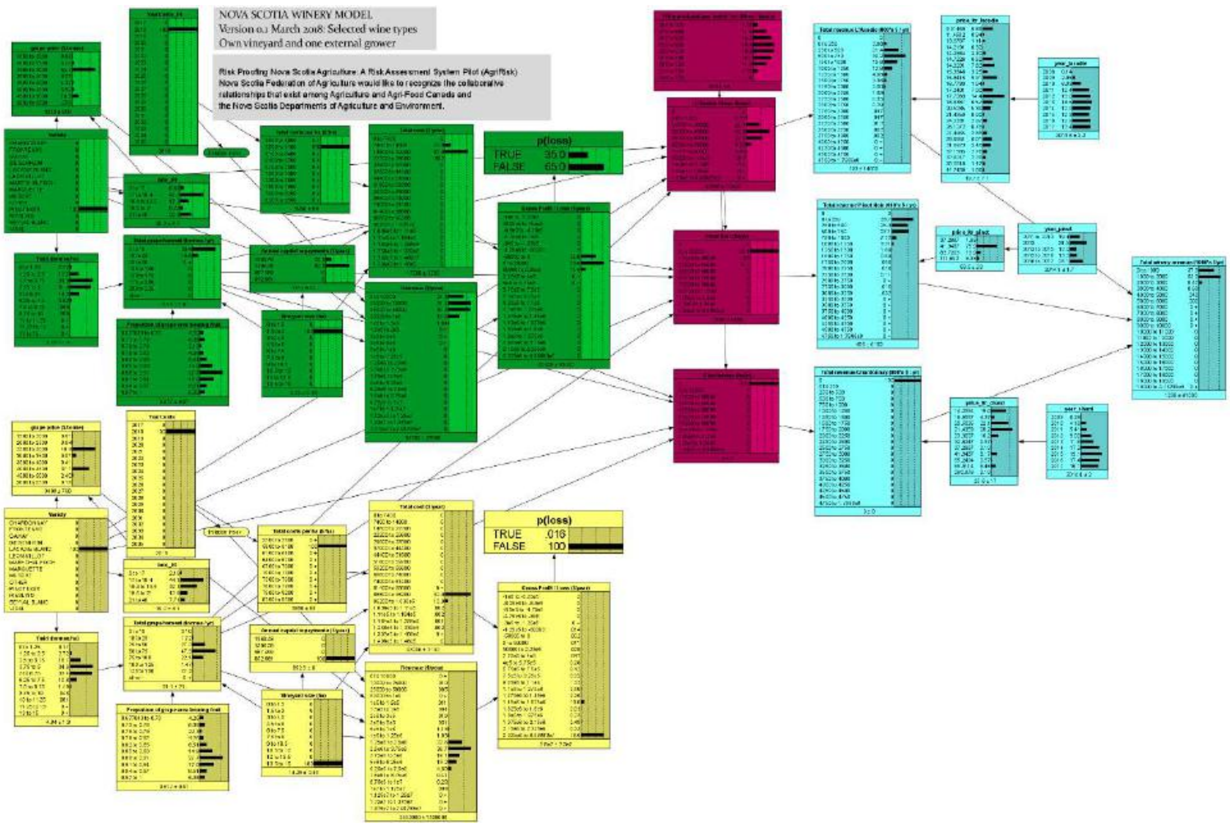

Figure 12. Image of the Nova Scotia winery BN Version 0.1. The model comprises two separate growers and a winery that (in this example) produces three types of wine (L'Acadie Blanc, Pinot Noir and Chardonnay). The green nodes are for the winery's own production and the yellow nodes are an external grower from whom they purchase grapes. The actual wine production components are the burgundy / red nodes which reflect production of L'Acadie Blanc, Pinot Noir and Chardonnay wines. The blue nodes reflect the revenue elements of the model in which total revenue from selling all wine to NSLC is estimated.

Sensitivity analyses of the winery model were carried out after setting the proportions of L'Acadie Blanc and Chardonnay to 0.5 each and the vineyard size to 10ha (thus having 5 ha of each variety) and then for the second grower setting the vineyard size to 5ha and all of it under Pinot Noir. The model thus comprised 5ha of each of the three varieties the winery used. As with previous sensitivity analyses the cost year was set to 2018. With these conditions the total winery revenue was most sensitive to additional information on total revenue from Pinot Noir (43%) followed by the production of Pinot Noir wine (17%) (Figure 13).

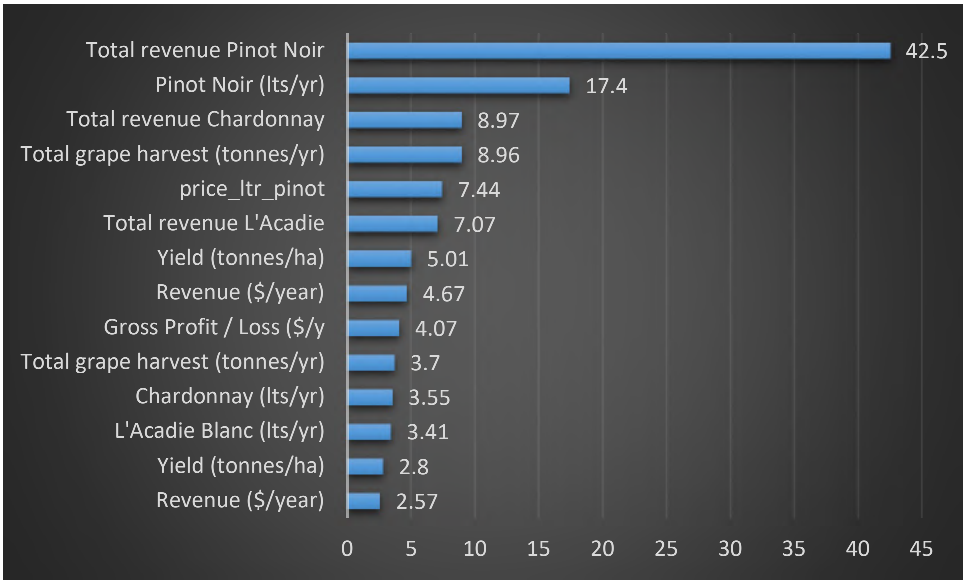

Figure 13. Sensitivity analysis (total winery revenue) of BN Winery model. Estimates based on mutual information: relative to the mutual information due to an additional finding at the target node (Gross Profit / Loss = 100%) how much (% change in mutual information) would be achieved by an additional finding (i.e. data point) at each of the other nodes in the network.

The current version of the winery model is a very large model that requires simplification to be more broadly usable.

Sales and distribution

The sales and distribution component of the BN model was developed from statistical analyses of NSLC annual sales data. Data was available for the period 2008 to 2017 (i.e. FY 2007/08 to 2016/17). The sales data was aggregated (i.e. the sum of dollar value or volume of sales for each year) for three categories of wine: 1) imported wine; 2) Nova Scotia bottled wine[5]; and 3) Nova Scotia wine[6].

The analyses that have been used to develop the wine sales relationships in the overall model are described a separate report[7]. The key values from that report that are used in the BN are the per capita consumption values that were derived from non-linear regression modelling.

---

[5] Wine that was bottled in Nova Scotia from a mixture of imported grape juice and Nova Scotia produced grape juice.

[6] Wine that was made from at least 85% Nova Scotia grape juice.

[7] Lynam , T. 2017. Wine sales in Nova Scotia: Preliminary analyses of NSLC data for fiscal years 2007-2008 to 2016-2017, submitted to NSFA.

Consumption

The consumption components of the BN have yet to be implemented in the working model. The base data and descriptions of the analyses for these data have been presented in a separate report submitted to NSFA[8].

Summary of gaps, opportunities and challenges

The final set of Bayesian Network models provide a suite of tools for the analysis of risk in the Nova Scotia grape and wine industry. Unfortunately, not all required data or sub-models were available at the time the final BN models were developed. Key gaps in the BN model include the disease models developed for the grower BN. These models should be available by the end of the AgriRisk project, but they were not available for inclusion in the final grower BN model. In addition, the grape model needs to be expanded to include a grape quality component.

A second missing sub-model is associated with the production processes of wine itself and in particular the blending of grapes to produce particular wines.

A third sub-model that would benefit the analysis of risk is the consumer preferences or tastes sub-model. Some preliminary analyses have been undertaken with the available data but a model that could be included in the BN has yet to be developed.

In addition to the missing model components sensitivity analyses with the existing BN models directs attention to uncertainties associated with some variables. Uncertainty in the model could be reduced through access to more or better data for the following:

- Grape yield. In particular grape yield as a function of variety and climate variables;
- Grape quality (i.e. brix, pH, acidity) as a function of yield, climate variables and site characteristics;
- Grape phenology. The dates or timing of key phenological stages (bud burst, flowering, version, harvest) as a function of GDD and site characteristics;
- Grape prices. The prices paid to growers for varieties and quality of grapes;
- Grape economics. The current grape economic analyses were based on relatively old data and do not vary by variety, site characteristics or management practices. There is considerable room for improvement in these data;
- Wine economics. The wine economics data were developed in Ontario as none currently exist for Nova Scotia wine production. Costs of production data for Nova Scotia wines would be very useful;

[8] Lynam, T and A. Lindo. 2018. AgriRisk: A brief examination of consumer preferences in relation to Nova Scotia wine.

- Farm gate sales of wines by wine type. The only farm gate sales data available to the AgriRisk team was highly aggregated and did not permit analyses of sales at farm gate to the level of varieties;
- Levels of inventory held by wineries and over what period. Information on the decisions wineries make about what to put in inventory and how long to keep it there as well as the price implications of storing wine would be useful to derive improved wine pricing relationships;

The BN modelling approach provides significant opportunities for expansion and refinement of the BN models. With models developed as modules that can be coupled together to form integrated models such as the current winery model there is an opportunity to develop models for specific wineries in Nova Scotia and provide the support to wineries to use the models with growers to develop strategies that sustain economic well being in the face of considerable risk and uncertainty.

References cited

Burkner, P.-C. (2017). {brms}: An {R} Package for Bayesian Multilevel Models using Stan.

CEL, C. E. L. (2018). *Grape Operations Model Final Report*. Retrieved from Halifax, Nova Scotia, Canada:

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning, 29*(2-3), 131-163. doi:10.1023/A:1007465528199

NSDA, N. S. D. o. A. (2014). *Nova Scotia Wine Grape Cost of Production and Cash Flow Analysis*. Retrieved from Nova Scotia, Canada:

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Appendices