Model Selection Strategy for Bayesian Networks

A Case Study of New Zealand Viticulture

Innocenter Amima

Supervisors: Dr. Beatrix Jones, Dr. Sarah Knight, Dr. Kate Lee
The University of Auckland

Australasian Applied Statistics Conference 2022

NEW ZEALAND WINE PURE DISCOVERY · MINISTRY OF BUSINESS, INNOVATION & EMPLOYMENT HIKINA WHAKATUTUKI · Plant & Food RESEARCH RANGAHAU AHUMARA KAI · THE UNIVERSITY OF AUCKLAND NEW ZEALAND SCIENCE · BRAGATO RESEARCH INSTITUTE NEW ZEALAND GRAPE AND WINE RESEARCH RANGAHAU KAREPE, WĀINA O AOTEAROA

The vineyard as an ecosystem (VE) programme

- Wine is New Zealand's 6th largest export good; in Nov 2020, it was valued at NZ$2B [1].
- Challenge: pests and diseases diminish the grape quality, yield and threaten the sustainability of the wine industry [1].

The vineyard as an ecosystem (VE) programme

- Wine is New Zealand's 6th largest export good; in Nov 2020, it was valued at NZ$2B [1].
- Challenge: pests and diseases diminish the grape quality, yield and threaten the sustainability of the wine industry [1].
- Investigate the long-term impacts of two distinct **management practices** on the longevity and sustainability of vineyards.

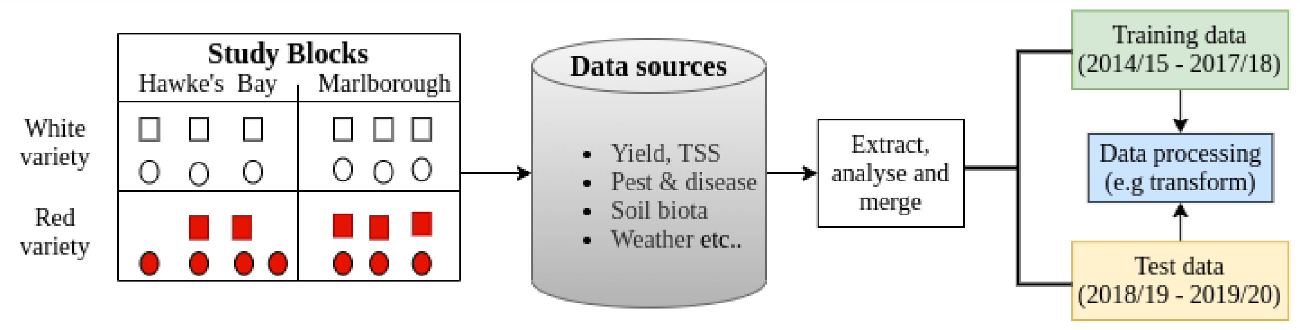
Contemporary: use synthetic sprays and herbicides to control under-vine weeds.

Future: use fewer synthetic sprays and cultivation to control weeds; can support biodiversity.

- Focus on identifying the factors influencing yield components, pests and diseases.

Data and data processing

- Study design:
  - 24 vineyard blocks in **Marlborough** (12) and Hawke's Bay (12).
  - Each group of 12 comprised 6 Sauvignon blanc blocks and 6 red varieties.
  - 11C and 13F managed blocks were monitored 1-3 times/season for 5 seasons.



- After data processing; n = 120, p = 131 variables representing the VE components.

Learning Bayesian networks (BNs)

- BNs express the conditional independence relationships among variables $\mathbf{X}$ via graphical separation [2].
  - Thus specifying the factorisation $\mathrm{P}(\mathbf{X}) = \prod \mathrm{P}(X_i \mid \Pi_{X_i}, \Theta_{X_i})$ [2].
- A BN is defined by a DAG $\mathcal{G} = (\mathbf{V}, A)$ and parameters Θ of $P(\mathbf{X})$ [2].

$$\underbrace{\mathrm{P}(\mathcal{G}, \Theta \mid \mathcal{D})}_{\text{learning}} = \underbrace{\mathrm{P}(\mathcal{G} \mid \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D})}_{\text{parameter learning}}.$$

- Greedy search algorithm to learn $\mathrm{P}(\mathcal{G} \mid \mathcal{D})$ from 10k bootstrap samples [5, 6, 7].
- Searches over the space of DAGs for a structure that maximises a score [5].

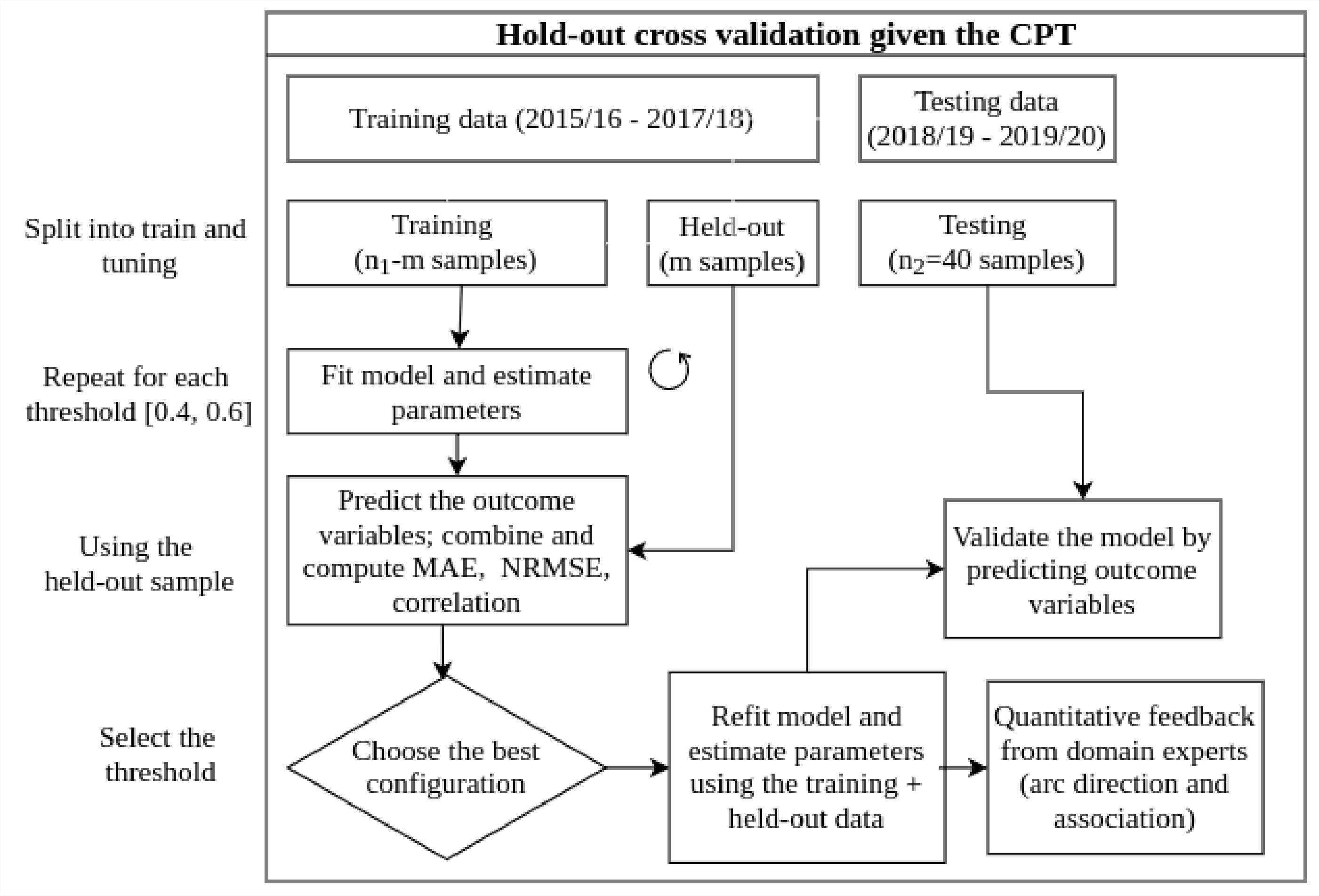
Motivation: Choice of threshold

- **Output**: CPT (17k arcs) contains the frequency and confidence in direction of each arc.

from	to	strength	direction
Syn_BotsFungicide	VI_H	0.65	0.85
MBCatch_InFl	MBCatch_BC	0.57	1.00
Management	TSS_H	0.48	1.00
Soft_PMFungicide	BunchPM_BC	0.44	0.67

- How can we choose the threshold to identify significant arcs when some variables are more important (yield, pests and diseases)?

Proposed model selection strategy for BNs



Results: choice of threshold

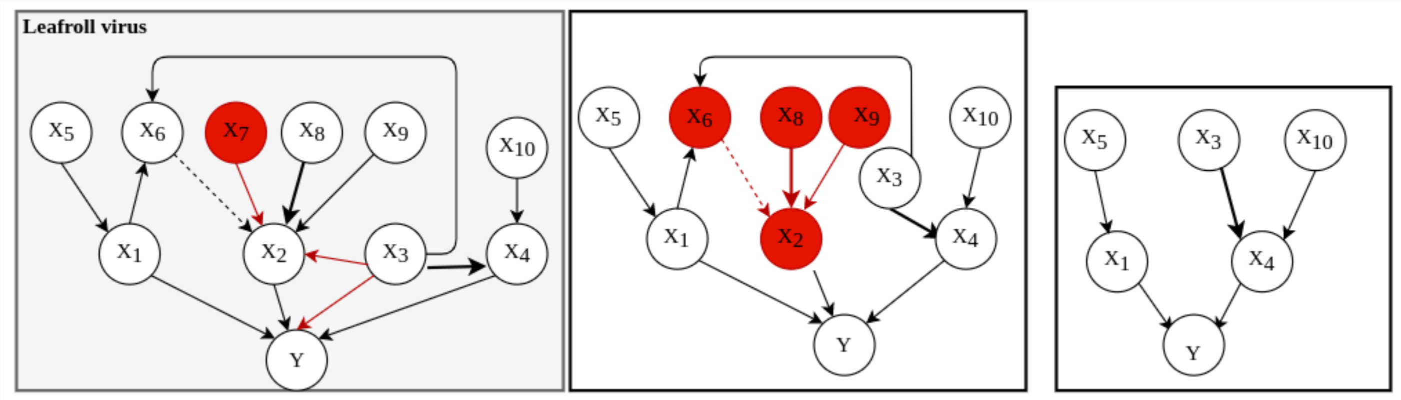
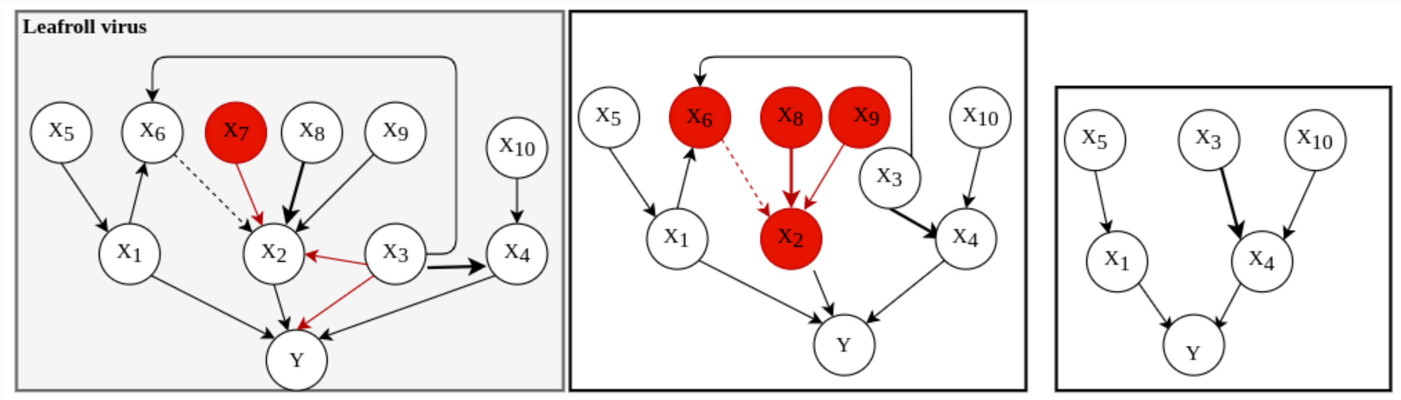


Diagram	**-**	**1**	**2**	**3**
Threshold	0.4	**0.47**	0.5	0.55
Average NRMSE (X_i = 12)	1.472	**0.772**	0.767	0.797
Average correlation (X_i = 12)	0.568	**0.578**	0.574	0.538
Directed arcs	471	**351**	303	250

Results: choice of threshold



A threshold = 0.47 provided a sparse, interpretable model with better prediction performance.

Diagram	**-**	**1**	**2**	**3**
Threshold	0.4	**0.47**	0.5	0.55
Average NRMSE (X_i = 12)	1.472	**0.772**	0.767	0.797
Average correlation (X_i = 12)	0.568	**0.578**	0.574	0.538
Directed arcs	471	**351**	303	250

Conclusion

- Built a consensus model with arcs that appear more than 47% in the 10k structures (351 directed arcs).
- Model validation: prediction correlation was [0.11, 0.84] and NRMSE [0.49, 2.33].
- Pests and diseases were influenced by management, weather, soil and region/variety.

Conclusion

- Built a consensus model with arcs that appear more than 47% in the 10k structures (351 directed arcs).
- Model validation: prediction correlation was [0.11, 0.84] and NRMSE [0.49, 2.33].
- Pests and diseases were influenced by management, weather, soil and region/variety.

Future work

- Perform what-if scenario analysis to predict the impact of climate change.
- Finalise the three-step framework for hypothesis generation, refinement and analysis.

Thank You

- Supervisors: Beatrix Jones, Sarah Knight and Kate Lee.
- Members past and present who have contributed to the vineyard ecosystem programme.
- Funding bodies (MBIE, NZ wine growers).

THE UNIVERSITY OF AUCKLAND NEW ZEALAND
SCIENCE

MINISTRY OF BUSINESS, INNOVATION & EMPLOYMENT
HĪKINA WHAKATUTUKI

NEW ZEALAND WINE
PURE DISCOVERY

Plant & Food RESEARCH
RANGAHAU AHUMARA KAI

BRAGATO RESEARCH INSTITUTE
NEW ZEALAND GRAPE AND WINE RESEARCH
RANGAHAU KAREPE, WĀINA O AOTEAROA

References

Greven, M.; Arnold, N.; Bell, V.; et al. (2017). Vineyard Ecosystems RA 1.1 Annual Report.

Pearl, J. (1988). Preface to the Fourth Printing. In: Pearl, J. (Ed.), Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann.

Friedman, N. (2013). The bayesian structural EM algorithm.

Scutari, M. (2010). Learning bayesian networks with the bnlearn r package. Journal of Statistical Software, Articles, 35 (3), 1–22. Retrieved from https://www.jstatsoft.org/v035/i03 doi: 10.18637/jss.v035.i03

Glover, F. (1990). Tabu Search: A Tutorial, Interfaces .

Claeskens, G., & Hjort, N. L. (2008). Model selection and model averaging. Cambridge University Press. Retrieved from https://EconPapers.repec.org/RePEc:cup:cbooks:9780521852258

Friedman, N., Goldszmidt, M., & Wyner, A. J. (2013). Data analysis with bayesian networks: A bootstrap approach. CoRR, abs/1301.6695. Retrieved from http://arxiv.org/abs/1301.6695