We thank Rio Tinto for the use of their data.

Graphical Models for Plant data

Margaret Donald
Joint work with AnneMarie Clements and Kerrie Mengersen

ABNMS2011
November 23, 2011

Graphical Models for Plant data

Data

Data are species presence/absence data

- for 89 quadrats
- 242 native and exotic species
- but limited for modelling purposes, to the 46 species occurring on at least 10 quadrats.

These data were collected as baseline data at sites with patches of Warkworth Sands Woodland (listed as an endangered ecologocial community) as part of a project to restore them to self-sustaining natural ecosystems.

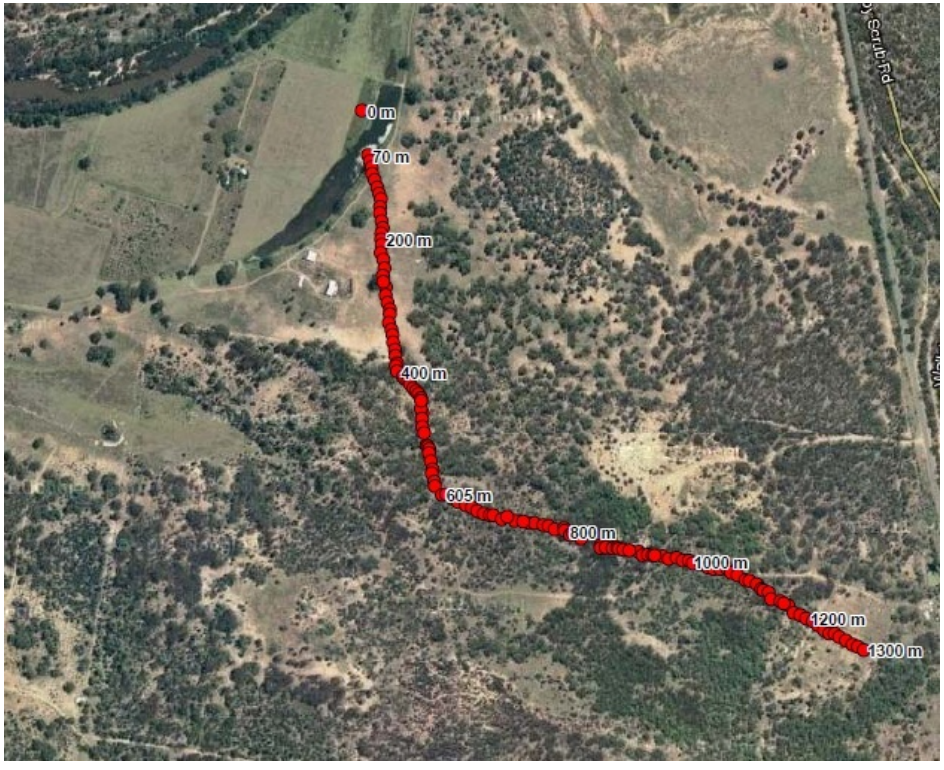


Figure: Transect in the southern Warkworth sands area.

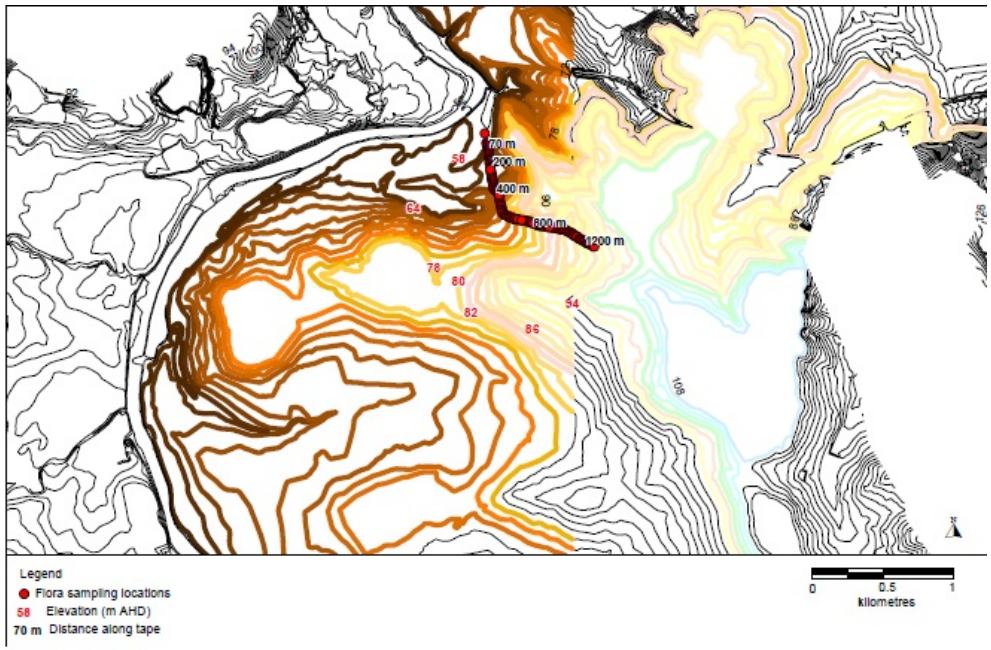


Figure: Transect in the southern Warkworth sands area together with 2 m contour curves.

Figure: Vegetation (*A.luhmannii*) at 160 m, transect 1.

Figure: Vegetation (mainly exotics) at 1200 m, transect 2.

Table: Species and variable names for species which occur in at least 10 quadrats.

	Grouping	*Species*	*Variable*
1	TREES	Acacia filicifolia	1
		Angophora floribunda	5
		Eucalyptus crebra	10
		Allocasuarina luehmannii	240
2	SHRUBS	Breynia oblongifolia	15
		Hibbertia linearis	21
		Pimelea linifolia	30

Table: Species and variable names for species which occur in at least 10 quadrats.

	Grouping	*Species*	*Variable*
3	GROUND COVER	Aristida ramosa	35
		Cheilanthes sieberi	54
		Chrysocephalum apiculatum	56
		Cynodon dactylon	61
		Dianella longifolia	68
		Digitaria diffusa	75
		Einadia hastata	79
		Einadia trigonos	81
		Eragrostis brownii	84
		Glycine clandestina	88
		Imperata cylindrica	95
		Laxmannia gracilis	100
		Lomandra filiformis	101
		Lomandra leucocephala	102
		Lomandra multiflora	104
		Microlaena stipoides	106

Table: Species and variable names for species which occur in at least 10 quadrats.

	Grouping	*Species*	*Variable*
4	EXOTIC	Acetosella vulgaris	144
		Anagallis arvensis	148
		Conyza parva	165
		Facelis retusa	170
		Galenia pubescens	171
		Gamochaeta antillana	172
		Gamochaeta pensylvanica	173
		Heliotropium amplexicaule	176
		Hypochaeris radicata	181
		Melinis repens	192
		Opuntia aurantiaca	200
		Opuntia humifusa	201
		Petrorhagia nanteuilii	206
		Polycarpon tetraphyllum	210
		Senecio madagascariensis	215
		Sonchus oleraceus	224

Undirected graphical modelling in R

- minet (Meyer et al 2008)
- gRapHD (de Abreu et al 2011)
- gRim (Hojsgaard 2011)
- dynamicGraph (Badsberg 2011)
- gRbase (Dethlefsen and Hojsgaard 2005)
- RBGL (Carey et al 2011)
- infotheo (Meyer 2011)
- dwig (Whittaker 2011)

DAGs for three variables

$A \perp C|B$

$A \perp C|B$

$A \perp C|B$

$A\neg \perp C|B$

Corresponding undirected graphical structures

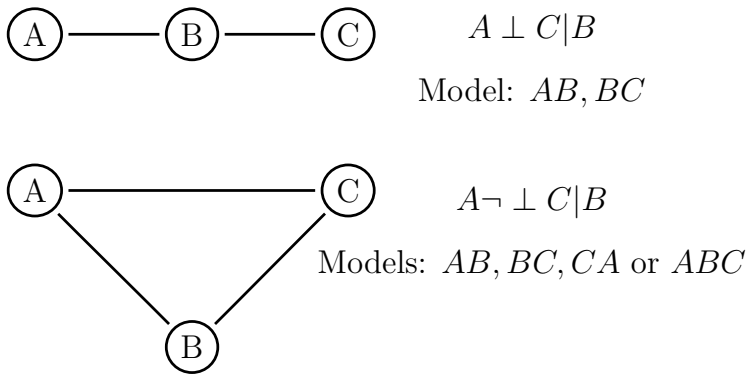


Figure: Corresponding undirected graphical structures

Modelling issues

- Seeking an undirected graphical structure
- in sparse data
- The number of possible log-linear models for these data is
- $2^{46} - 1 - 46 = 7.04 \times 10^{14}$

Hence, we use a two-way saturated interaction model, in order to compare models for adequacy. (And note that for 46 variables, this corresponds to a model with ${}^{46}C_2$ or 1035 df.)

Algorithms/Techniques used

Algorithms and techniques used

- Chow-Liu tree algorithm (Chow and Liu 1968)
- ARACNE algorithm (Margolin et al 2006)
- together with backwards/forwards selection
- Bootstrapping

Additionally as an exploratory technique, we use chain graphs (Wermuth and Cox 1998), using the "dwig" package of Kao & Whittaker (2011).

Assessing model adequacy

Models can be assessed via the change in deviance, but here we assess model adequacy by comparing the mutual information (measured in millibits) of the chosen model as a proportion of the 'total' information explained by the saturated two-way model.

Ground cover models

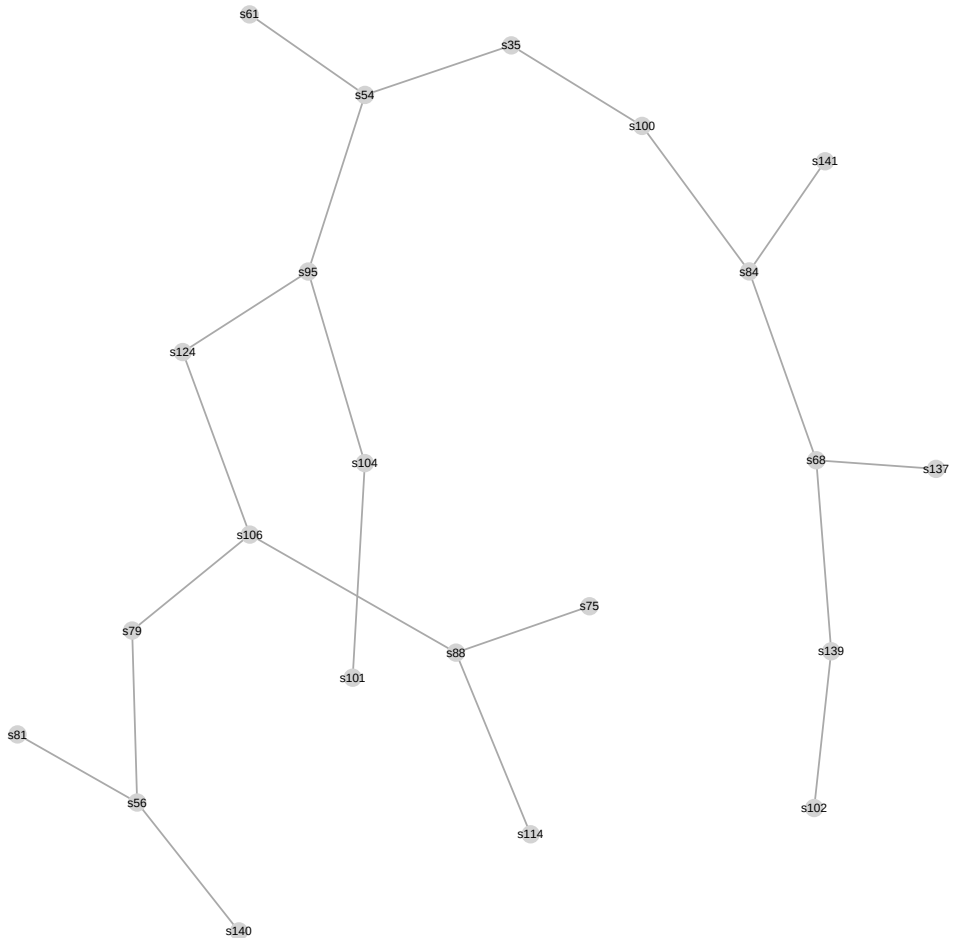


Figure: Chow Liu tree algorithm reconstruction for 22 ground covers.

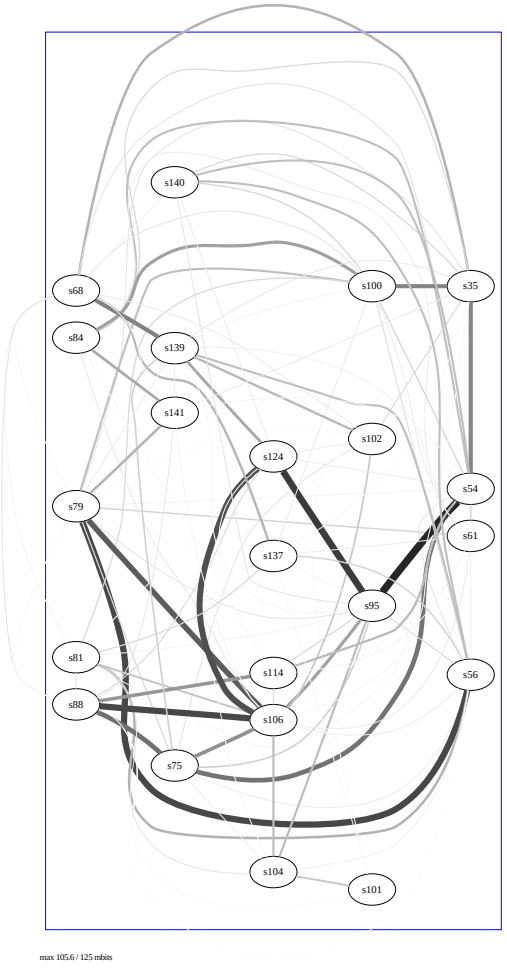


Figure: Divergence weighted independence graph for the 22 ground cover species: Marginal mutual information.

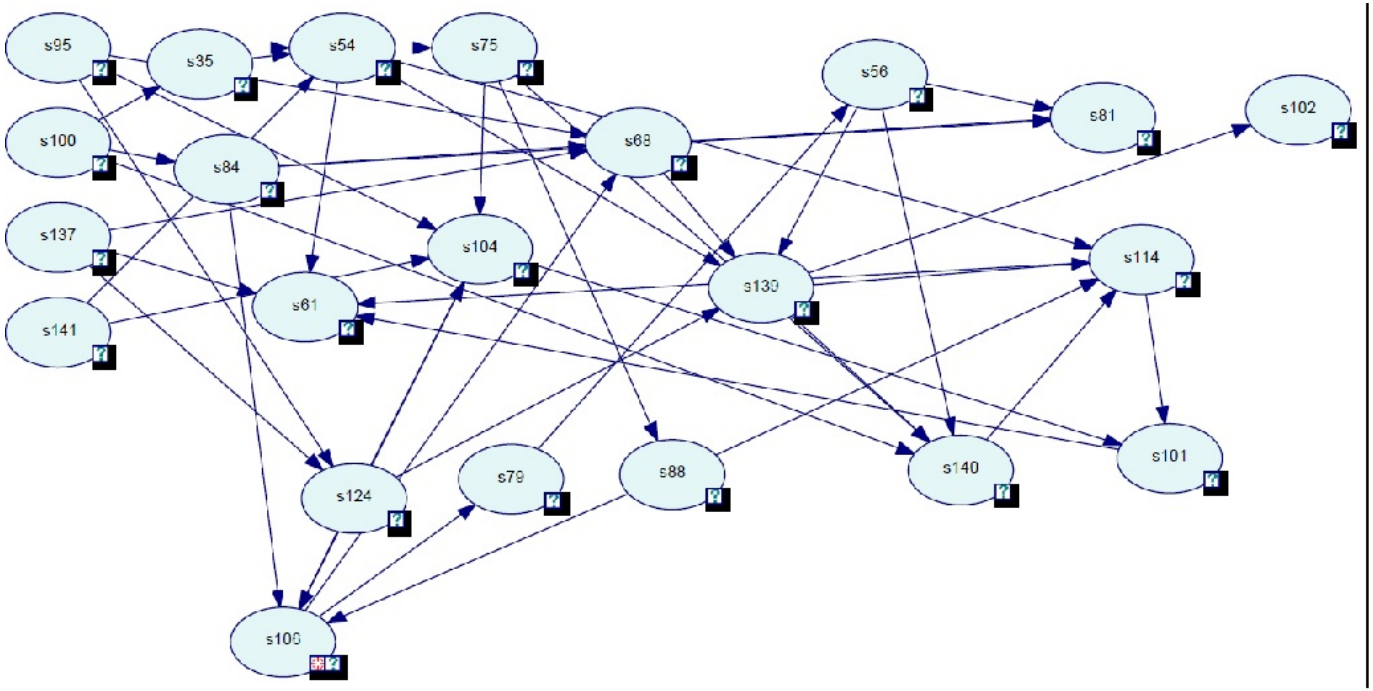


Figure: 22 Ground Cover variables: BN structure found using K2 algorithm in Genie

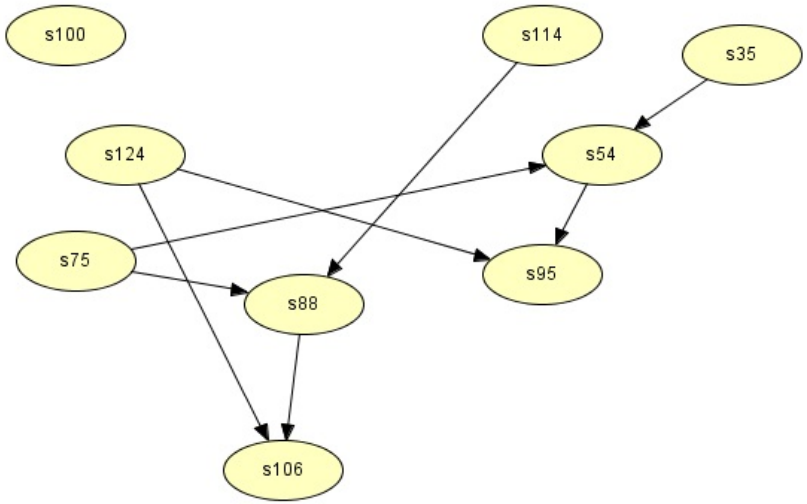


Figure: 9 Ground Cover variables: BN structure found using NPC algorithm in Hugin

Within the undirected framework of these packages, I can find

- a best forward model from the independence model (not necessarily two-way)
- a best backward model from the two-way model
- an ARACNE model
- a tree model

But with so few data, I would like a robust model. Hence, I bootstrap one of my possible models (the ARACNE model), to derive a robust graphical model.

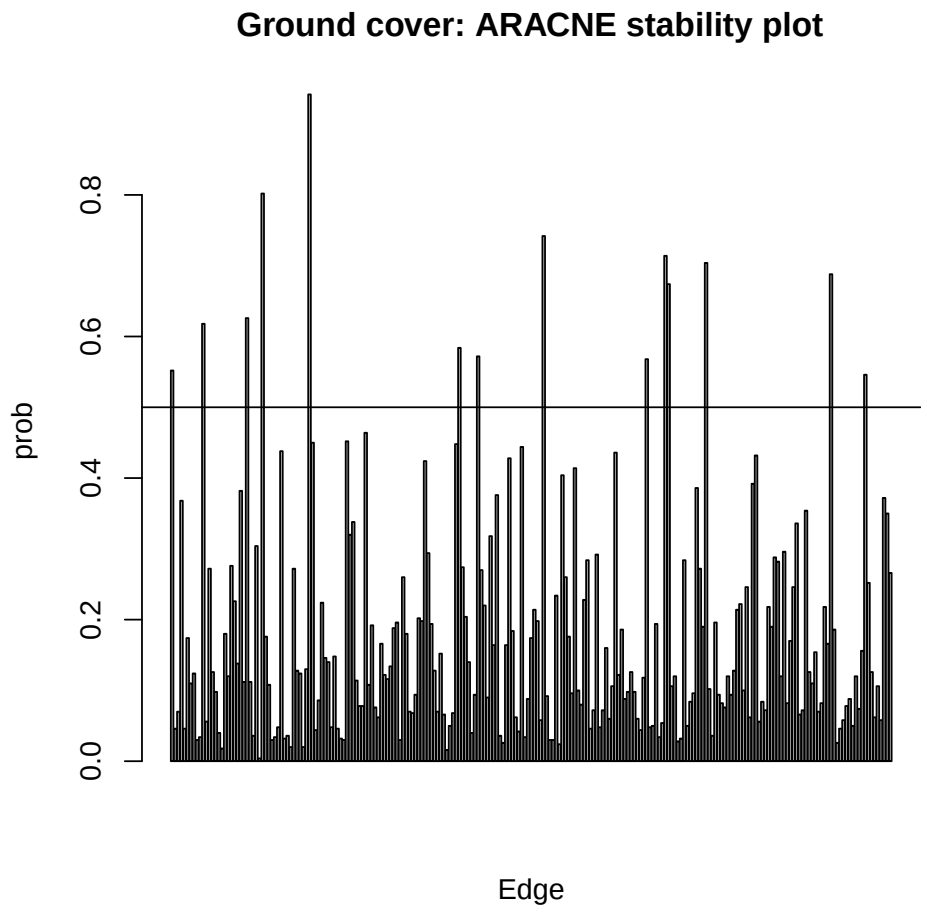


Figure: Bootstrapping the ARACNE model for 22 ground cover species: Proportions of edge retention in the model.

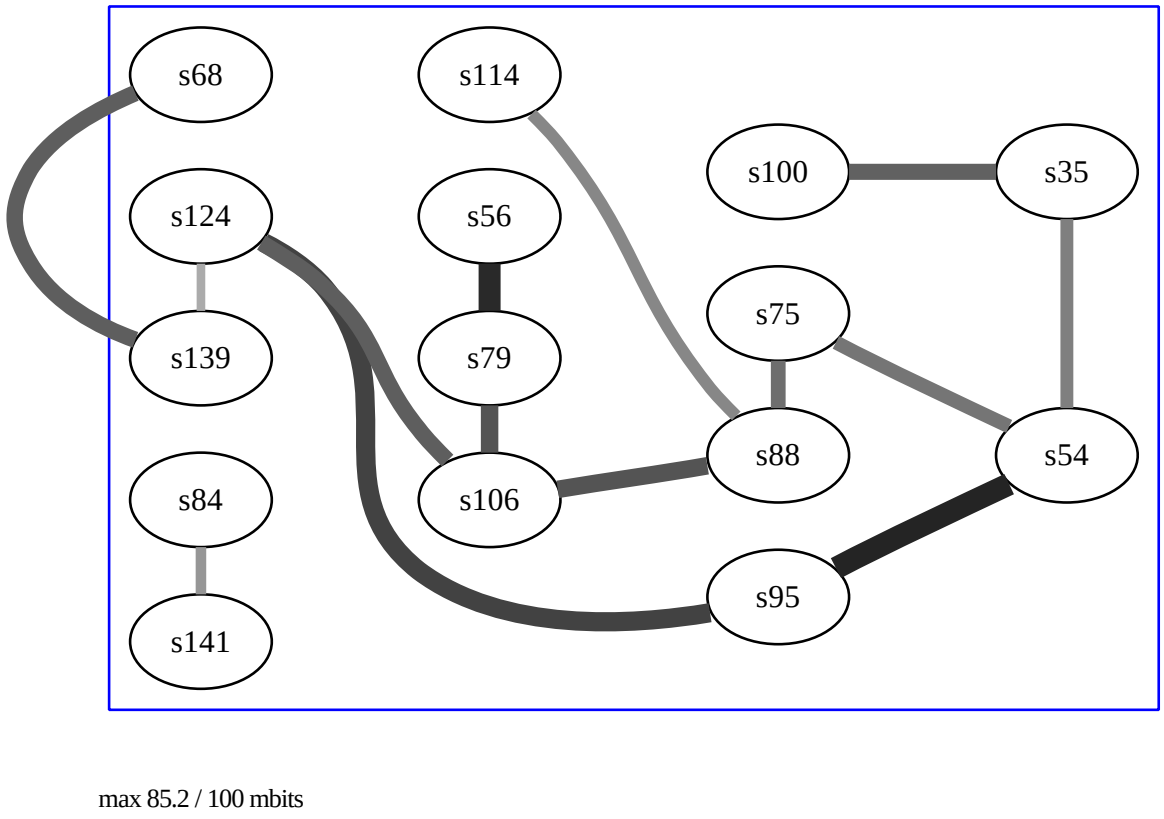


Figure: The bootstrapped ARACNE model for 22 ground cover species (Independent nodes not shown are s61, s81,s101, s137, s102, s104, s140).

Aracne Groundcover model: Observed vs Fitted

Table: Observed vs Predicted frequencies when s106=1

s106	*s79*	*s88*	*s124*	*Frequency Observed*	*Predicted*
1	0	0	0	13	10.38
1	1	0	0	3	3.46
1	0	1	0	7	9.87
1	1	1	0	4	3.29
1	0	0	1	6	8.37
1	1	0	1	3	2.79
1	0	1	1	10	7.38
1	1	1	1	2	2.46

s106: Microlaena stipoides
s124: Pteridium esculentum
s79: Einadia hastata
s88: Glycine clandestina

Aracne Groundcover model: Marginal & Conditional probabilities

Table: Probability that s106=1.

s79	*s88*	*s124*	*p(s106=1,s124,s88,s79)*	*p(s106=1|s124,s88,s79)*
0	0	0	0.12	0.26
1	0	0	0.04	0.82
0	1	0	0.11	0.66
1	1	0	0.04	0.96
0	0	1	0.09	0.67
1	0	1	0.03	0.96
0	1	1	0.08	0.92
1	1	1	0.03	0.99

s106: Microlaena stipoides
s124: Pteridium esculentum
s79: Einadia hastata
s88: Glycine clandestina

Chain Graphs

In these graphs,

- we assume precedence of one set of species before another
- either in time, or causally..

Thus species in the first block cannot be predicted by those in later blocks. This gives a combination of directed and undirected edges, since species in the later block may be predicted by those from an earlier block.

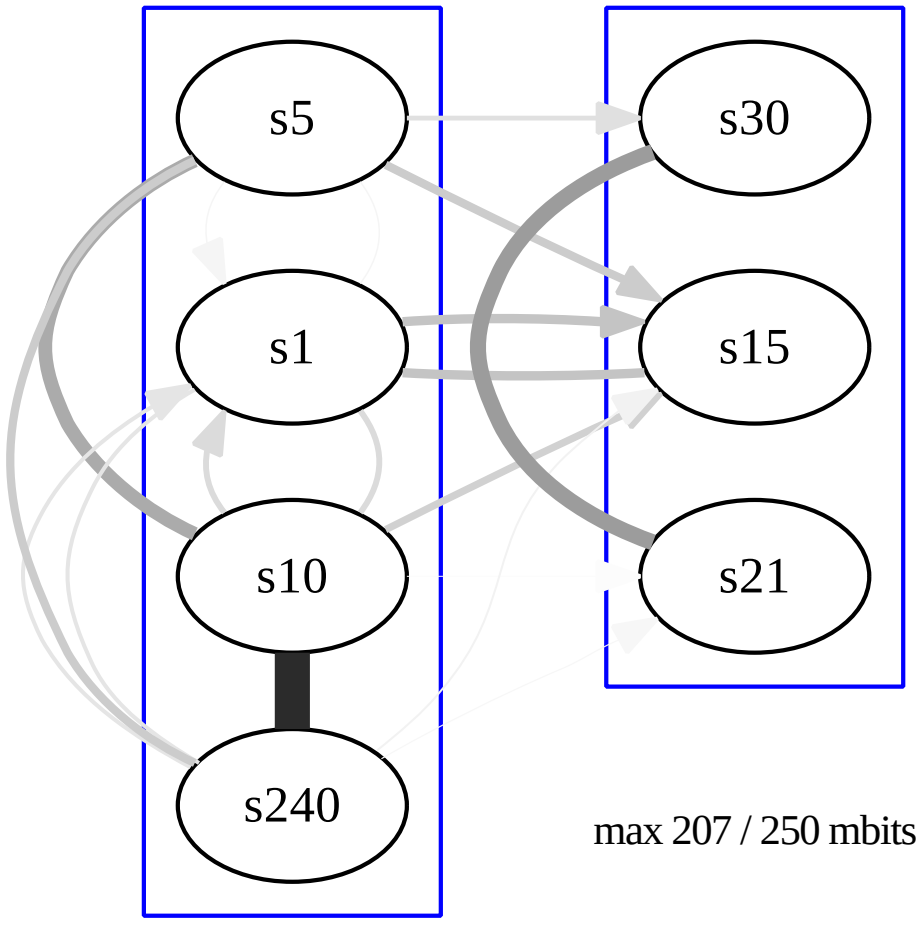


Figure: Chain graph with shrubs dependent on trees.

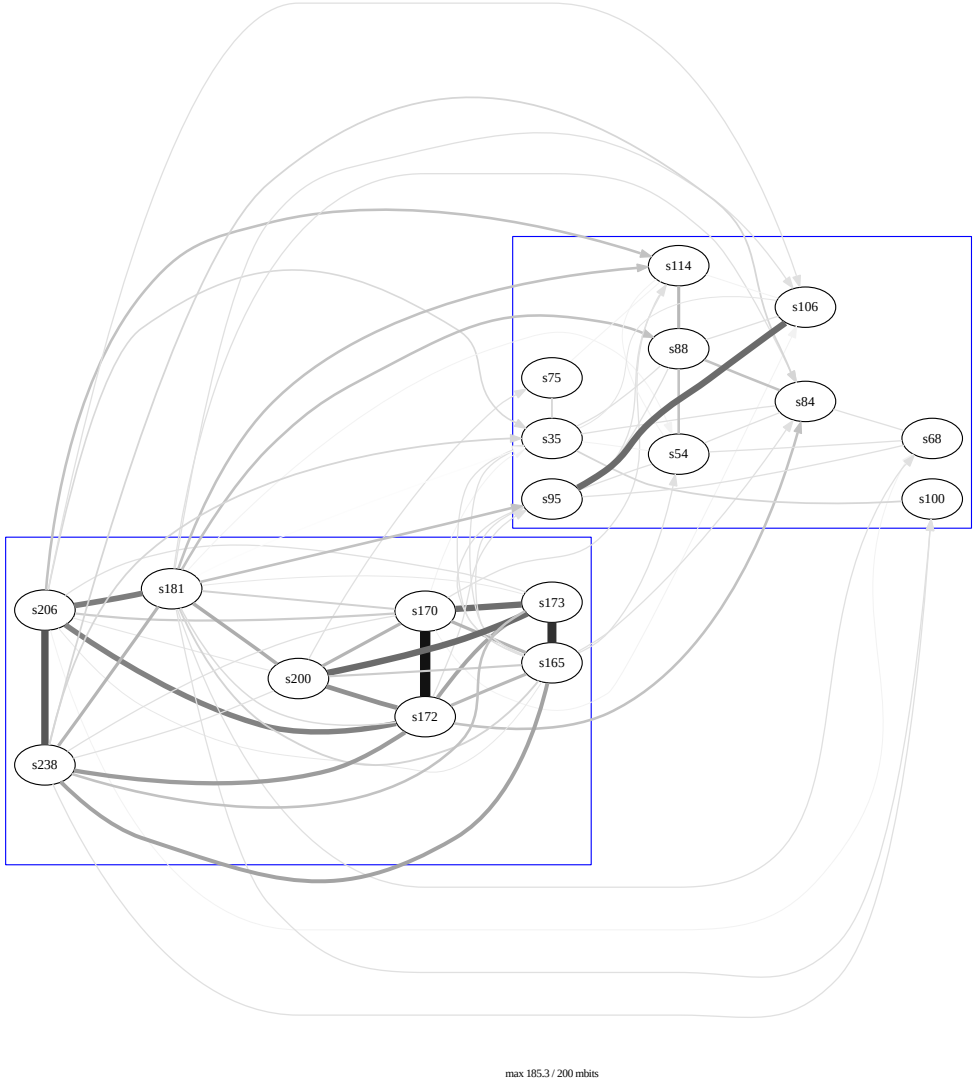
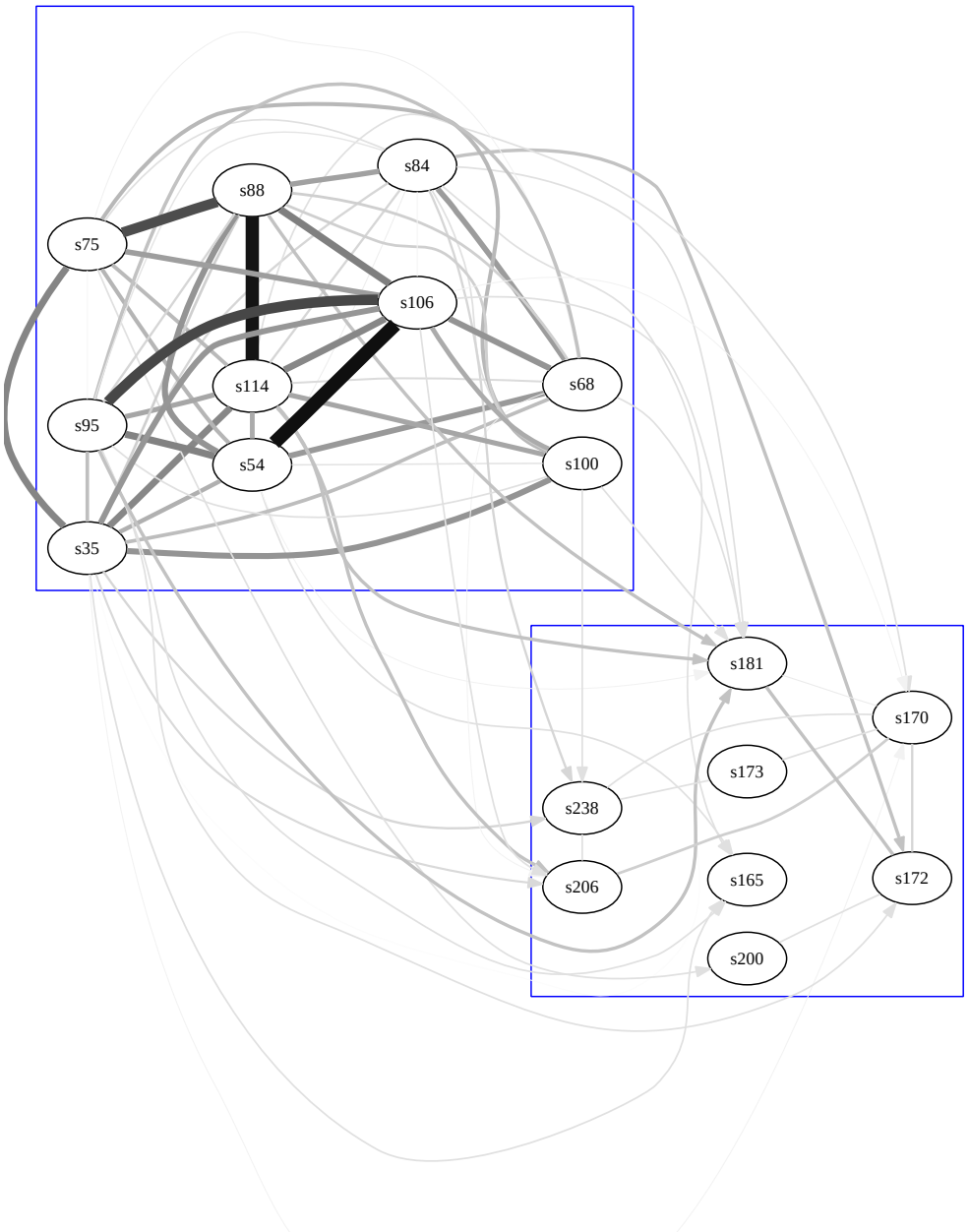


Figure: Exotics dependent on exotics; Ground cover on all other variables.


s84
s88
s75
s106
s114
s68
s95
s54
s100
s35
s181
s170
s173
s238
s206
s165
s172
s200

Chain graphs

Ideally, we wanted to build our models such that

- Trees $\Rightarrow$ Shrubs
- Trees $\Rightarrow$ Groundcover
- Exotics $\Rightarrow$ Groundcover, or should it be,
- Groundcover $\Rightarrow$ Exotics
- Trees $\Rightarrow$ Exotics

The 'dwig' package did not always permit this. Thus, the following graph is based on the idea of trees predicting ground cover, but it is not a chain graph.

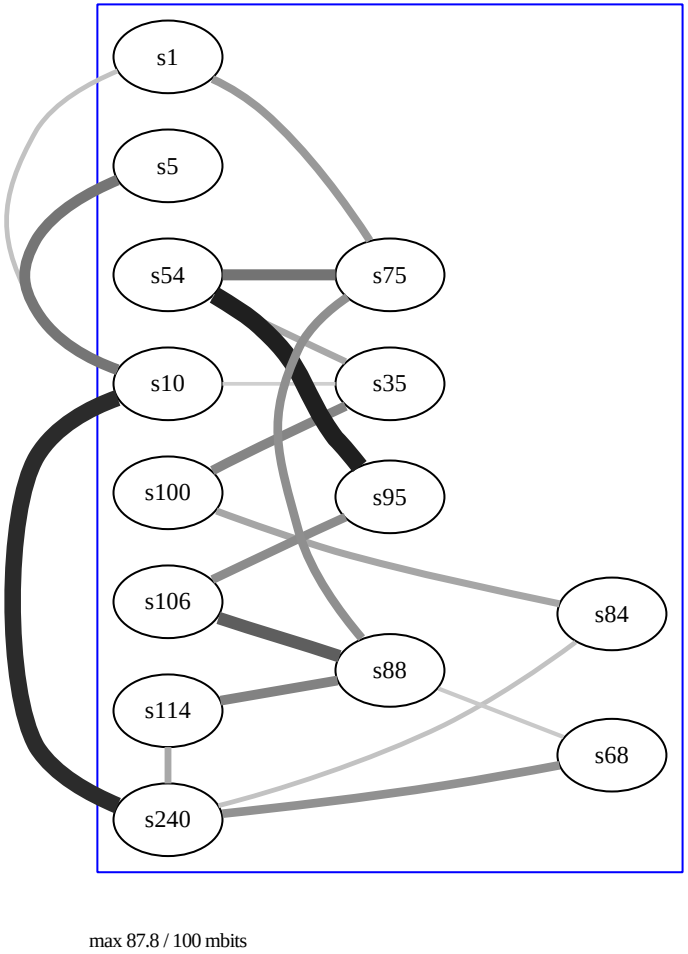


Figure: ARACNE model for Trees and 10 Ground Cover species.

Model comparisons for the Trees & Ground cover model (14 species)

Table: Deviances for models for 4 Trees and 10 Ground Covers.

Model	*Model deviance*	*Model df*	*Deviance*	*df*
Independent	0	0	596.58	16369
Minimum forest	142.21	13	454.37	16356
ARACNE	184.02	20	412.56	16349
Saturated two-way	271.68	91	324.90	16278

Model comparisons for the Trees & Ground cover model (14 species)

Table: Explained Information for the ARACNE model

	Information	*df*
ARACNE	1527.3	20
Residual	727.5	71
Total (saturated two-way)	2254.8	91

Final comments

- Modelling using gRim and gRapHD works well with the 22 or so variables used here
- The information theoretic packages grind to a halt when getting entropies for fitted models of $\sim$20 variables
- Hugin models found using the NPC algorithm look very like the undirected models, but are often forced to shed an edge or two by the constraints imposed by being a DAG
- With 10-15 binary predictors, all the packages work well.

Final comments

These packages are worth considering for modelling, because they permit

- easily accessible criteria for model choice
- easy scrutiny of graphical structure
- unconstrained modelling of graphical structure

But more importantly (and not demonstrated here), they

- allow modelling of graphical Gaussian models for multivariate normal data
- and of models with both categorical and Gaussian data

Selected References

Badsberg, J. H.: 2011, R package dynamicGraph.
URL: *http://cran.r-project.org/web/packages/dynamicGraph/index.html*

Carey, V., Long, L. and Gentleman, R.: 2011, R package RBGL.
URL: *http://cran.r-project.org/web/packages/RBGL/index.html*

Chow, C. and Liu, C.: 1968, Approximating discrete probability distributions with dependence trees, *IEEE Transactions on Information Theory* **14**(3), 462–467.

de Abreu, G. C. G., Labouriau, R. and Edwards, D.: 2011, R package: gRapHD.
URL: *http://cran.r-project.org/web/packages/gRaphHD/index.html*

Dethlefsen, C. and Hjsgaard, S.: 2005, A common platform for graphical models in R: The gRbase Package, *Journal of Statistical Software* **14**(17), 1–12.

Hojsgaard, S.: 2011, R package 'gRim Graphical Interaction Models.
URL: *http://cran.r-project.org/web/packages/gRim/index.html*

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. and Califano, A.: 2006, Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC bioinformatics* **7**(Suppl 1), S7.

Selected References

Meyer, P. E.: 2011, R package infotheo Information-theoretic measures.
URL: *http://cran.r-project.org/web/packages/infotheo/index.html*

Meyer, P. E., Lafitte, F. and Bontempi, G.: 2008, Minet: An open source R/bioconductor package for mutual information based network inference, *BMC Bioinformatics* **9**.
URL: *http://www.biomedcentral.com/1471-2105/9/461*

Wermuth, N. and Cox, D. R.: 1998, On association models defined over independence graphs, *Bernouilli* **4**(4), 477–495.

Whittaker, J.: 1990, *Graphical Models in Multivariate Statistics*, Wiley, Chichester (England); New York.

Whittaker, J.: 2008, Bootstrapping divergence weighted independence graphs for design based survey analysis.
URL: *http://www.newton.ac.uk/programmes/SCH/seminars/031911002.pdf*

Whittaker, J.: 2011, *An Introduction to Graphical Models using R*. Course, University of Queensland, June 2011.

Thank you for listening. Questions?

Figure: Exotic: *Opuntia humifusa*

Figure: Native: *Calandrinia balonensis*