Centre for Data Science

AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Bayesian Modelling and Analysis of Challenging Data

Kerrie Mengersen
School of Mathematical Sciences
QUT

PC Mahalanobis Lecture Series
January 2021

Programme of Lectures

January 27th:

- Lecture 1: 10-1045am IST (230pm-3:15pm AEST)
  Identifying the Intrinsic Dimension of High-Dimensional Data

- Lecture 2: 11-11:45am IST (3:30pm-4:15pm AEST)
  Finding Patterns in Highly Structured Spatio-Temporal Data

January 29th:

- Lecture 3: 10-1045am IST (230pm-3:15pm AEST)
  Describing Systems of Data

- Lecture 4: 11-11:45am IST (3:30pm-4:15pm AEST)
  Making New Sources of Data Trustworthy

Bayesian Modelling and Analysis of Challenging Data

Lecture 3:
Describing Systems of Data

Sandra Johnson, Paul Wu,
Charisse Farr, Fabrizio Ruggeri

Everything is a complex system!

Case study 1: Bayesian network modelling of lyngbya

University of Queensland

What are the main factors that influence the initiation of lyngba?
What management approaches are most effective?

Caboolture Shire Council

Case Study 2: “Beyond compliance”

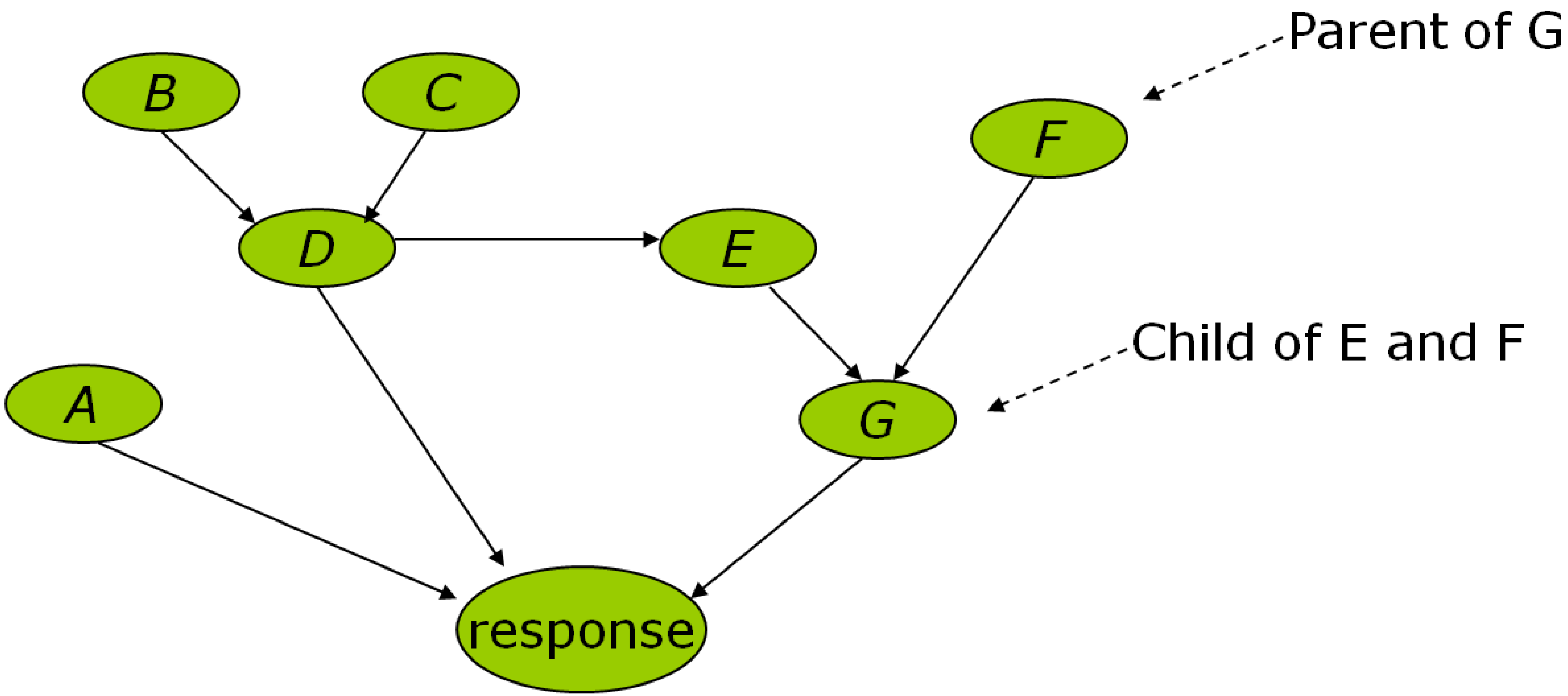
- Requires pest risk mitigation (biosecurity) measures
    - Subject to international standards (ISPMs). Must be based on pest risk, scientifically justified, proportional to risk and least trade-restrictive
- Pest risk mitigation measures are usually single, e.g. pest area freedom or chemical treatment. These can:
    - Be difficult (or impossible) to achieve
    - Damage the commodity
    - Carry health and environmental risks
    - Halt the whole trade on a minor failure
    - Convey a power imbalance between trading partners

Case study 3: from BN to DBN

Can we find “ecological windows” for dredging to reduce the impact on seagrass?

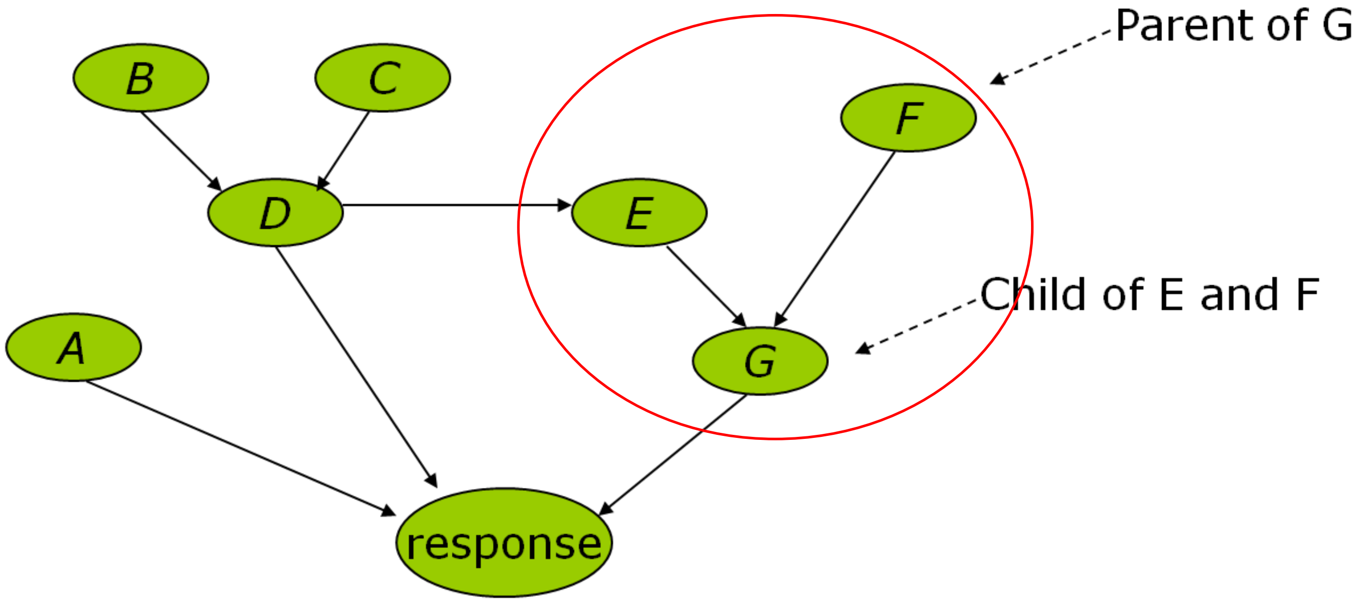
Case Study 4: Wayfinding – Combining expert information

Using Bayesian Networks to model complex systems

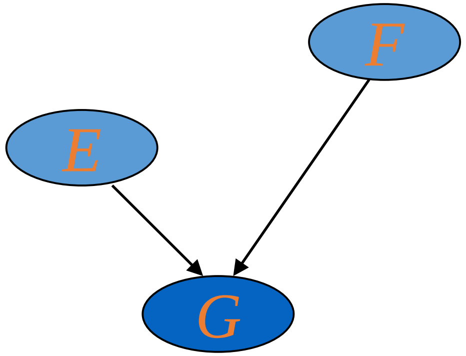


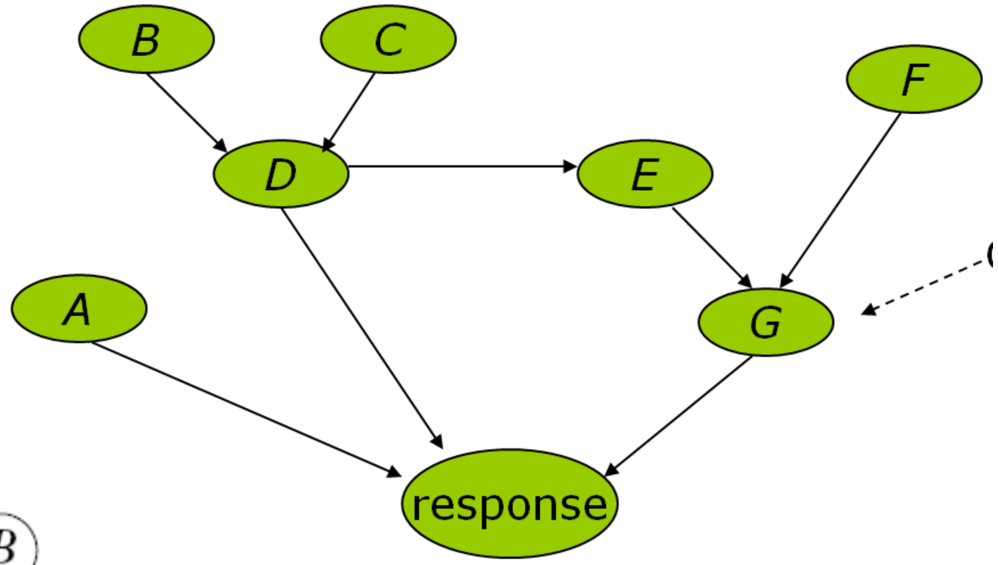
Markov Assumption: (1st order)

If we know the present, then the past
has no influence
on the future

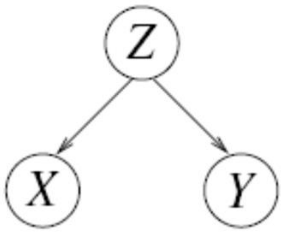


Markov Blanket
children, parents, children’s parents

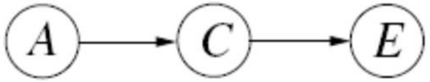




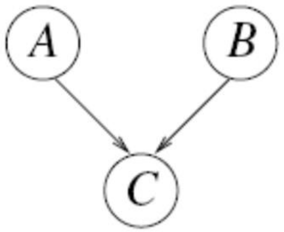
- 3 types of structural connections:



Diverging connection
(common cause)



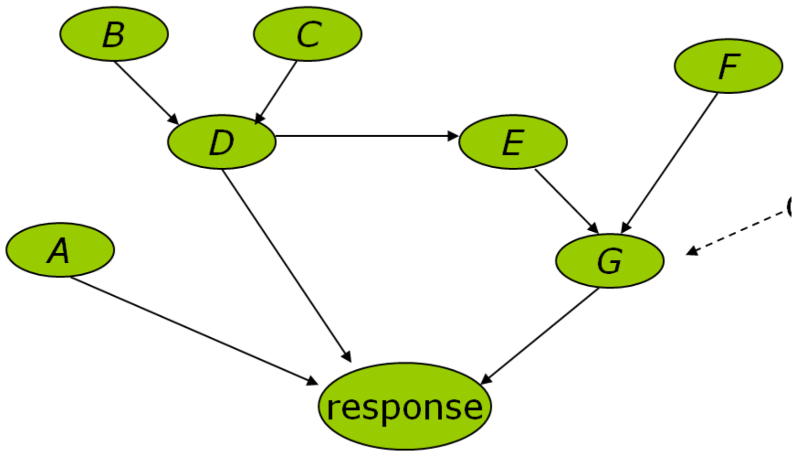
Serial connection



Converging connection
(common effect)

- *d-separation:*
  If nodes X & Y are d-Separated given Z, then $X \perp Y \mid Z$

BN Chain Rule



- Instead of calculating the joint probability distribution across all the nodes using the multiplication law of elementary probability theory:

$$P(X_1, . . .,X_n) = P(X_1)P(X_2/X_1)P(X_3/X_1, X_2).....P(X_n/X_1,...,X_{n-1})$$

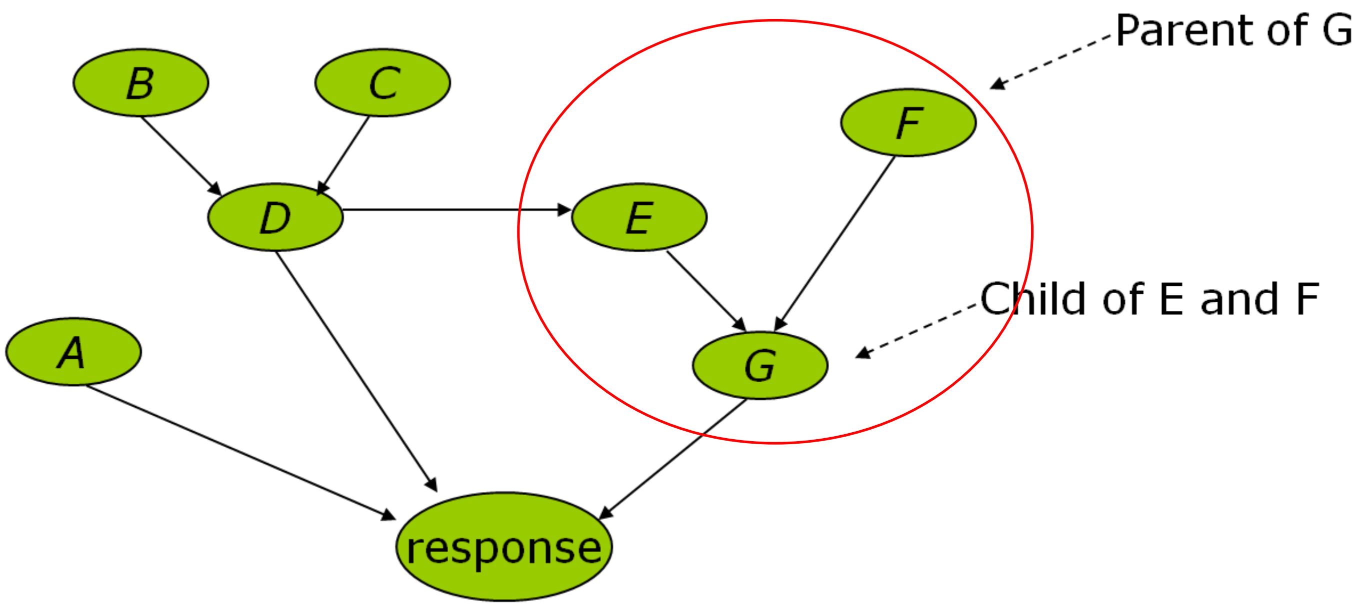
- By using d-Separation (i.e. conditional independence) and the Markov property this simplifies into the well known BN chain rule:

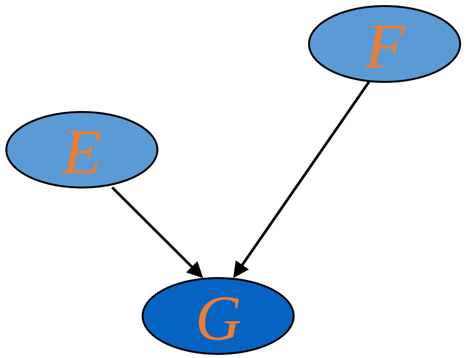
$$P(X_1, . . .,X_n) = \prod_{i=1}^{n} P(X_i \mid \mathbf{PA}(X_i)) \qquad [\mathbf{PA}(X_i) \text{ parents of } X_i]$$

Probabilistically quantify the BN using 'evidence':

- data
- literature
- model outputs
- expert judgement

etc





		G	
E	*F*	*normal*	*high*
	low	0.4	0.6
yes	*medium*	0.2	0.8
	high	0.1	0.9
	low	0.5	0.5
no	*medium*	0.6	0.4
	high	0.4	0.6

Extensions to BNs

- Object-oriented Bayesian networks
  - Used to model large, complex hierarchical systems
- Dynamic Networks
  - Used to model beliefs changing over time
  - Hidden Markov Models and Kalman Filters are special cases.
- Decision Networks (Influence Diagrams)
  - Used for decision making

Why Bayesian Networks?

1. Bring together disparate scientific knowledge
2. Create a ‘conceptual map’ of the scientific drivers
3. Quantify the map with data, model outputs, expert knowledge, etc
4. Identify key drivers
5. Explore scenarios of change
6. Understand impact of management and policy decisions

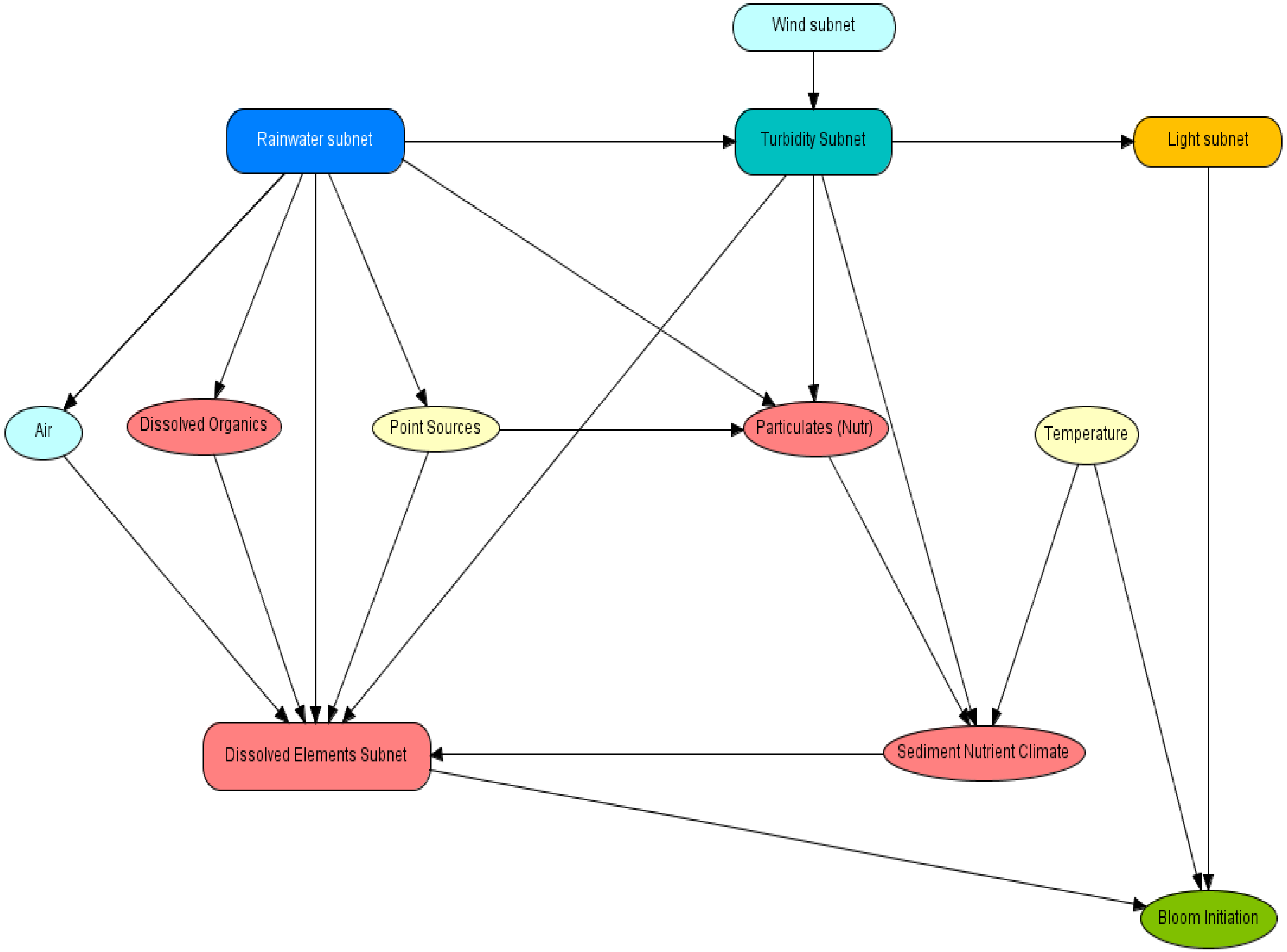
Case study: Bayesian network modelling of lyngbya

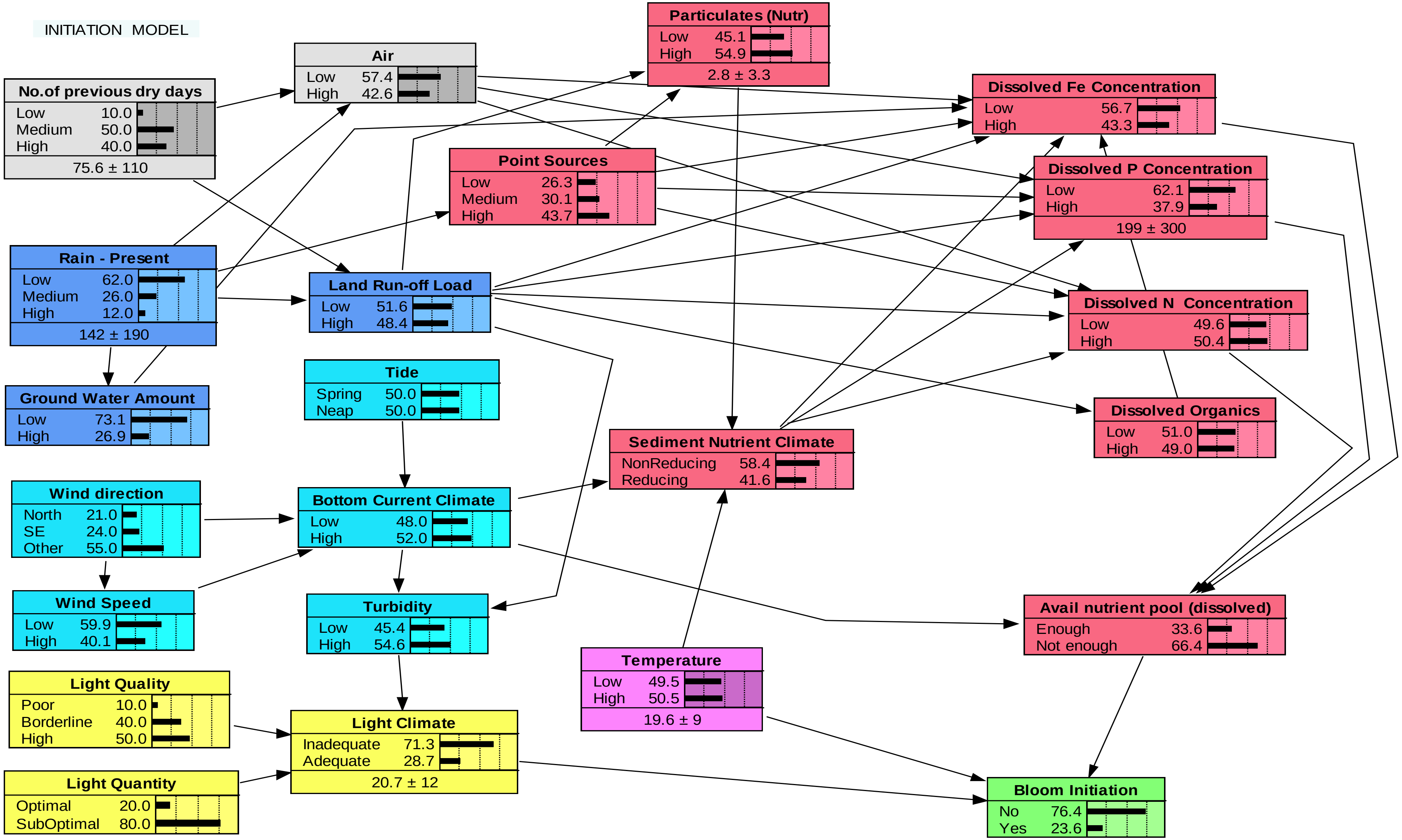
University of Queensland

What are the drivers of lyngbya?

What management actions should be taken?

Caboolture Shire Council


Wind subnet
Rainwater subnet
Turbidity Subnet
Light subnet
Air
Dissolved Organics
Point Sources
Particulates (Nutr)
Temperature
Dissolved Elements Subnet
Sediment Nutrient Climate
Bloom Initiation


INITIATION MODEL
No.of previous dry days
Low 10.0
Medium 50.0
High 40.0
75.6 ± 110
Air
Low 57.4
High 42.6
Particulates (Nutr)
Low 45.1
High 54.9
2.8 ± 3.3
Dissolved Fe Concentration
Low 56.7
High 43.3
Point Sources
Low 26.3
Medium 30.1
High 43.7
Dissolved P Concentration
Low 62.1
High 37.9
199 ± 300
Rain - Present
Low 62.0
Medium 26.0
High 12.0
142 ± 190
Land Run-off Load
Low 51.6
High 48.4
Dissolved N Concentration
Low 49.6
High 50.4
Ground Water Amount
Low 73.1
High 26.9
Tide
Spring 50.0
Neap 50.0
Dissolved Organics
Low 51.0
High 49.0
Sediment Nutrient Climate
NonReducing 58.4
Reducing 41.6
Wind direction
North 21.0
SE 24.0
Other 55.0
Bottom Current Climate
Low 48.0
High 52.0
Wind Speed
Low 59.9
High 40.1
Turbidity
Low 45.4
High 54.6
Avail nutrient pool (dissolved)
Enough 33.6
Not enough 66.4
Temperature
Low 49.5
High 50.5
19.6 ± 9
Light Quality
Poor 10.0
Borderline 40.0
High 50.0
Light Climate
Inadequate 71.3
Adequate 28.7
20.7 ± 12
Light Quantity
Optimal 20.0
SubOptimal 80.0
Bloom Initiation
No 76.4
Yes 23.6

Most influential factors

1. Available Nutrient Pool
2. Bottom Current Climate
3. Sediment Nutrients
4. Dissolved Iron
5. Dissolved Phosphorous
6. Light
7. Temperature

MANAGEMENT ACTIONS

“What-if” scenarios

Factor	Change in P(Bloom) (%)
Available Nutrient Pool	77 (3% - 80%)
Bottom Current Climate	28 (15% - 43%)
Sediment Nutrient Climate	17 (21% - 38%)
Dissolved Fe	16 (21% - 37%)
Dissolved P	15 (23% - 38%)
Light Climate	14 (18% - 32%)
Temperature	14 (21% - 35%)
Dissolved N	13 (22% - 35%)
Rain – present	10 (25% - 35%)
Light Quantity	9 (21% - 30%)

Translating science to management

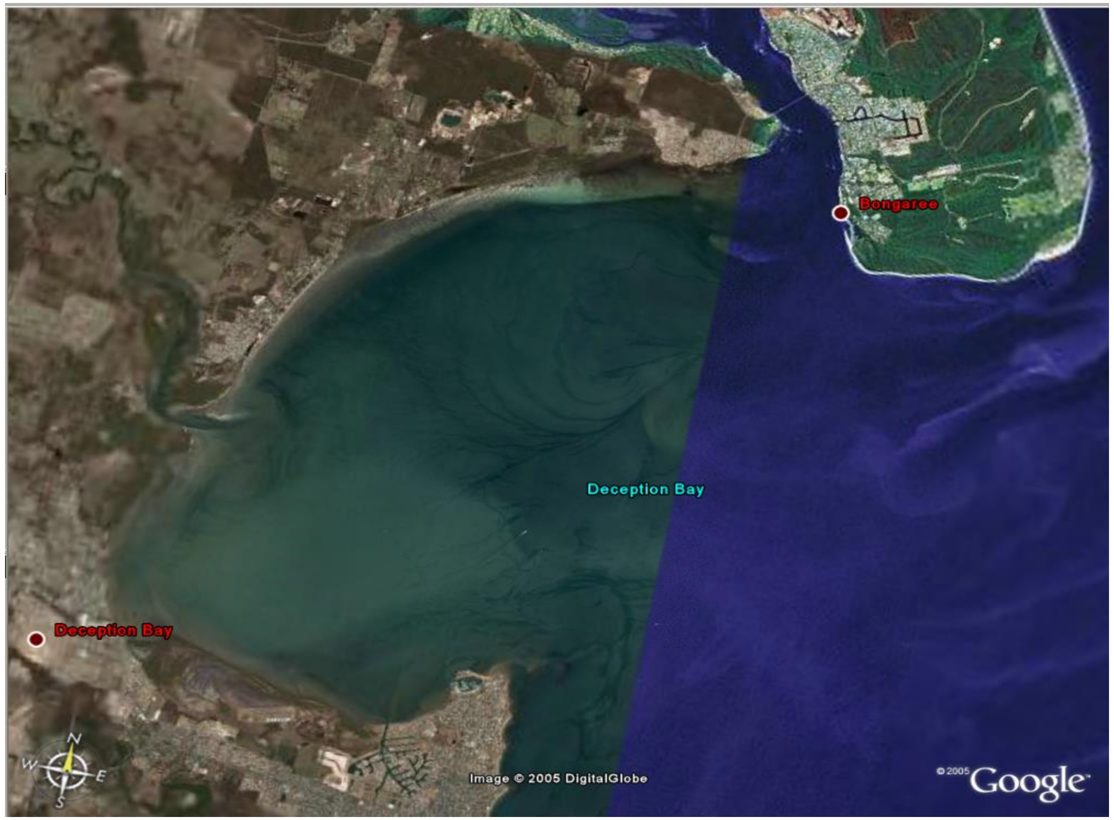
Management Model

↓

Science BN Model

↓

P(Bloom Initiation)

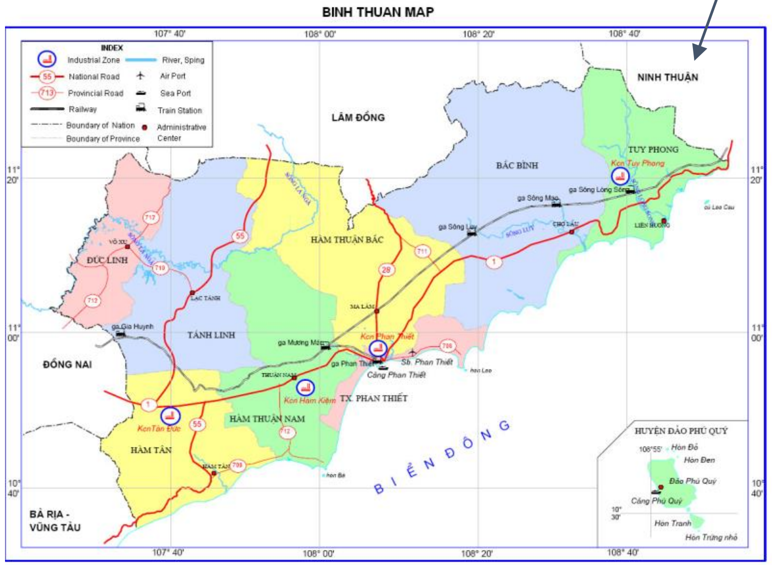


Evaluation of factors, scenario assessment
Integration of information, adaptive updates

Case Study 2: Vietnam case study - Dragon fruit

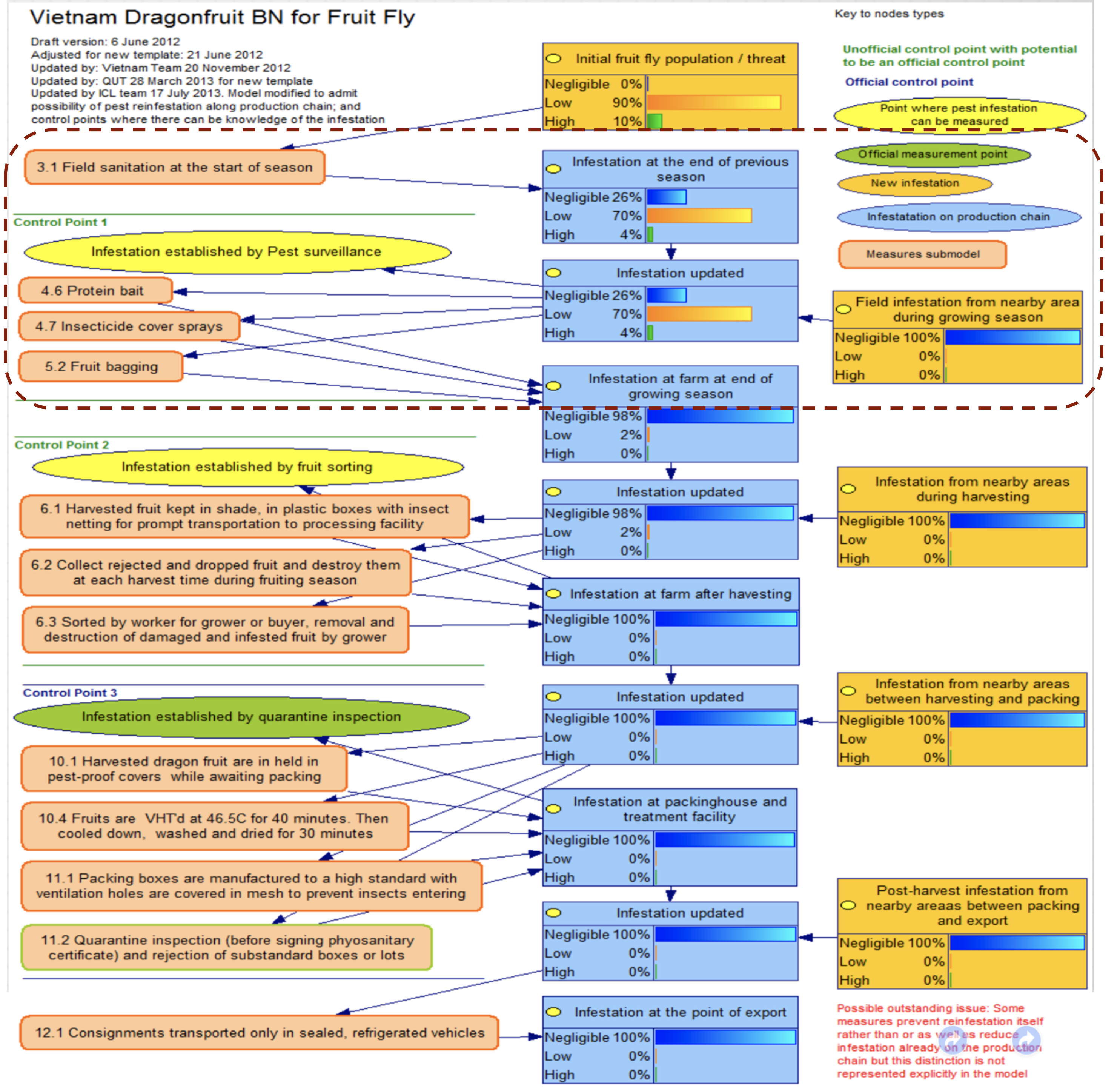
Binh Thuan province:
- 20,000 ha
- 18,000 growers
- 500,000 tonnes/yr
- $0.20-$1.00/kg farm gate

Large semi-official export to China
Small, high quality markets in Korea, EU, US

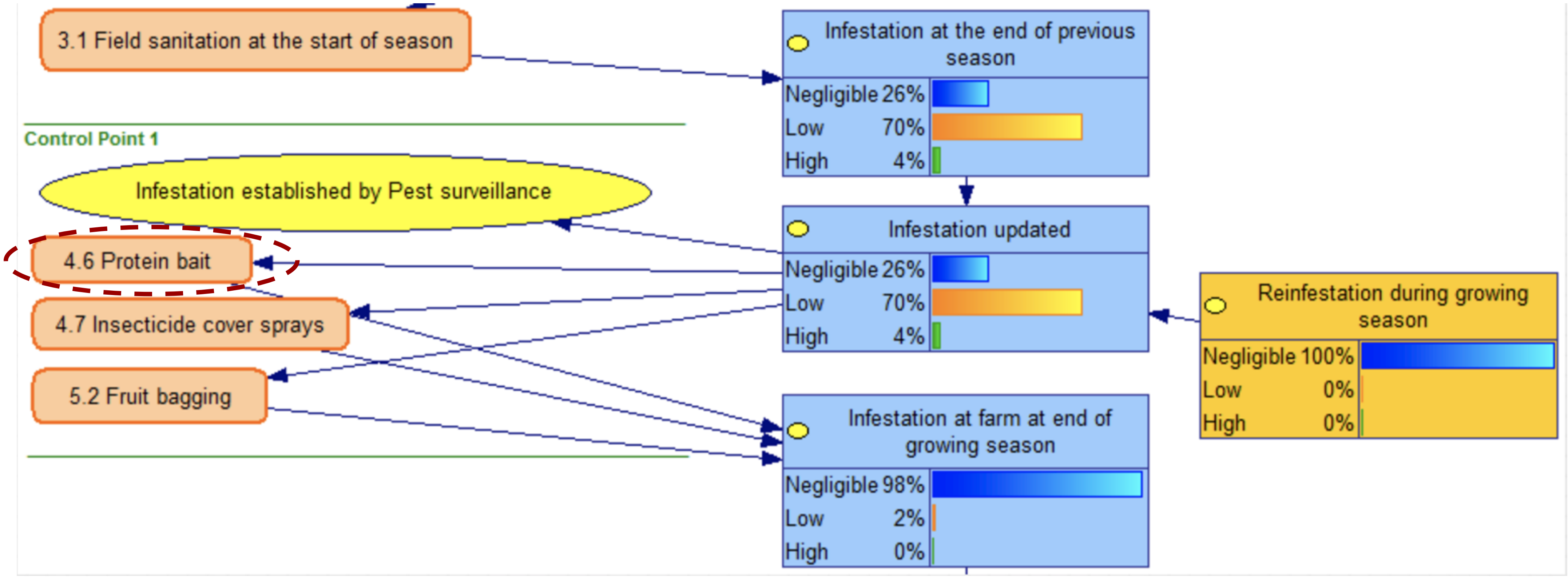


Control point BN

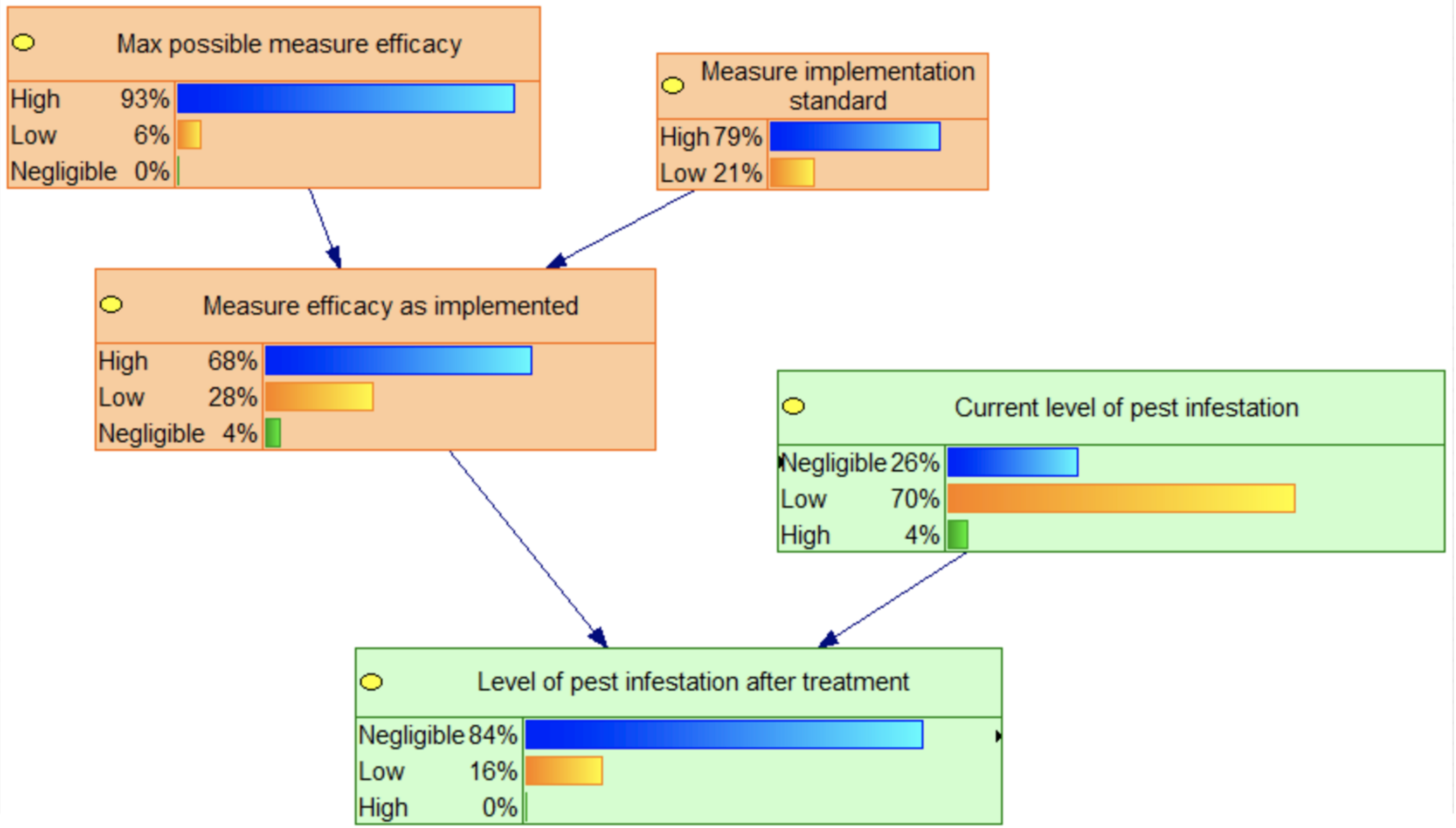
1. Field measures, with pest monitoring
2. Harvest sorting and hygiene
3. Inspection and sorting at packing



BN control point 1 - field measures with monitoring



Protein bait BN sub-model



Decision support spreadsheet

Efficacy & implementation from stake-holder/expert elicitation

	Efficacy	Uncertainty		Implementation	Uncertainty	
4.7 Protein bait **Bait**	High	Low		Easy	Low	

Systematic elicitation for all 14 measures on common scales

TABLE C1. Description of candidate measures (these may be used alone or with other measures)

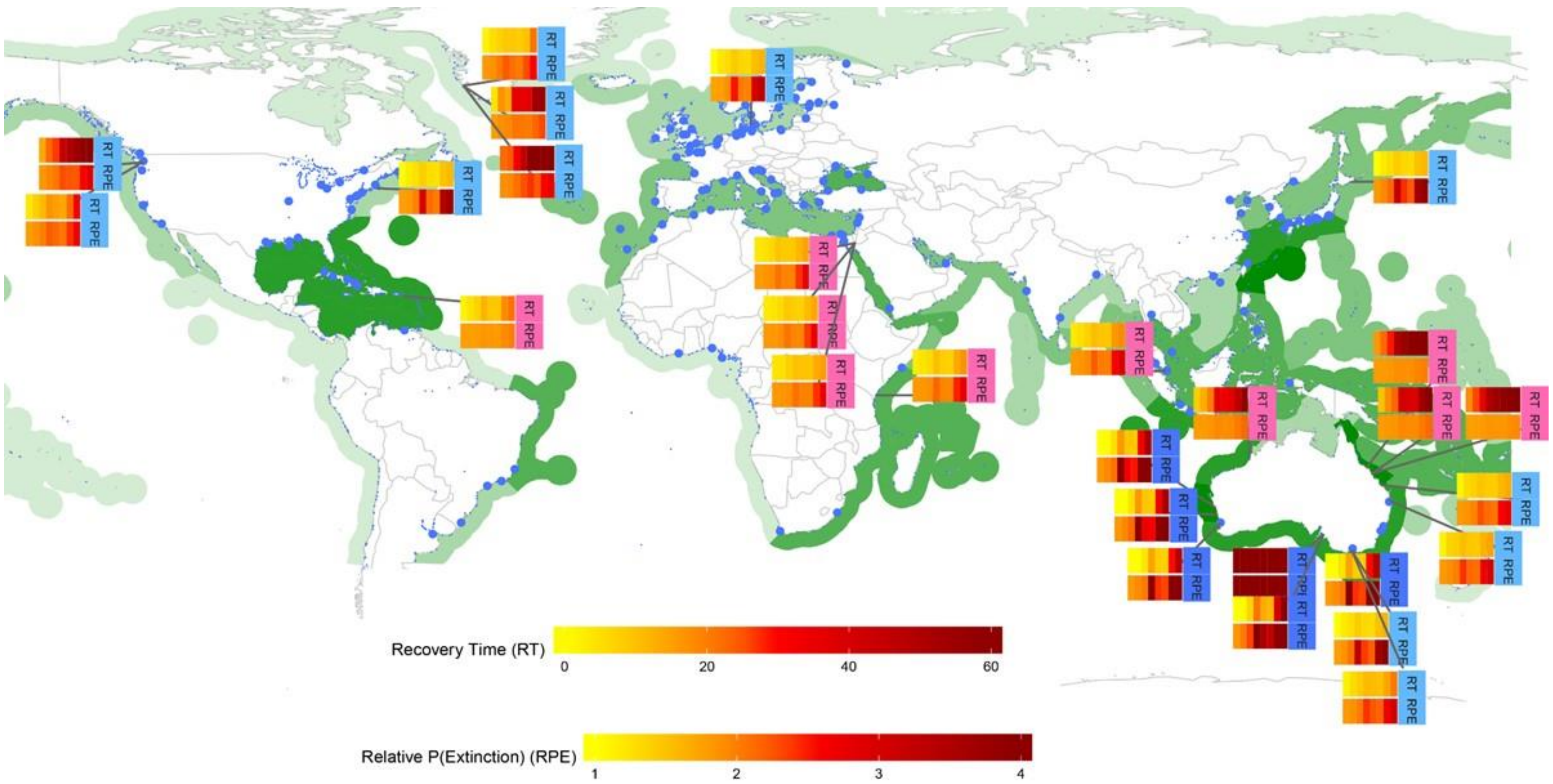
	Risk management measures available (automatically read in from Table B2)	Efficacy: 1.1 a) What is its potential contribution to risk reduction?	1.1 b) Uncertainty	Graphic	Verification: 1.2 a) The measure can be verified?	1.2 b) Uncertainty	Graphic
i	*3.1 Field sanitation at the start of season*	Low	Low		Easy	Low	
ii	*4.5 Pests surveillance*	Medium	Low		Easy	Low	
iii	*4.6 Pheromone traps/male annihilation technique (MAT)*	High	Low		Easy	Low	
iv	*4.7 Protein bait*	High	Low		Easy	Low	
v	*4.8 Insecticide cover spray*	Low	Low		Easy	Low	
vi	*5.2 Fruit bagging*	High	Very low		Very easy	Very low	
vii	*6.1 Harvested fruit kept in shade, in plastic boxes with insect netting for prompt transportation to processing facility*	High	Low		Easy	Low	
viii	*6.2 Collect dropped fruit and destroy them at each harvest time during fruiting season*	High	Medium		With some difficulty	Low	
ix	*6.3 Sorted by woker for grower or buyer, removal and destruction of damaged and infested fruit by grower*	Low	Low		With some difficulty	Low	
x	*10.1 Harvested dragon fruit are in held in pest-proof covers while awaiting packing*	Medium	Low		Easy	Low	
xi	*10.4 Fruits are VHT'd at 46.5C for 40 minutes . Then cooled down, washed and dried for 30 minutes.*	Very high	Very low		Very easy	Very low	
xii	*11.1 Packing boxes are manufactured to a high standard with ventilation holes are covered in mesh to prevent insects entering.*	Very high	Low		Easy	Low	
xiii	*11.2 Quarantine inspection (before signing phyosanitary certificate)*	Very high	Low		Very easy	Low	
xiv	*12.1 Consignments transported only in sealed, refrigerated vehicles.*	High	Low		Easy	Low	

Case Study 3: Seagrass Case Study

- Seagrass ecosystems
  - Habitat, $1.9 trillion in ecosystem services, carbon, declining at a rate of ~110km^2 since 1980 (Waycott et al, 2009)

- Need to manage threats to marine ecosystems
  - Urban and agricultural runoff
  - Infrastructure development
  - Dredging

Courtesy of: Gary Kendrick



World distribution of seagrasses (green) & ports (blue dots).

Heat maps show average recovery time (bottom panel) & average ratio of extinction risk to baseline risk (top panel).

Bars correspond to dredging periods: 1, 2, 3, 6, 9, 12 mths

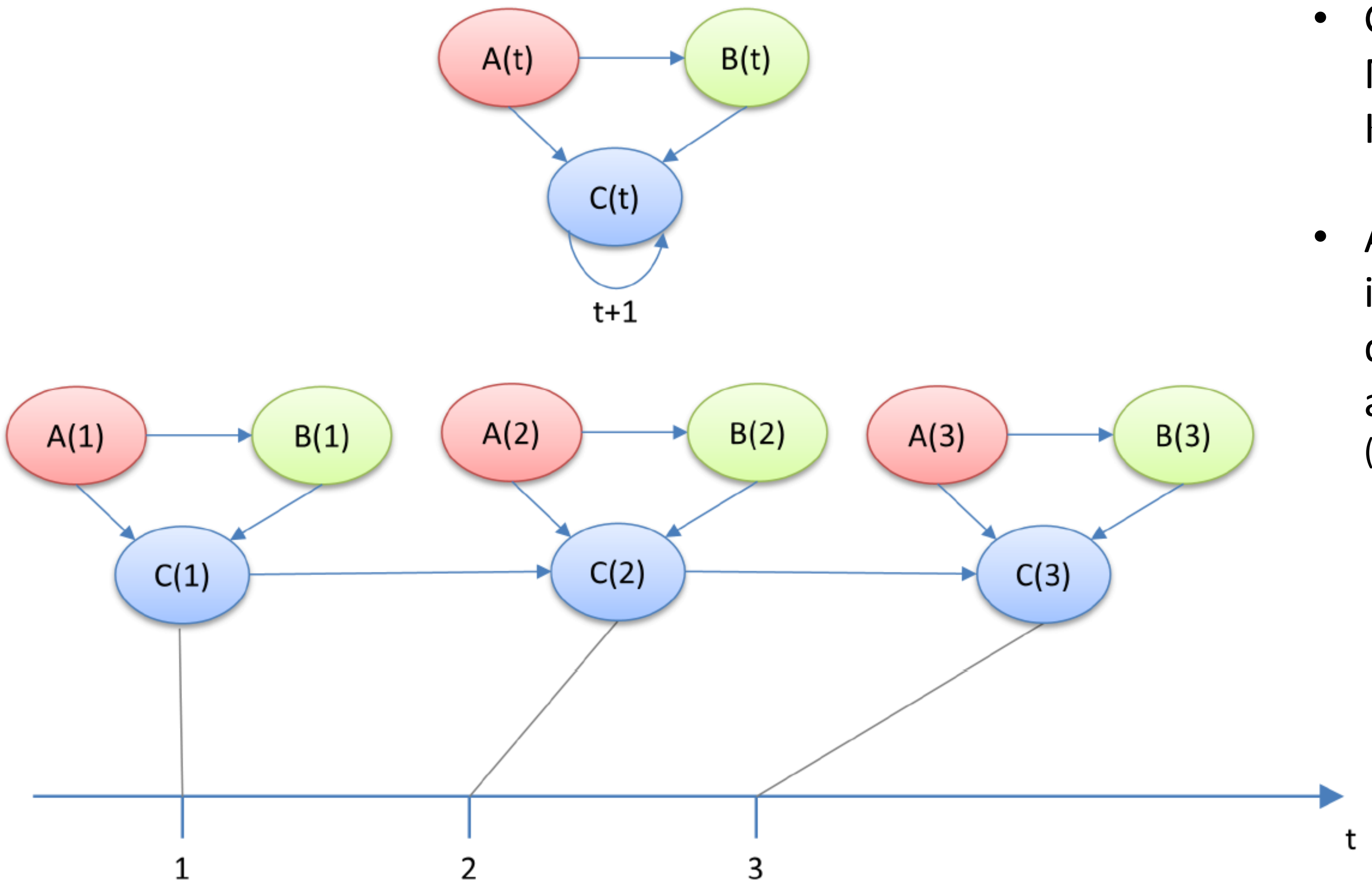
Labels coloured by genera – *Halophila*, *Zostera*, *Amphibolis*.

Key outcome is resilience

Resilience (Levin, 2008; Holling, 1973)

- Resistance, loss of individuals and/or species as the result of stress
- Recovery, expected recovery time
- Persistence, risk of local extinction (probability of zero population of species)

Dynamic Bayesian Networks

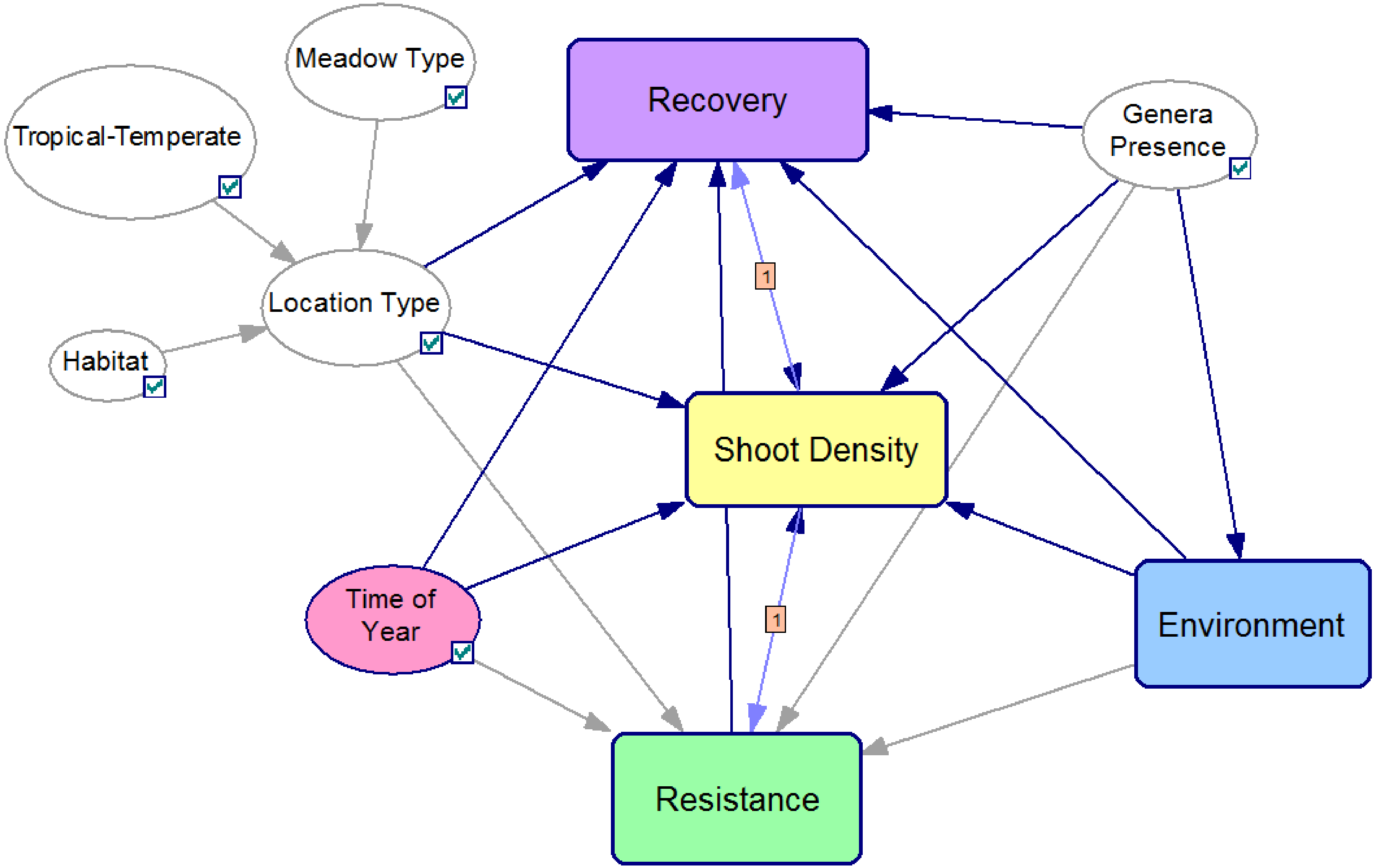


- Generalised form of Hidden Markov Models (HMMs) and Kalman filter models

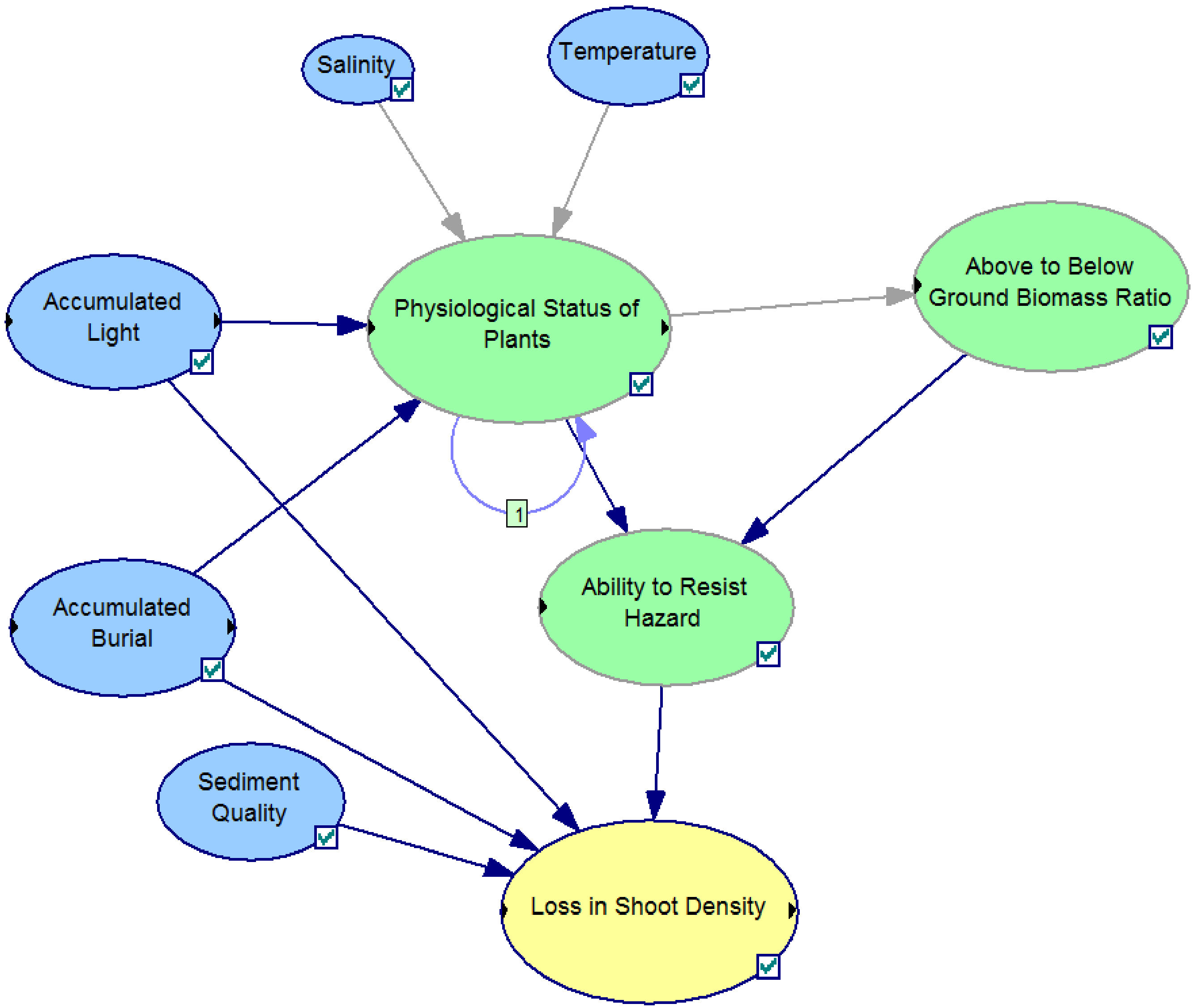
- Allows state space representation in factored form rather than single discrete random variable, arbitrary probability distributions (Murphy, 2002)

Conditional probabilities

$$P(\mathbf{ShootDensity}(t)|\{\mathbf{genera},\mathbf{location},\mathbf{light},...\})$$



Seagrass Model



Forwards and backwards inference

Forwards

$$\pi(s_t|\boldsymbol{y}_t) = P_{y_t,s_t} \sum_{j=1}^{K} \lambda_{j,s_t} \pi(S_{t-1} = j|\boldsymbol{y}_{t-1})$$

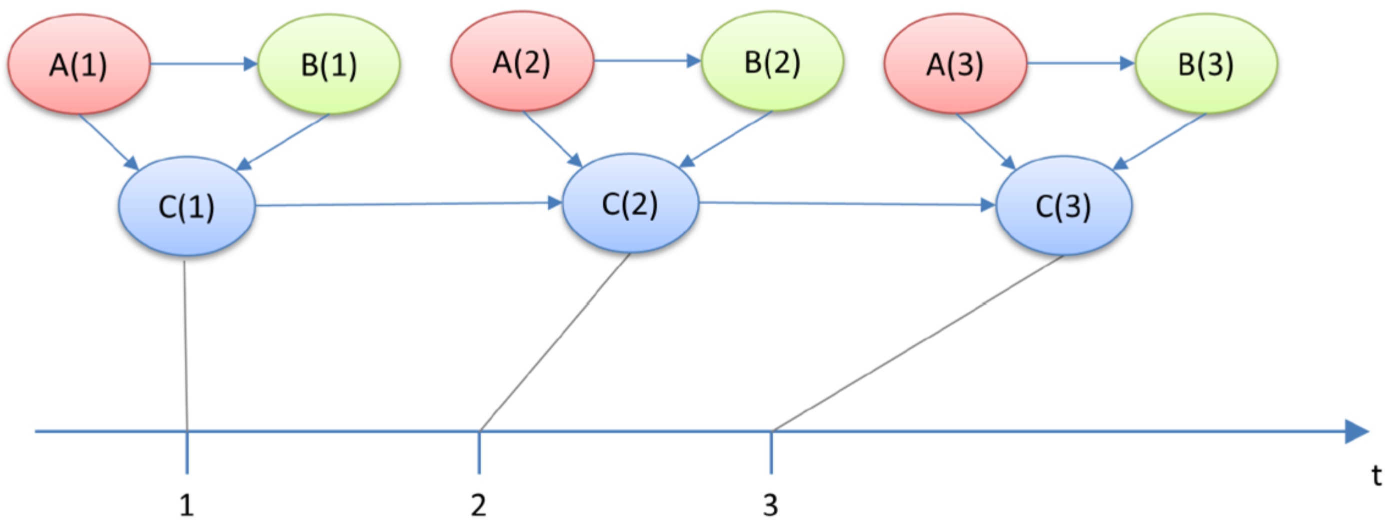
$$\pi(s_{t+1}|\boldsymbol{y}_t) = \sum_{j=1}^{K} \lambda_{j,s_{t+1}} \pi(S_t = j|\boldsymbol{y}_t)$$

Backwards

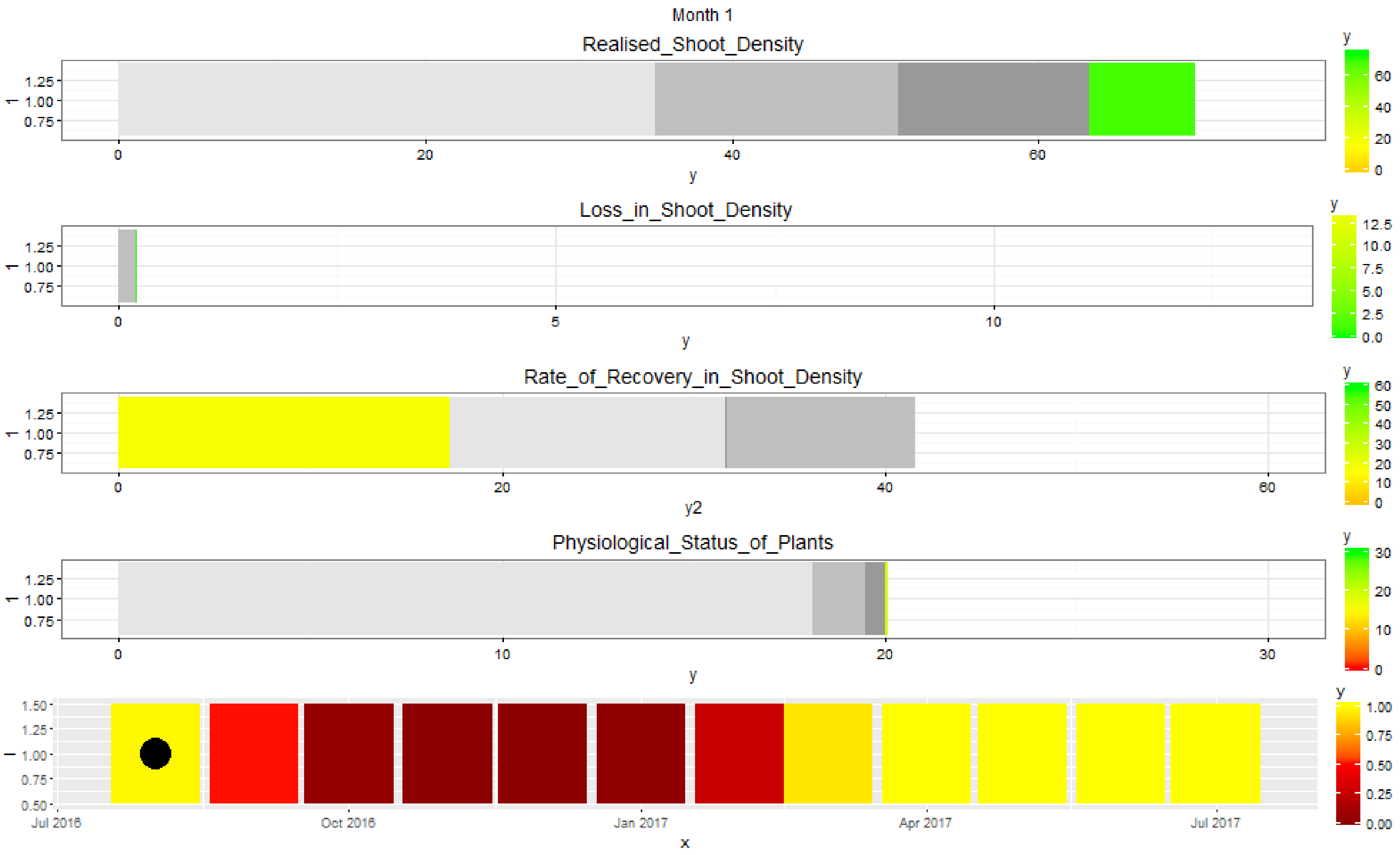
$$\pi(s_t|s_{t+1}, \boldsymbol{y}_{t+1}) = \lambda_{s_t,s_{t+1}} \frac{\pi(s_t|\boldsymbol{y}_t)}{\pi(s_{t+1}|\boldsymbol{y}_t)}$$

Link Node Based Non-Homogeneous DBN Inference

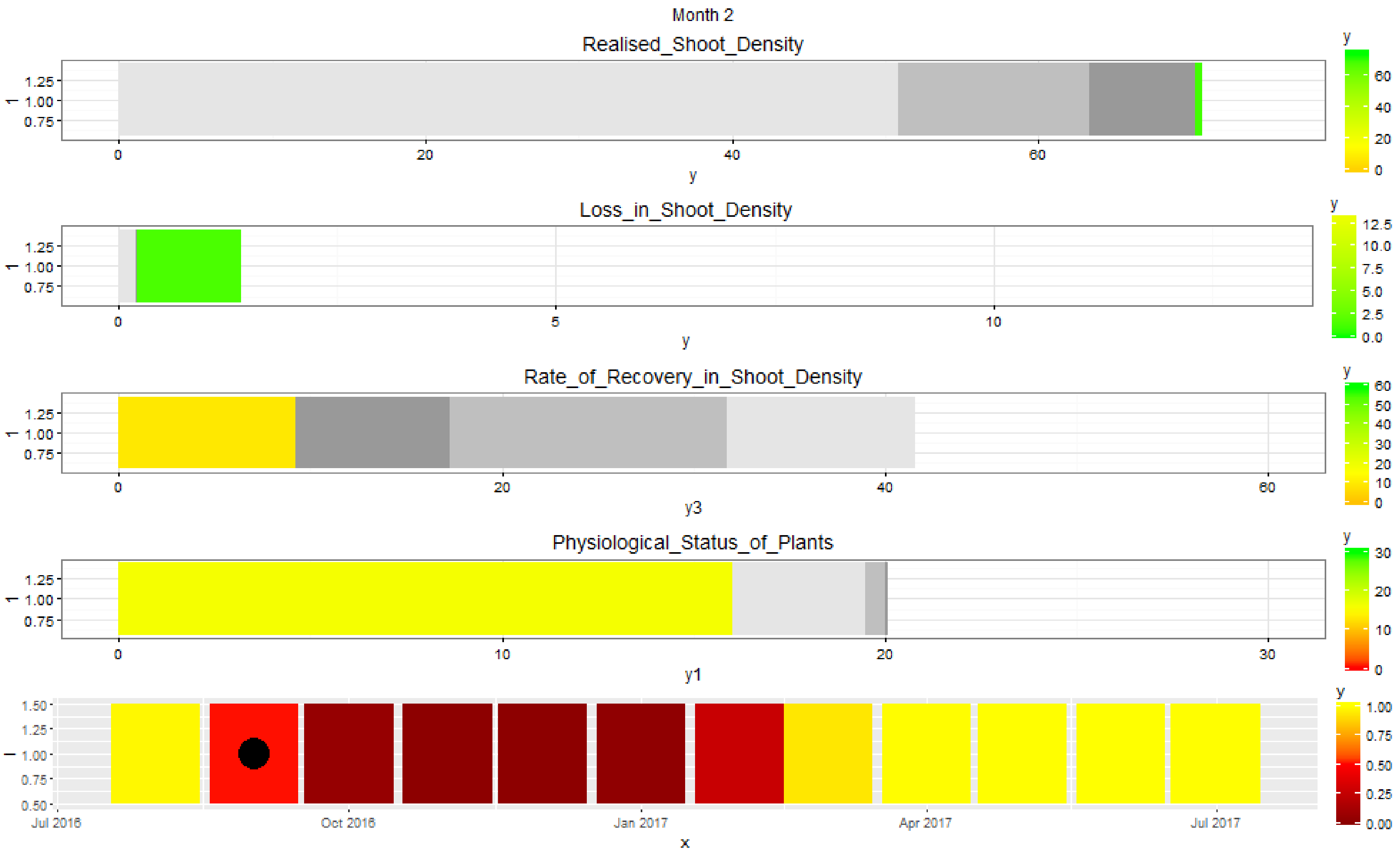
- Complex systems characterised by “small world networks” (Watts, 1998)
- Use link nodes (small number of nodes connecting between time slices)
- Use dynamic forward-backward algorithm



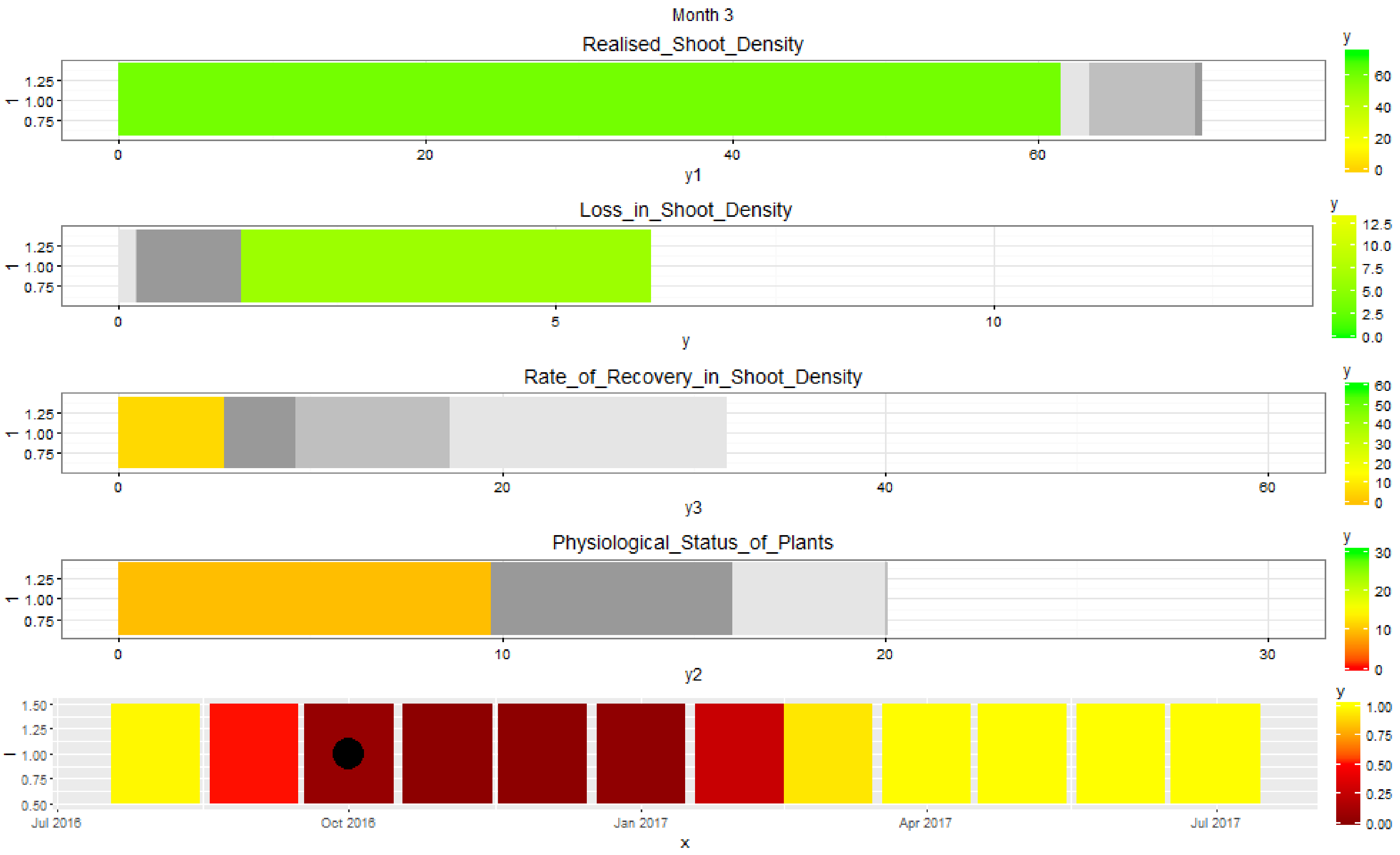
Results

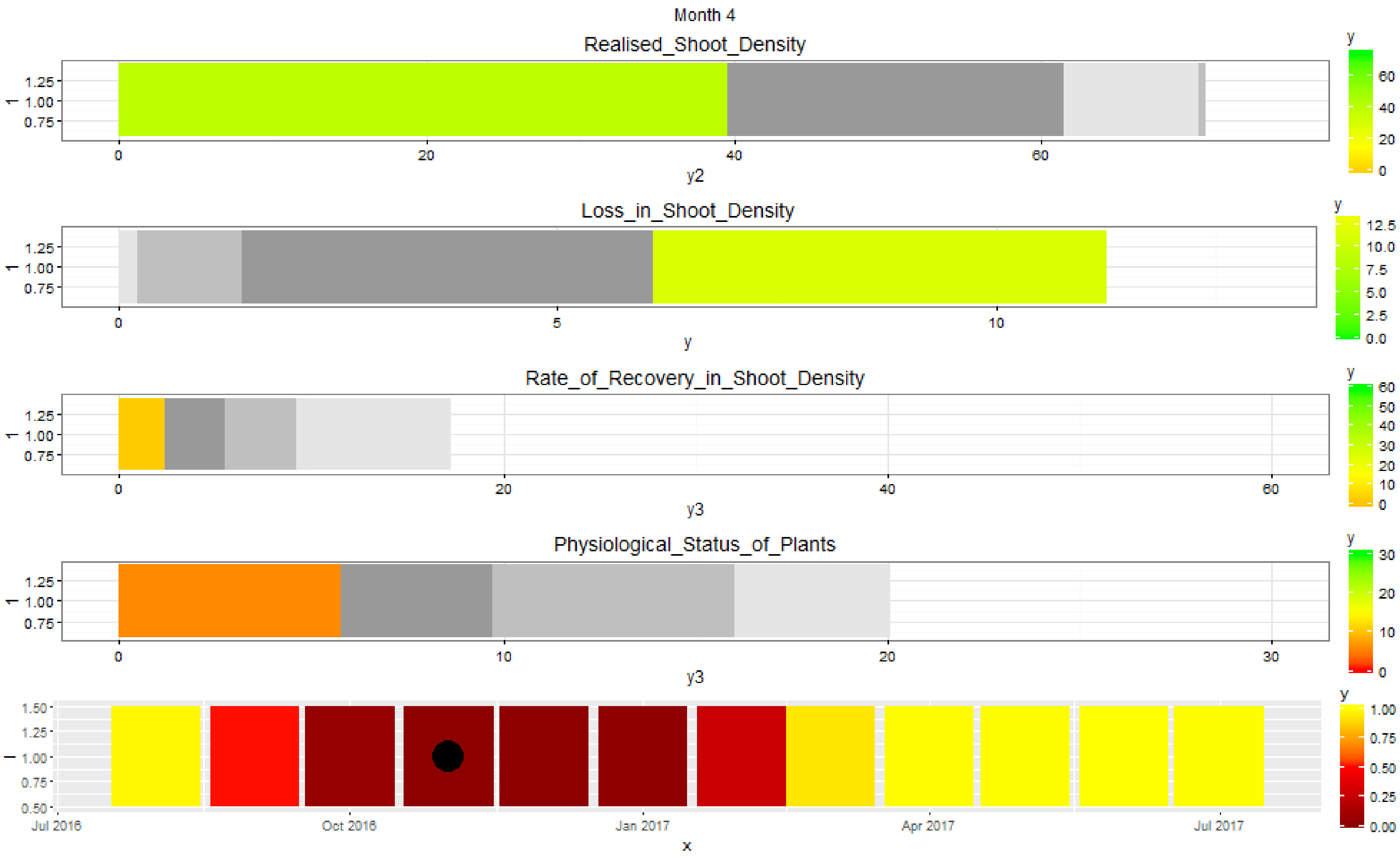


Month 1
Realised_Shoot_Density
Loss_in_Shoot_Density
Rate_of_Recovery_in_Shoot_Density
Physiological_Status_of_Plants
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017

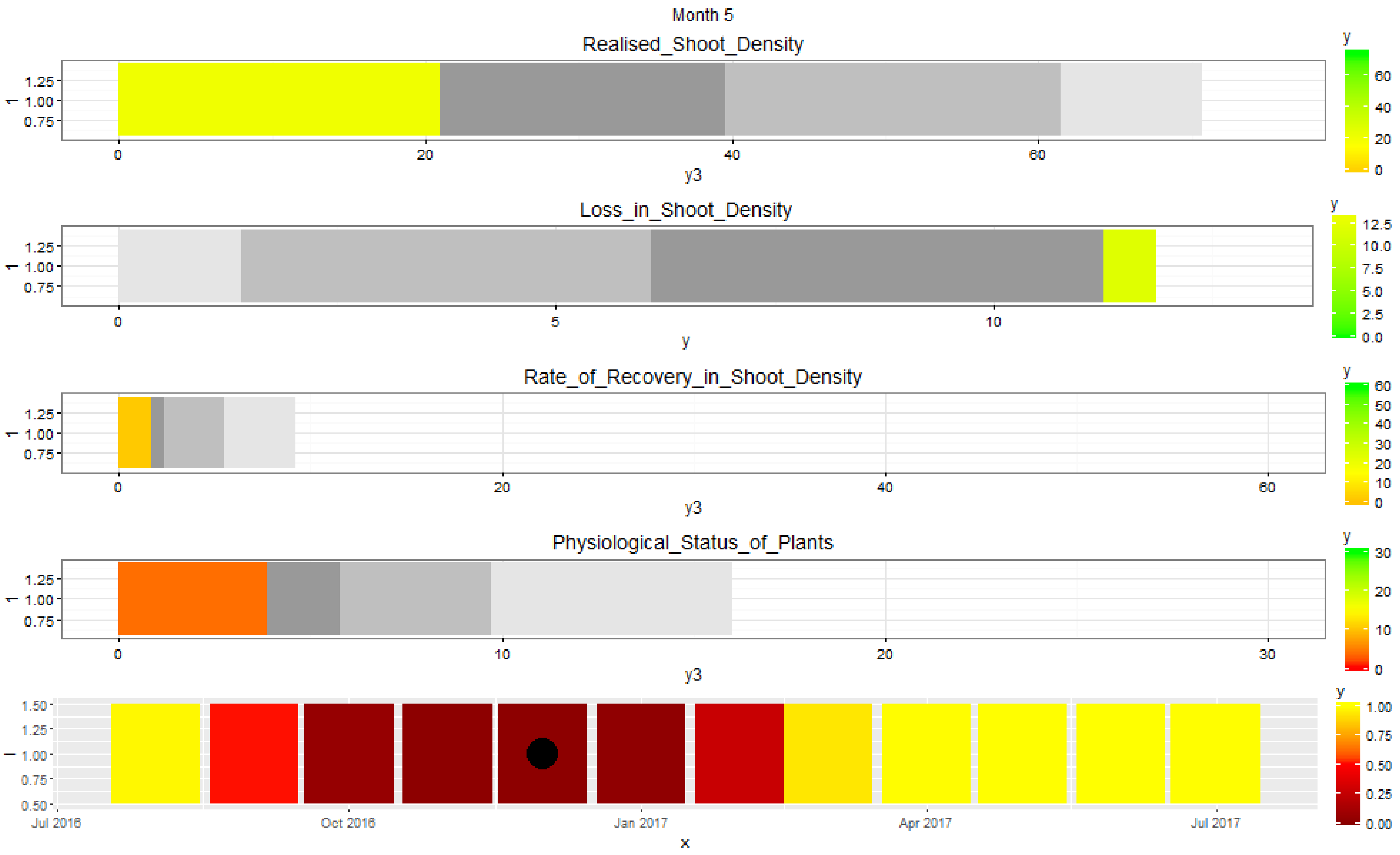


Month 2
Realised_Shoot_Density
Loss_in_Shoot_Density
Rate_of_Recovery_in_Shoot_Density
Physiological_Status_of_Plants
y
y3
y1
x
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017

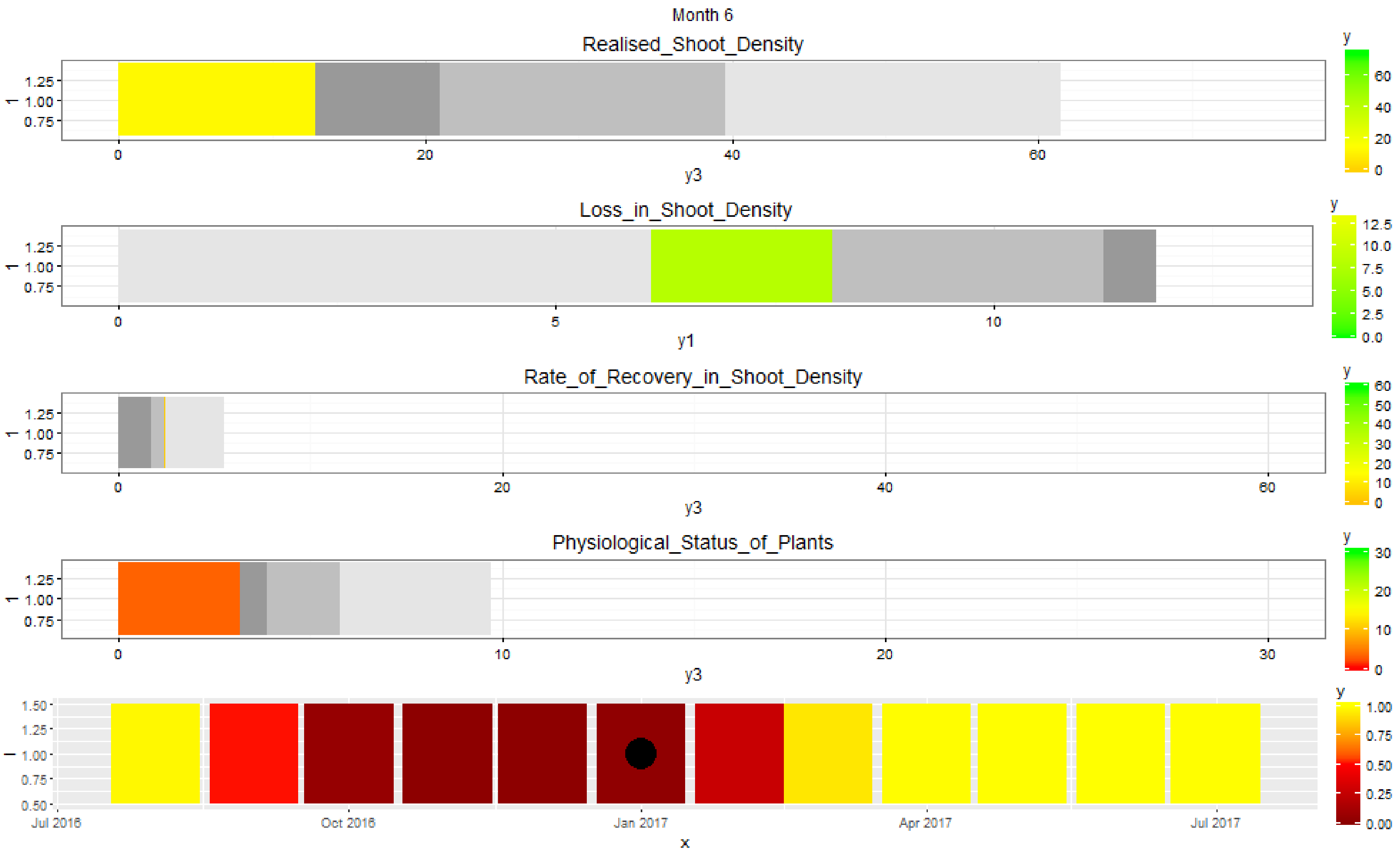

Month 3
Realised_Shoot_Density
y1
Loss_in_Shoot_Density
y
Rate_of_Recovery_in_Shoot_Density
y3
Physiological_Status_of_Plants
y2
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017
x



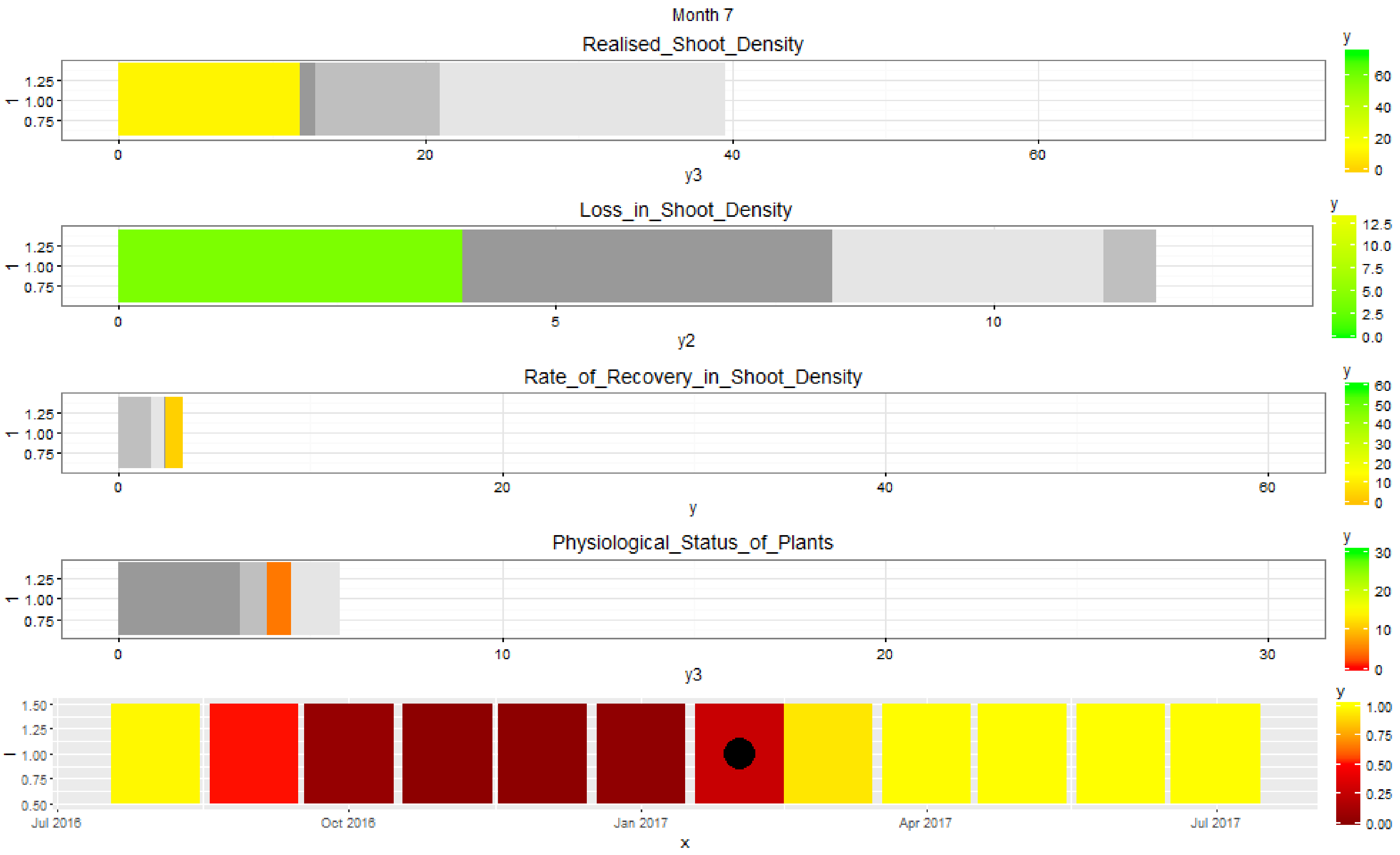
Month 4
Realised_Shoot_Density
y2
Loss_in_Shoot_Density
y
Rate_of_Recovery_in_Shoot_Density
y3
Physiological_Status_of_Plants
y3
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017
x



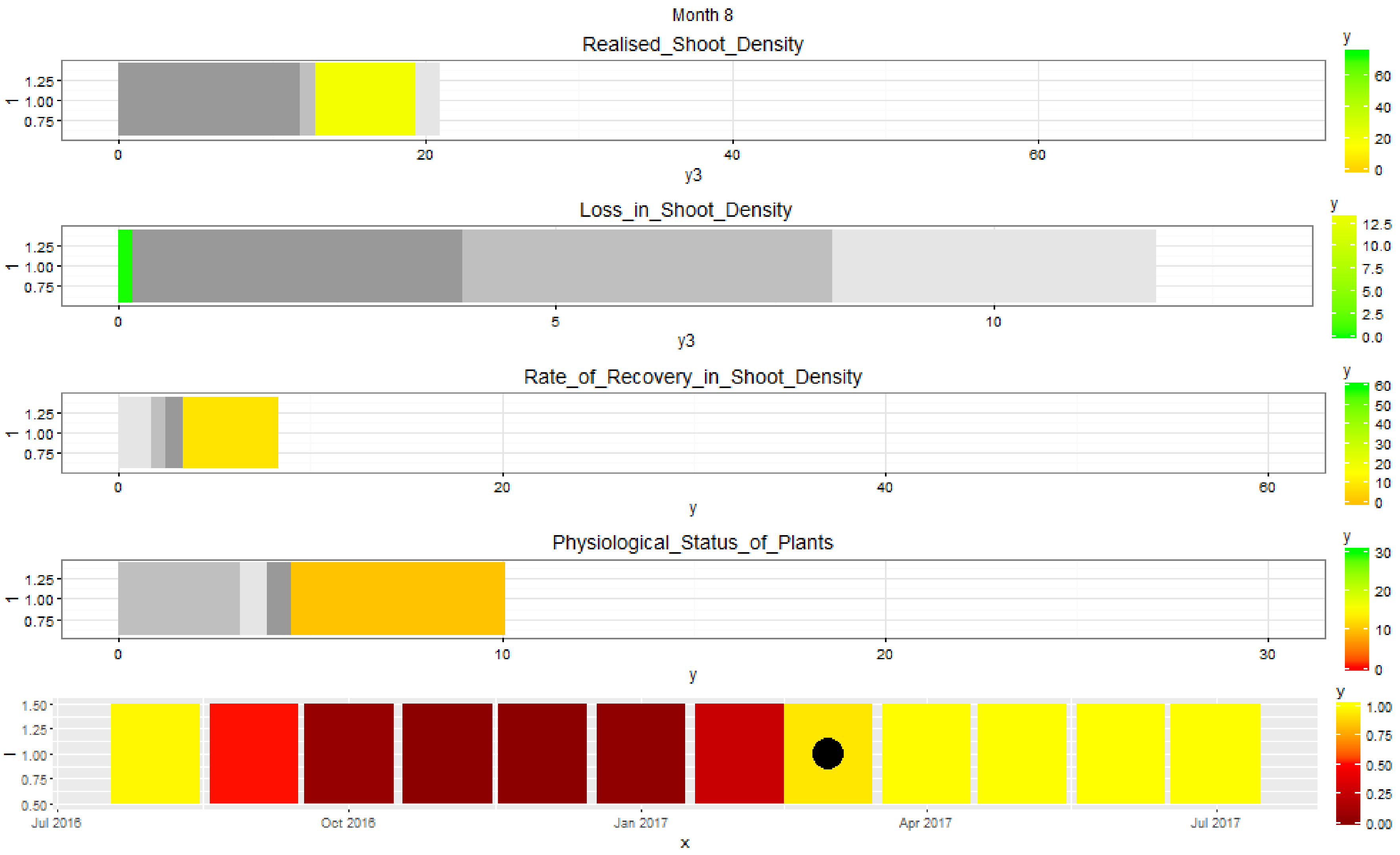
Month 5
Realised_Shoot_Density
Loss_in_Shoot_Density
Rate_of_Recovery_in_Shoot_Density
Physiological_Status_of_Plants
y3
y
x
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017



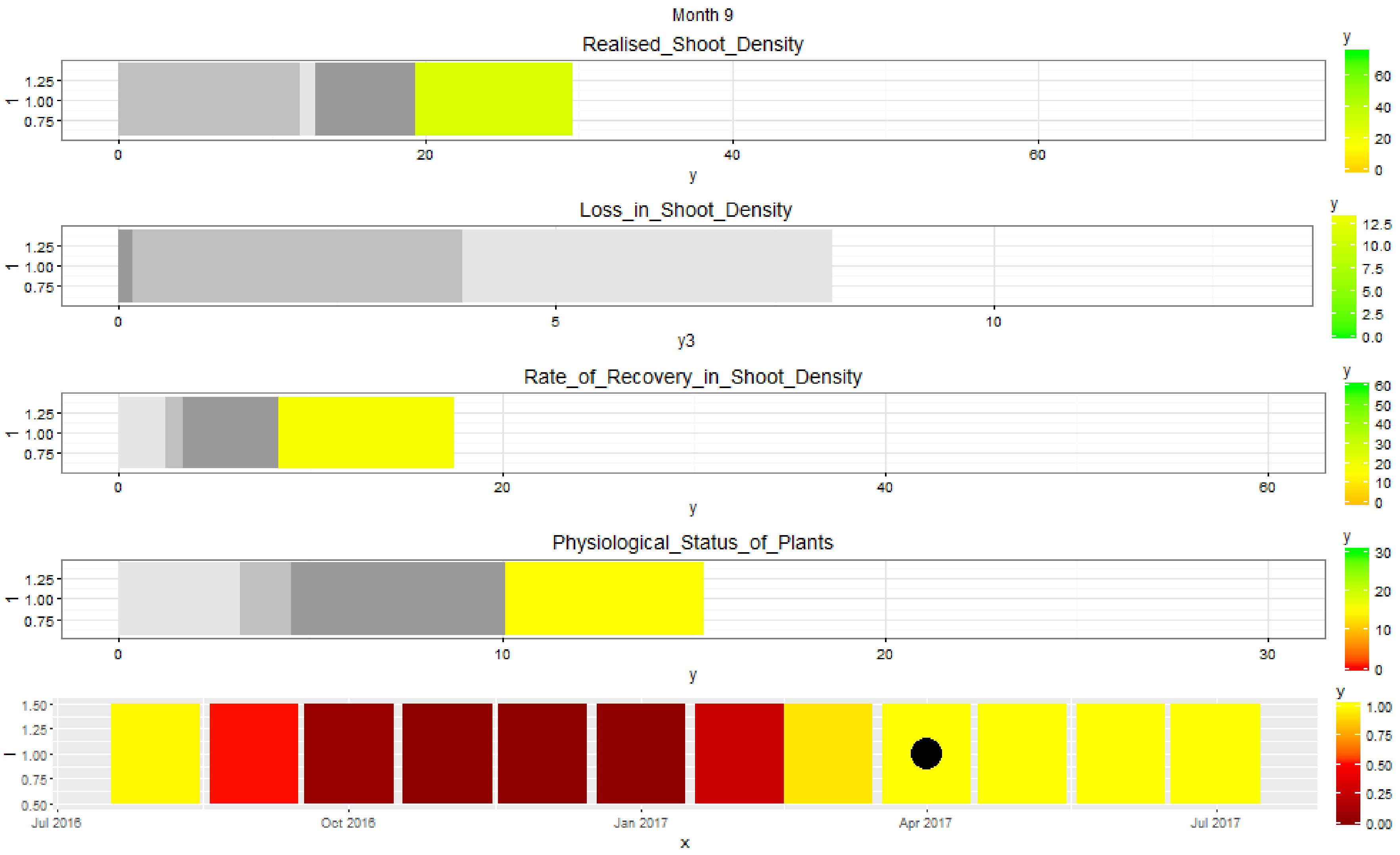
Month 6
Realised_Shoot_Density
Loss_in_Shoot_Density
Rate_of_Recovery_in_Shoot_Density
Physiological_Status_of_Plants
y3
y1
x
y
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017



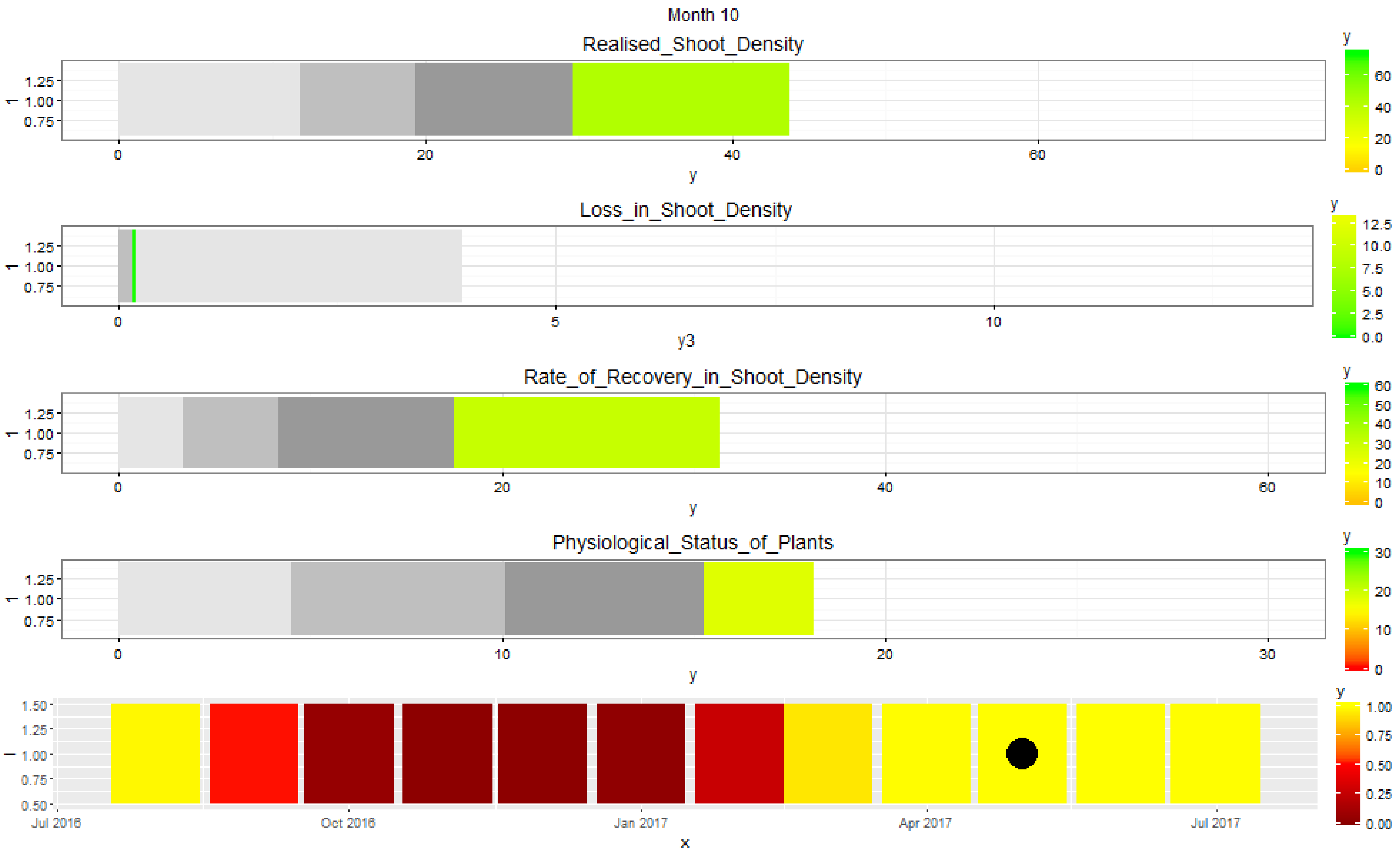
Month 7
Realised_Shoot_Density
y3
Loss_in_Shoot_Density
y2
Rate_of_Recovery_in_Shoot_Density
y
Physiological_Status_of_Plants
y3
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017
x

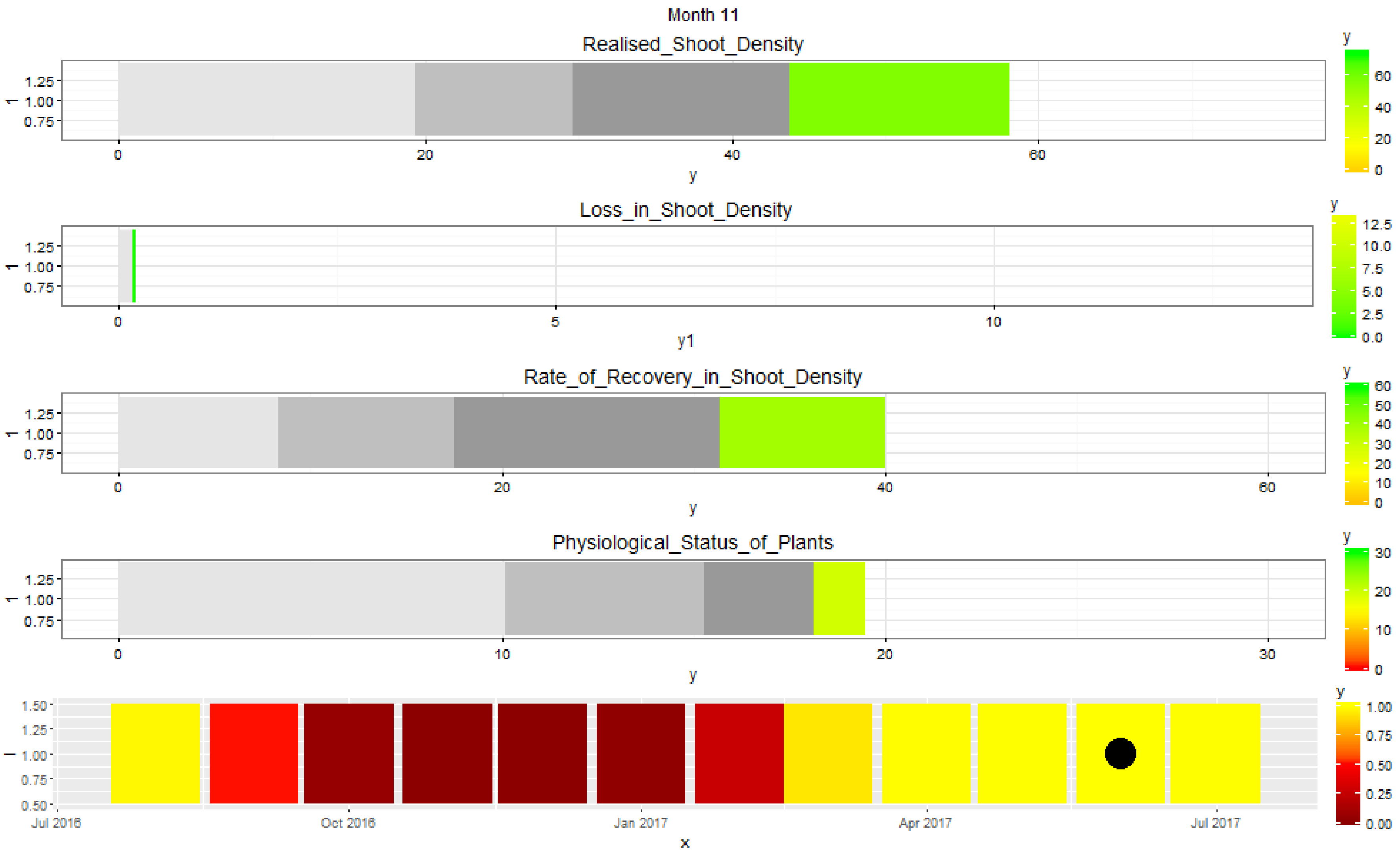


Month 8
Realised_Shoot_Density
Loss_in_Shoot_Density
Rate_of_Recovery_in_Shoot_Density
Physiological_Status_of_Plants
y3
y
x
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017

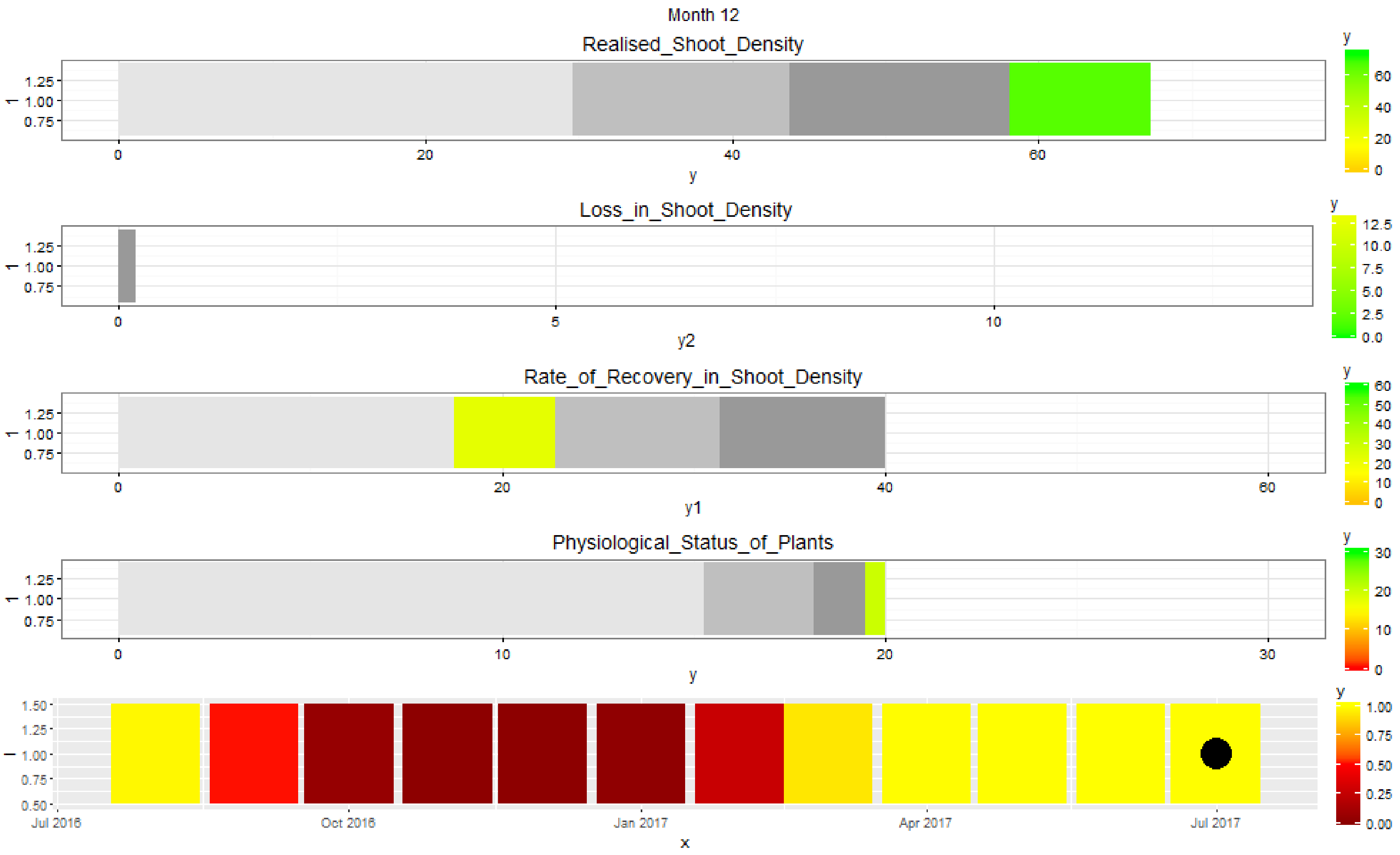


Month 9
Realised_Shoot_Density
Loss_in_Shoot_Density
Rate_of_Recovery_in_Shoot_Density
Physiological_Status_of_Plants
y
y3
x
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017

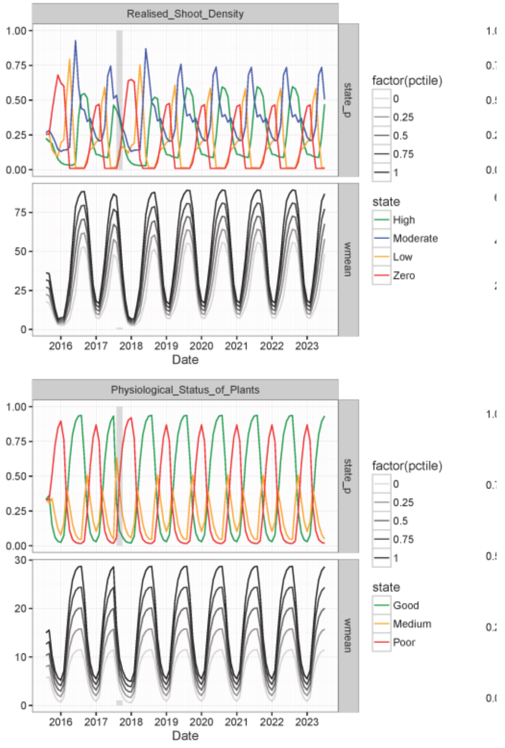

Month 10
Realised_Shoot_Density
Loss_in_Shoot_Density
Rate_of_Recovery_in_Shoot_Density
Physiological_Status_of_Plants
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017
x
y
y3

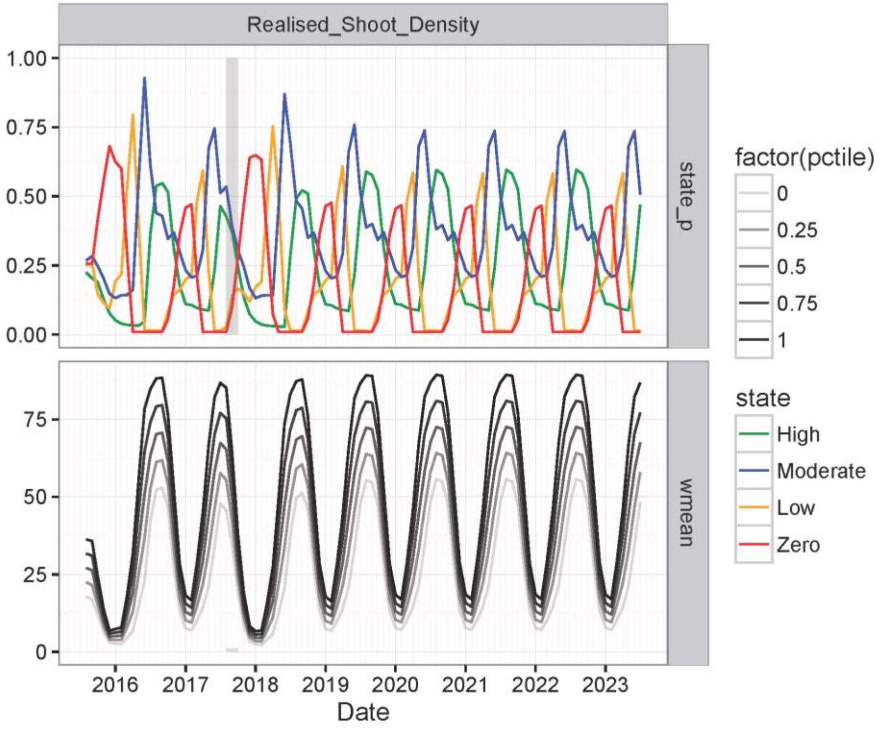


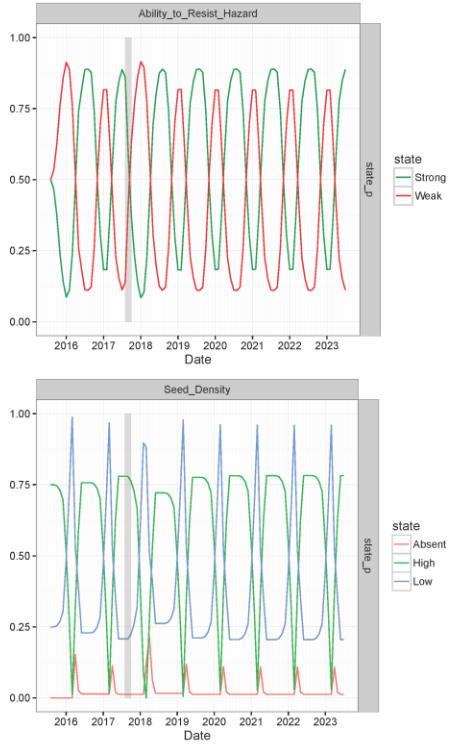
Month 11
Realised_Shoot_Density
Loss_in_Shoot_Density
Rate_of_Recovery_in_Shoot_Density
Physiological_Status_of_Plants
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017


Month 12
Realised_Shoot_Density
y
Loss_in_Shoot_Density
y2
Rate_of_Recovery_in_Shoot_Density
y1
Physiological_Status_of_Plants
y
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017
x

Whole-of-Systems Response



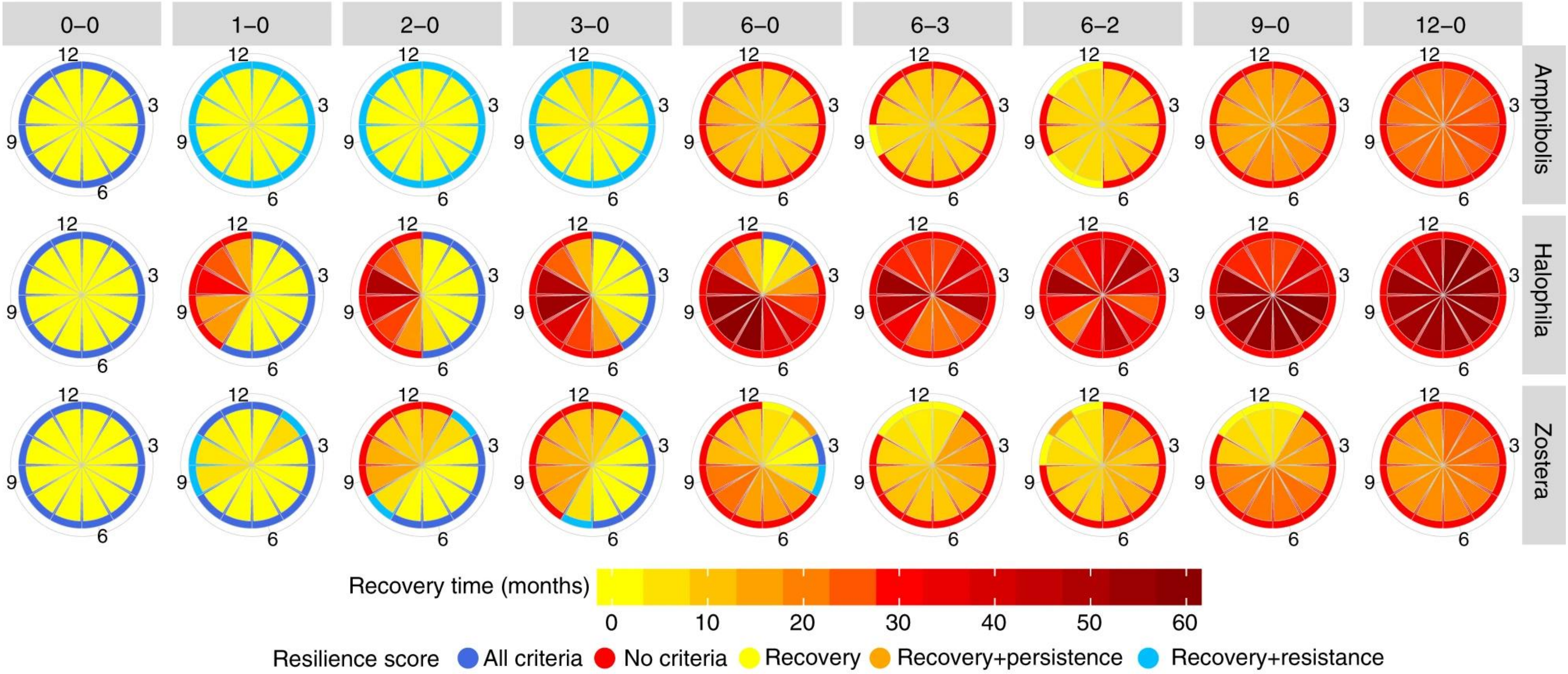




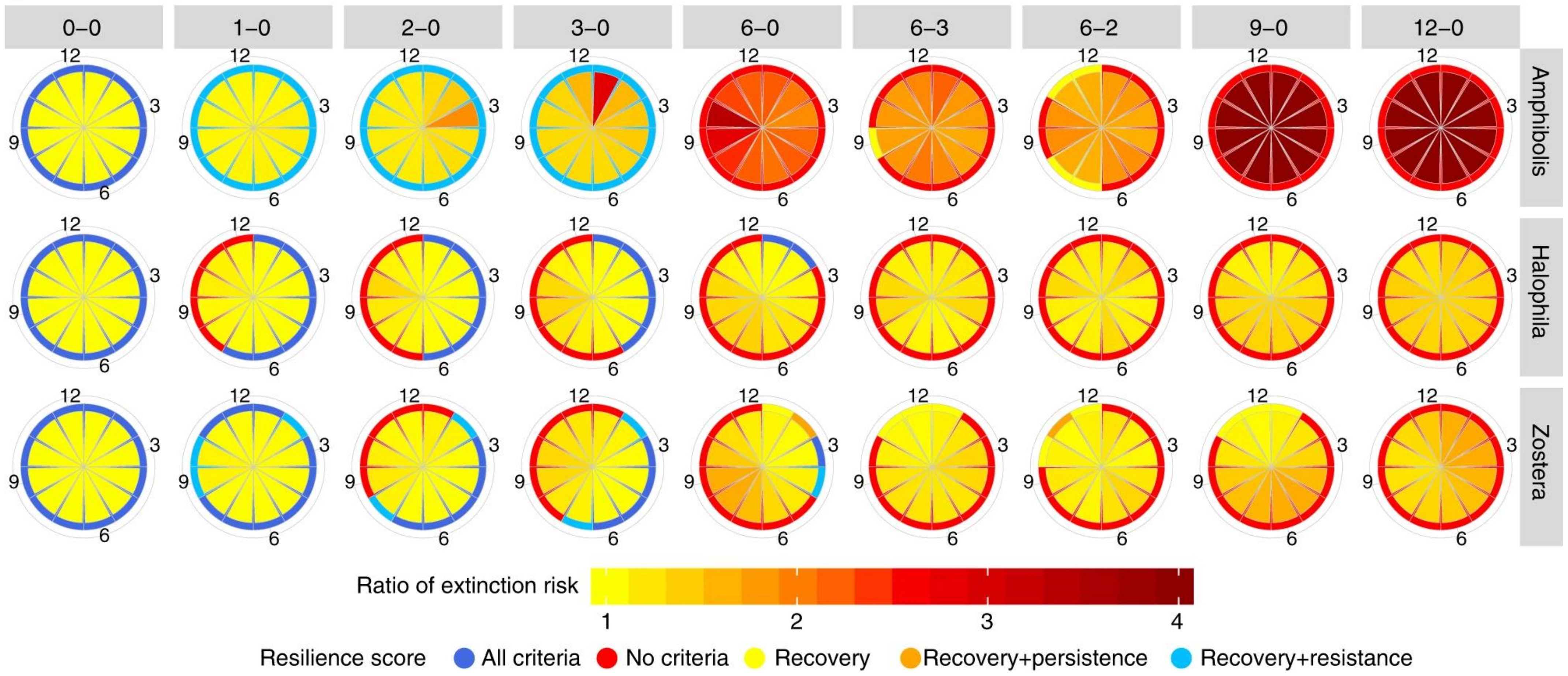
Resilience Criteria

- Resilience criteria using baselines (Levin, 2008, Halpern, 2007)
  - Resistance, loss of individuals and/or species as the result of stress
    - **80% of baseline population in that month**
  - Recovery, expected recovery time
    - **Within 6 months**
  - Persistence, risk of local extinction (probability of zero population of species)
    - **Within 2.5% of baseline risk of zero**

Ecological Windows



Ecological Windows



So What?

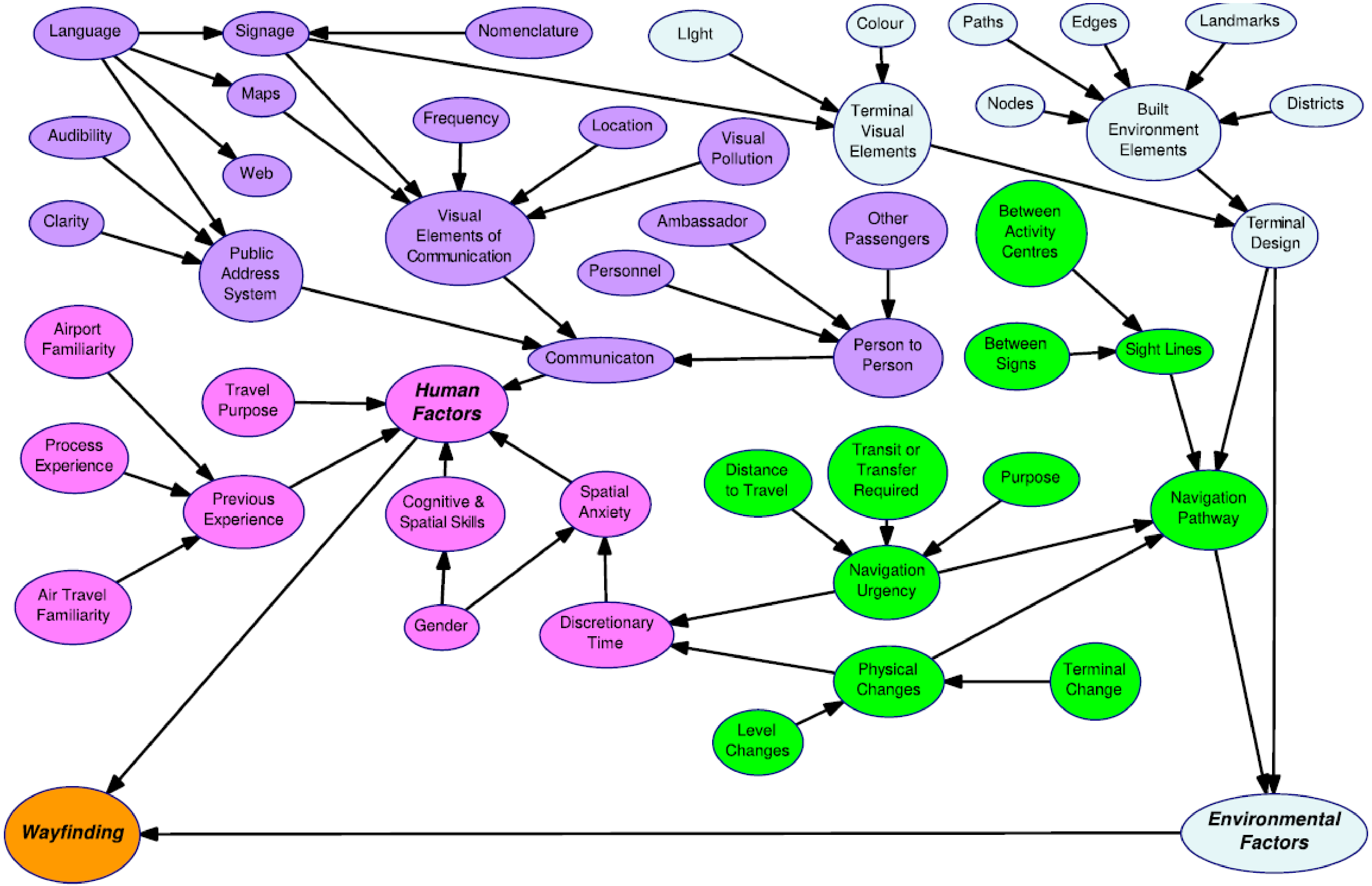
- Resilience is dynamic in space and time
  - Windows emerge from interactions of life histories, local conditions and growth patterns

- Ecological windows can enable up to four-fold reduction in recovery time, 35% reduction in extinction risk

- Consistent windows for greater robustness
  - Tend towards Autumn and Winter

- We can manage resilience much more effectively with planned scheduling of dredging

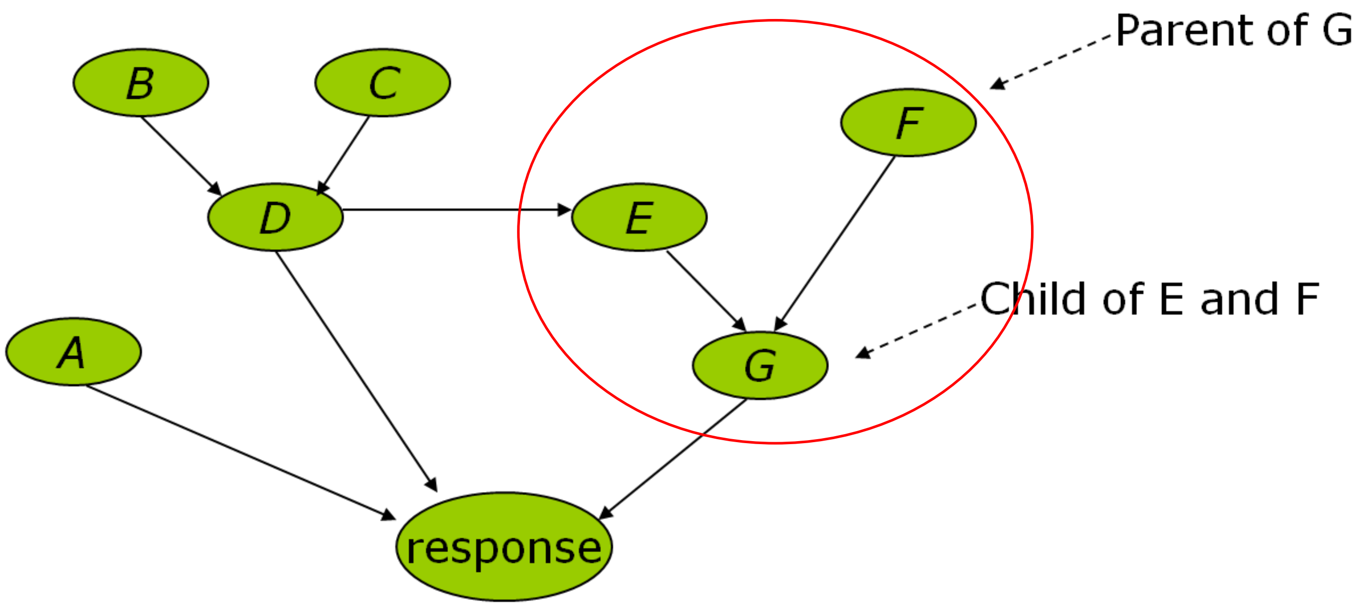
Case Study 4: Wayfinding

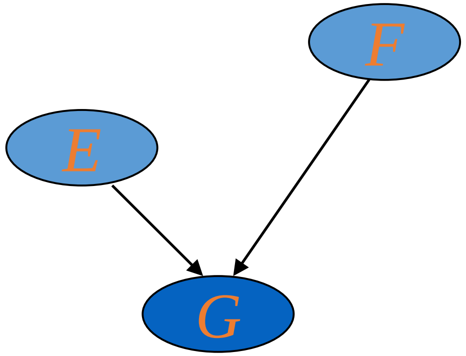


Probabilistically quantify the BN using ‘evidence’:

- data
- literature
- model outputs
- expert judgement

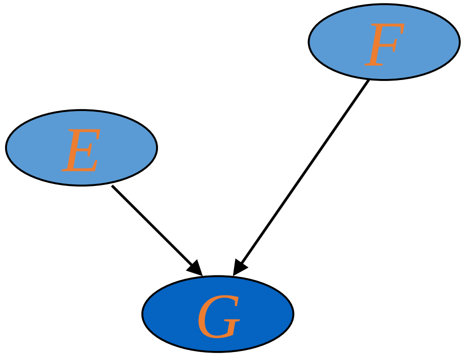
etc

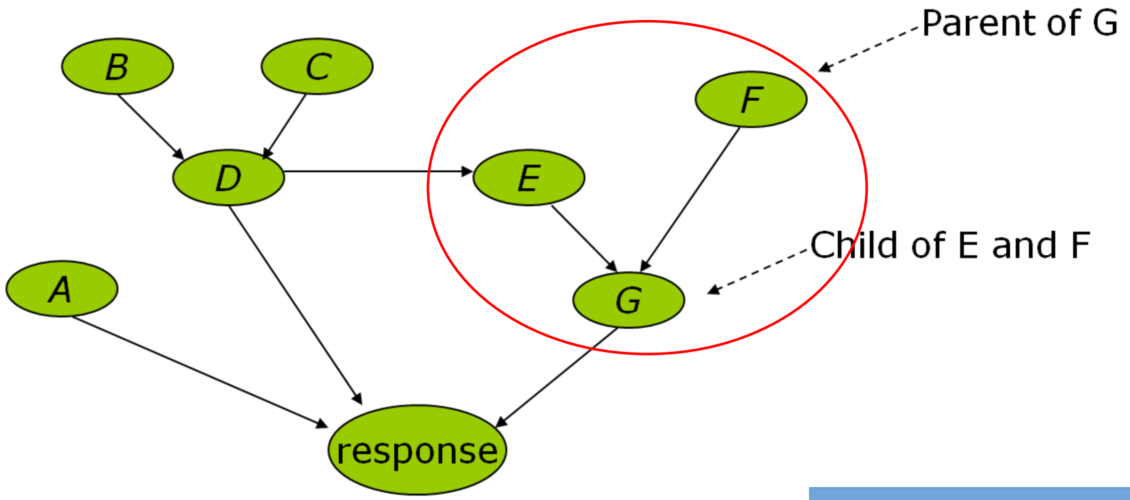




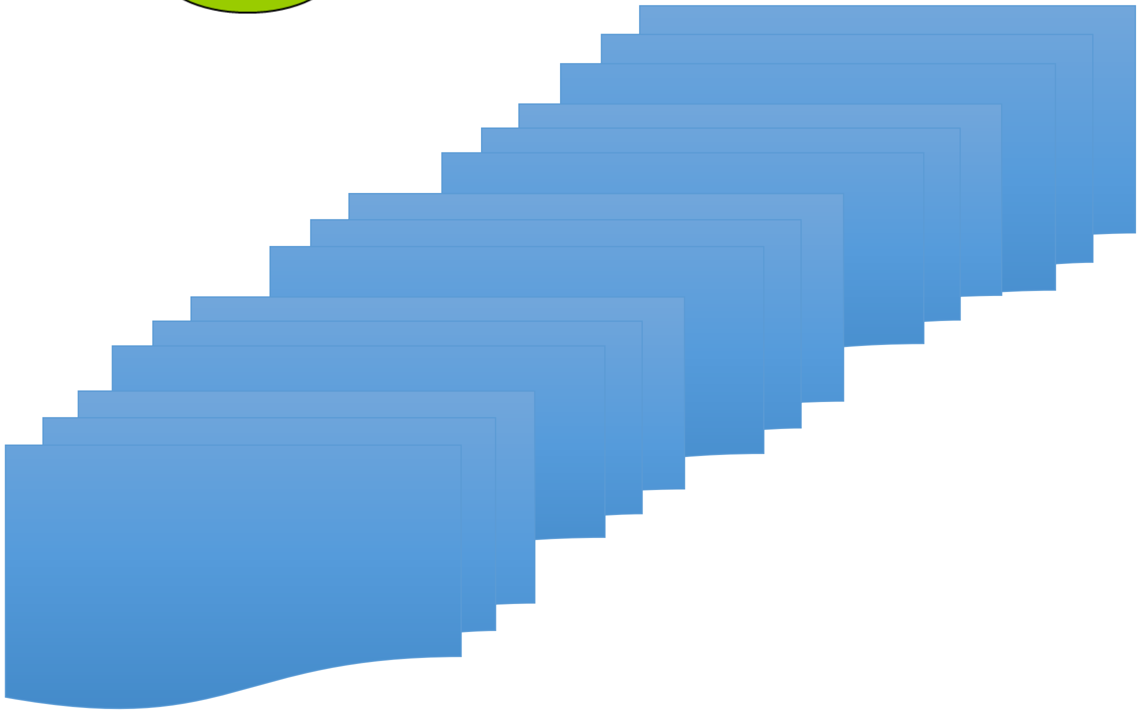
		G	
E	F	*normal*	*high*
yes	*low*	0.4	0.6
	medium	0.2	0.8
	high	0.1	0.9
no	*low*	0.5	0.5
	medium	0.6	0.4
	high	0.4	0.6

But what about *multiple* experts?





		G	
E	F	*normal*	*high*
yes	*low*	0.4	0.6
	medium	0.2	0.8
	high	0.1	0.9
no	*low*	0.5	0.5
	medium	0.6	0.4
	high	0.4	0.6



Linear Pooling

Information about node *X* provided by *n* experts $E_1, \ldots, E_n$

Joint model:

$$\mathrm{P}(\mathrm{P}(\mathrm{X}|E_1) \ldots \mathrm{P}(\mathrm{X}|E_n))$$

Pooled approximation:

$$P(X) = \sum_{i=1}^{n} \lambda_i P_i(X) \; ; \; \sum_{i=1}^{n} \lambda_i = 1$$

If all experts are equally weighted:

$$\lambda_i = 1/n$$

$$P(X) = \sum_{i=1}^{n} P_i(X)/n$$

Linear Pooling - Options

- Prior linear pooling:
  - probabilities pooled within each node (apply   at this stage)
  - resultant CPTs are then propagated through network to find marginal probabilities for nodes of interest

- Posterior linear pooling
  - quantify and compute BN for each expert separately
  - pool the marginal probability distributions for the final nodes in the *n* BNs (apply   at this stage)

Drawbacks of Linear Pooling

- Point estimate for the consensus; variation in expert opinions is lost.

- Does not follow from a coherent probability model (de Finetti, 1964: can only be considered an estimator if each observation is indept & Gaussian)

- The different methods can result in different outcomes for the nodes of interest.

- The conditional independence structure of the BN is not reflected in the way in which the expert opinions are combined particularly for prior linear pooling.

Measurement Error Approach - Univariate

- Consider a single node of interest.

- Model the marginal probability for expert $E_i, i = 1, \dots, n$ as

$$p_i \sim \mathrm{Beta}(a_i, b_i)$$

- Allow for variation between experts:

$$\mathrm{logit}\left(\frac{a_i}{a_i + b_i}\right) = \mathrm{logit}\left(\frac{a_i}{b_i}\right) + \mu + \epsilon_i$$

$$\mu \sim N(0, \tau_u^{-1}), \ \ \epsilon_i \sim N(0, \tau_\epsilon^{-1})$$

- Since E(logit…) = 0, this implies $a_i = b_i$, so an alternative is to model:

$$a_i + b_i \sim \mathrm{Gamma}(\alpha_0, \beta_0)$$

Measurement Error Approach - Multivariate

- Form consensus for multiple nodes $j = 1, \ldots, m$ in the BN.
- Then

$$p_{ij} \sim \mathrm{Beta}(a_{ij}, b_{ij})$$

- MVN random effect for each expert + extra between-expert variation:

$$\mathrm{logit}\left(\frac{a_{ij}}{a_{ij} + b_{ij}}\right) = \mu_j + \epsilon_{ij}$$

$$\mu_j \sim N\left(0, \tau_\mu^{-1}\right), \;\; \boldsymbol{\epsilon}_i \sim N\left(0_m, Q^{-1}\right), \qquad i = 1, \ldots, n; \;\; \boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im})$$

$$a_{ij} + b_{ij} \sim \mathrm{Gamma}(\alpha_\tau, \beta_\tau)$$

Multivariate Measurement Error Approach - Comments

- Structure of random effect term:

$$\boldsymbol{\epsilon} = \mathbf{Rs}$$

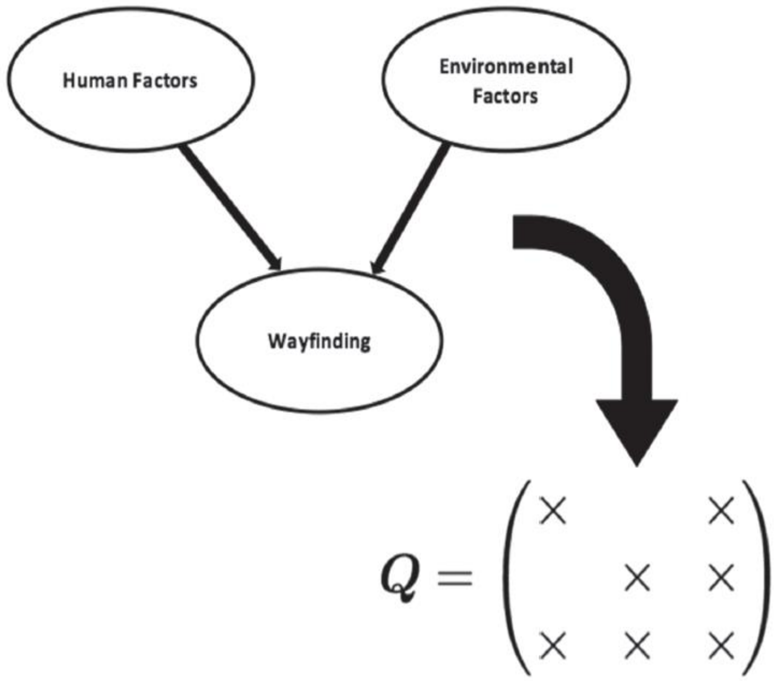
R: Cholesky decomposition of precision matrix **Q**

s: vector of i.i.d. standard normals, i.e. $\boldsymbol{s} = \mathrm{N}(0, I)$

- Hence by definition, $\boldsymbol{\epsilon}$ has a precision matrix s.t. $\boldsymbol{Q} = (\mathbf{RR}^{\mathrm{T}})^{-1}$.
- Implication: if **R** has the correct sparsity required, then **Q** will also have the correct sparsity structure.
- Hence $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_m, \mathbf{RR}^{\mathrm{T}})$
- Similarly, by finding **R**, we can give **Q** the right structure to reflect the conditional independence of a BN.

Improvement on usual approach (linear pooling)

Example



- 99 experts: P(good), P(good), P(effective) for H,E,W
- Model allows for combination of all opinions.
- Coherence is maintained under reordering of the independent expert opinions.
- $\mathbf{Q}$ allows the model to ‘borrow strength’ from other parts of the model: information can travel up and down the levels of the hierarchy.
- $\mathbf{Q}$ also ensures that the conditional independence structure of the BN is reflected when combining opinions: $\mathbf{Q}_{ij} \neq 0$ iff node *i* depends on node *j* in the BN.

Need to construct **R**, Cholesky decomposition of $\mathbf{Q}$: use expert priors on $\mathbf{Q}$. Write:

$$\operatorname{logit}(\mu_H) = \mu_H + \beta_1\epsilon_W + \epsilon_H$$
$$\operatorname{logit}(\mu_E) = \mu_E + \beta_2\epsilon_W + \epsilon_E$$
$$\operatorname{logit}(\mu_W) = \mu_W + \beta_3\epsilon_H + \beta_4\epsilon_E + \epsilon_W$$

μ_H, μ_E, μ_W : mean opinions for nodes H, E, W.

β terms indicate how much of the random effects comes from the other nodes.

Hence

$$\operatorname{logit}(\boldsymbol{\mu}_X) = \boldsymbol{\mu}_X + \mathbf{Rs}$$

$$\mathbf{R} = \begin{pmatrix} \tau_{\mathrm{H}}^{-1/2} & 0 & \beta_1\tau_{\mathrm{W}}^{-1/2} \\ 0 & \tau_{\mathrm{W}}^{-1/2} & \beta_2\tau_{\mathrm{W}}^{-1/2} \\ 0 & 0 & \tau_{\mathrm{W}}^{-1/2} \end{pmatrix}, \qquad \mathbf{Q} = \begin{pmatrix} \tau_{\mathrm{H}} & 0 & \beta_1\tau_{\mathrm{H}} \\ 0 & \tau_{\mathrm{E}} & \beta_2\tau_{\mathrm{E}} \\ -\beta_1\tau_{\mathrm{H}} & -\beta_2\tau_{\mathrm{E}} & \beta_1^2\tau_H\tau_{\mathrm{W}}^2 + \beta_1^2\tau_H\tau_{\mathrm{W}}^2 + \tau_{\mathrm{W}} \end{pmatrix}$$

$$\tau_X \sim \mathrm{Gamma}(1, 5\times 10^{-5}), \beta_X \sim \mathrm{N}(0, 5\times 10^{-5})$$

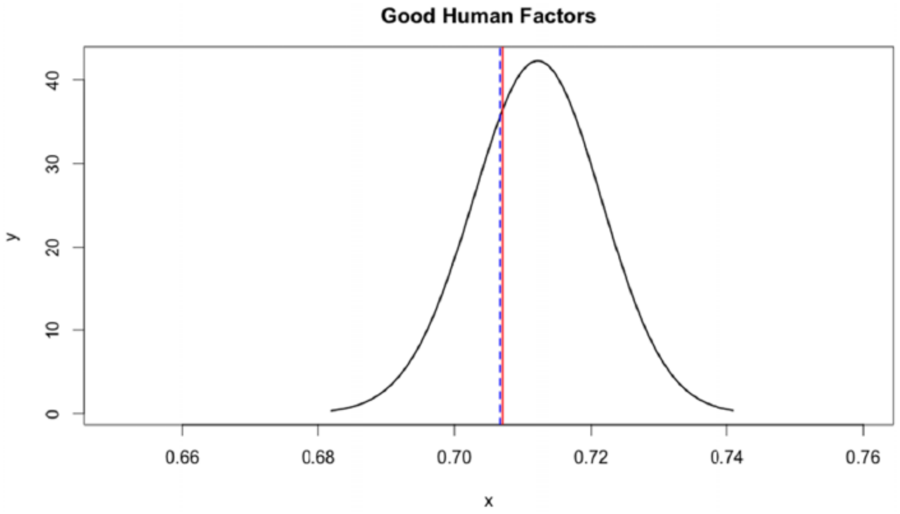
Priors:

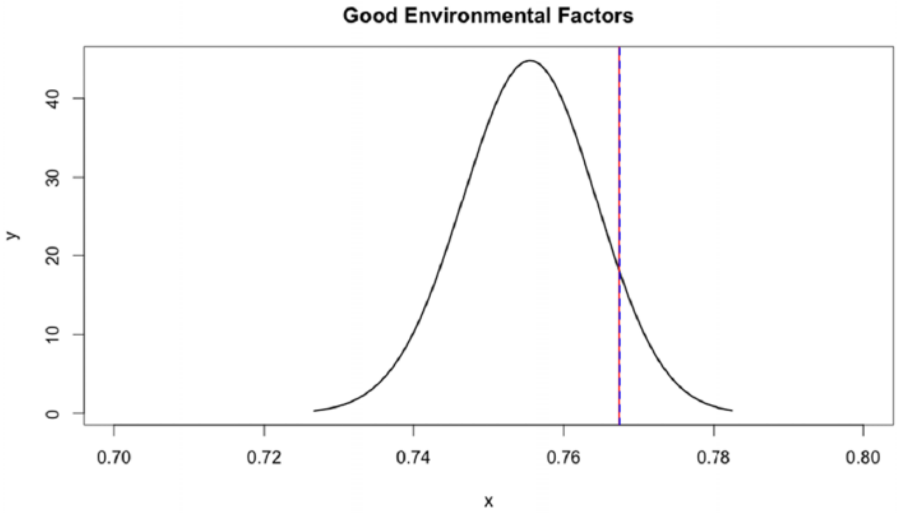
- Gamma(1,0.1) for $a + b$
  [2.5% & 97.5% interval for anticipated values for *p* implied by this prior is (0.53,1)]
- Proper but relatively uninformative priors for μ so ~ uniform distribution on p_i: $\tau_\mu^{-1} = 10^4$
- $(a_\tau, b_\tau) = (1{,}5 \times 10^5)$: small contribution of measurement error to overall value of p_i

Analysis:

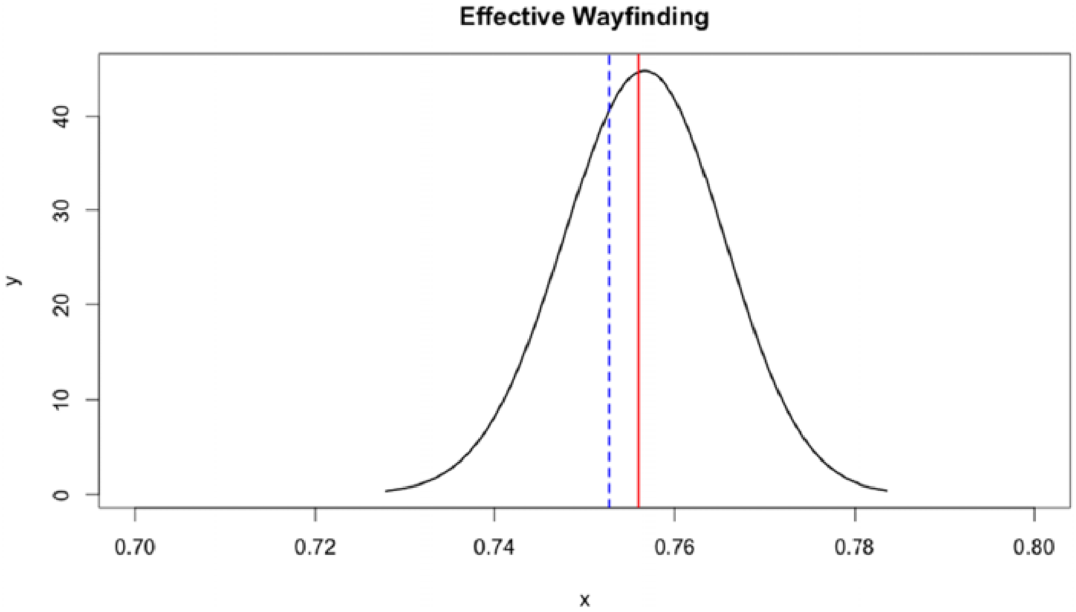
- In R INLA
- Utilises deterministic Laplace approximations by fitting Gaussian conditional posteriors via an optimisation step for latent Gaussian models.
- Faster, more accurate alternative to simulation-based MCMC schemes in many cases.
- Applicable here because of the sparsity of $\boldsymbol{Q}$

Results





Bayesian MME:
Posterior linear pooling:
Prior linear pooling:



Summary

- New measurement error approach for combining opinions in Bayesian networks
- Results indicate improved performance and increased inferential capability compared with current approaches such as linear pooling
- Can extend model to include bias and additional covariates.

e.g. to investigate effect of experienced (E) and inexperienced (I) travellers, modify μ_j (overall mean for node *j*) as

$$\mu_I \sim \mathrm{N}\left(\mu_j - \delta_I, \sigma_I^2\right); \quad \mu_E \sim \mathrm{N}\left(\mu_j + \eta_I, \sigma_E^2\right)$$

References

Caley MJ, O'Leary RA, Fisher R, Low-Choy S, Johnson S, Mengersen K (2013) What is an expert? A systems perspective on expertise. *Ecology and Evolution.*

Farr C, Mengersen K, Ruggeri F, Simpson P, Yarlagadda P (2020) Combining opinions for use in Bayesian Networks: a measurement error approach. *International Statistical Review.*

Santos-Fernandez E., *et al.* (2020) Correcting misclassification errors in crowdsourced ecological data: a Bayesian perspective. *Journal of the Royal Statistical Society Series C – under revision.*

Santos-Fernandez E., *et al.* (2020) Bayesian item response models for citizen science ecological data. *Methods in Ecology and Evolution – under revision.*