Predicting albumen gland length of grain crop pest snails

Kathy Ruggiero

Department of Statistics
The University of Auckland

28 November 2023

Email: k.ruggiero@auckland.ac.nz

Outline

The problem with snails

- Damage crops by feeding on plant heads and stalks
- Snail-fouled crops clog and damage harvest machinery
- Juveniles pass through machinery, contaminating grain
- Grain downgraded or even rejected at receival

- Minimising economic losses requires effective snail population control

Controlling snail populations

Requires a combination of methods:

- Cultural
  - Field hygiene, e.g. weed control and removal of refuges
- Biological
  - Predators, e.g. beetles, lizards, and birds (ducks, chickens or guinea fowl)
- Chemical
  - Baits

Baiting efficacy

- Limited field life
  - Rainfall affects physical integrity and dilutes concentration of active ingredients
  - Temperature-related degradation
- Requires re-application every 2–4 weeks
- Timing is critical
  - Rule of thumb: Bait in the autumn when snails are actively feeding and prior to egg laying
  - Idea: Preventing egg laying to mitigate risks of harvest contamination by juveniles (due to their small size)

Baiting efficacy

Can efficacy be improved by optimising timing of bait application?

What triggers reproduction activity?

Snail reproduction activity

- Hermaphrodites but "prefer" to mate with another
- Mature in 1-2 years
- Aestivate through summer, generally dormant
- Rain triggers snail activity, temporary feeding
- Autumn rains prompt feeding, mating, and egg-laying
- Lay up to 6 batches of 80 eggs each, hatching after 2 weeks

Snail reproduction activity

- Albumen gland swells when reproductively active

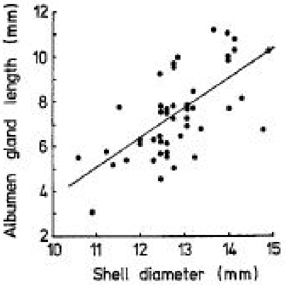


Fig. 7. Length of albumen gland as a function of shell diameter for *Cernuella virgata* collected in the pasture at Mt Benson during March 1985. Equation of the regression line is: $y=1\cdot32x-9\cdot37$, $r_{50}=0\cdot681$, $P<0\cdot05$.

Image source: Baker, G. H. (1988). The life-history, population-dynamics and polymorphism of *Cernuella virgata* (Mollusca, Helicidae). *Australian Journal of Zoology*, **36**, 497–512.

Study: Biology and ecology of pest snails

- GRDC[1]-funded project conducted by SARDI[2]
- Aimed at understanding pest snails in agricultural regions of southern and western Australia
- Four snail species and two slug species
- Several sites across southern and western Australia
- The rest of this talk will focus on a single snail species from a single site
- The methods are applicable to all four species across all sites

[1] Grain Research and Development Corporation

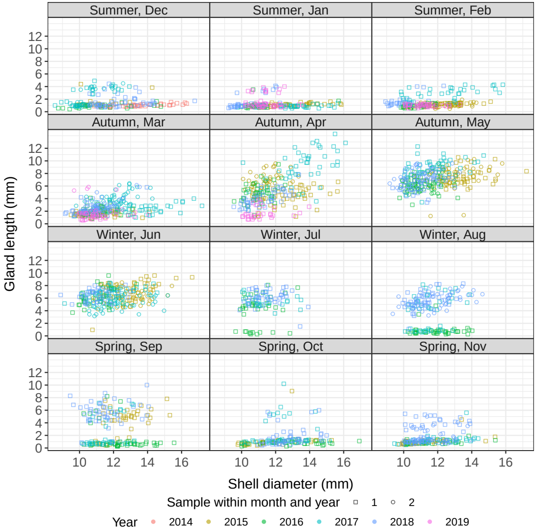
[2] South Australian Research and Development Institute

The data

- Quota sampling
- Approximately monthly samples gathered for size and reproductive trait analysis
- Longest and most comprehensive dataset:
  - Dec 2014 to Apr 2019, covering 64 sampling occasions
  - Sample sizes ranged from 17 to 45 snails per sample
  - Total 2498 animals during the study period
- Daily Austrailian Bureau of Meterology data
- Micro-climate data logger (30-minute intervals)

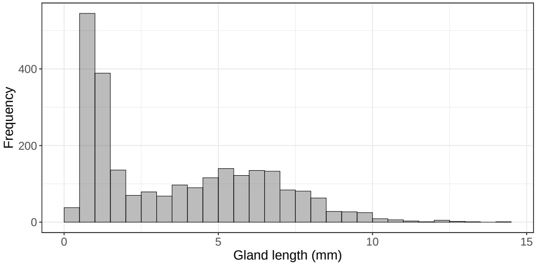
SARDI snail data: Single species, single site

December 2014 to April 2019



SARDI snail data: Single species, single site

Distribution of albumen gland length



- Gland lengths primarily cluster around two modes: around 1.0 mm and 5.8 mm.
- Bimodalilty plausibly explained by two reproductive states:
  - Higher mode (swollen glands) indicates reproductive activity (State A).
  - Lower mode indicates reproductive inactivity (State I).

Snail Data Analysis: Gland Length Modeling

- Baker (1988) modelled gland length as a function of shell diameter
- SARDI data, exhibiting bimodal gland length distribution, indicates we need a model accommodating two reproductive states (A and I)
- Unsupervised clustering required due to the lack of state labels

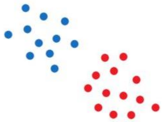
Objectives

- Separate snails into reproductively active and inactive states (today)
- Identify environmental (climate) variables which trigger reproductive activity

Unsupervised clustering of reproductive state

k-Means Clustering

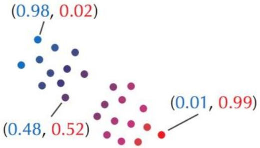
- Each snail assigned to one cluster (hard clustering)



- Tends to produce spherical and equally sized clusters
- Sensitive to outliers and initial centroids

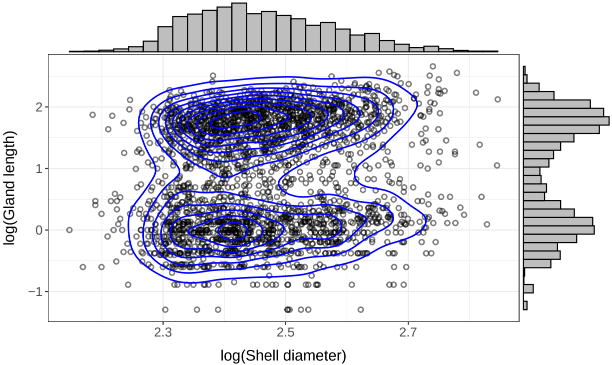
Gaussian mixture models

- Assigns probabilities to cluster membership (soft clustering)



- Accommodates varying cluster shapes and sizes

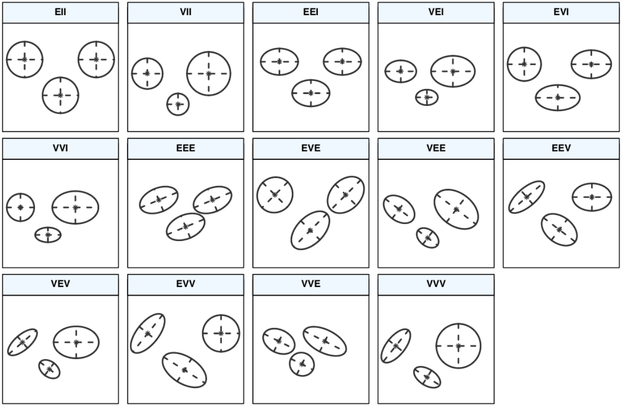
Log-transformed data characteristics



- Mixture of bivariate normals
- Ellipsoidal covariance matrix; variable orientation along coordinate axes
- Moderate (State A) and weak (State I) correlation between variables

Multi-dimensional Gaussian Mixture Model

Parametrizations of covariance matrices[3] (volume, shape, and orientation)



[3] Implemented in the Mclust (version 5) R package

Model selection

Multi-dimensional Gaussian Mixture Model

- How many components should be included in the mixture?
- Which covariance matrix should be adopted?

Bayesian Information Criterion (BIC)

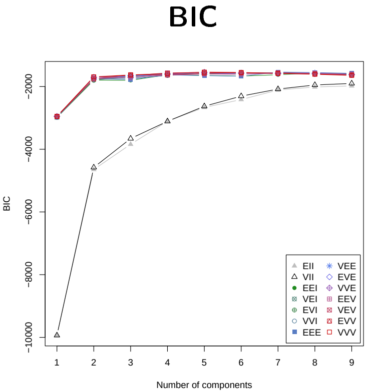
- Penalises model complexity using the number of parameters and sample size
- Tends to favour simpler models and may overlook certain complex structures

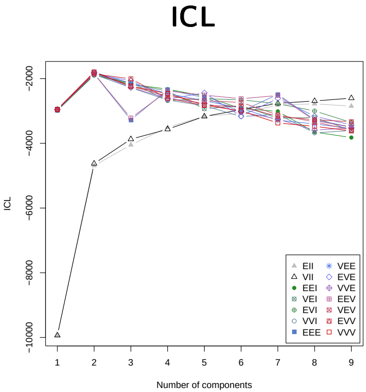
Integrated Complete-data Likelihood (ICL) criterion

- Penalises BIC by incorporating an entropy term which quantifies the overlap of observations between clusters
- Tends to favour solutions with clearly separated clusters

Information criteria for model selection

Multi-dimensional Gaussian Mixture Model





Best BIC values:

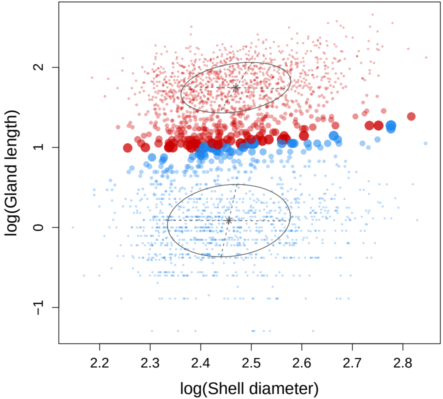
	VVE,5	VEE,5	VVV,6
BIC	-1545.15	-1548.56	-1549.53
BIC diff	0.00	-3.41	-4.39

Best BIC values:

	VEV,2	VVV,2	VEE,2
ICL	-1794.9	-1798.75	-1802.59
ICL diff	0.0	-3.85	-7.69

2-Cluster uncertainty plot

Multi-dimensional Gaussian Mixture Model



- Uncertainty is given by size of point
- 94.7% of cases have probability >0.9 of belonging to the assigned cluster

```
null device
          1
```

Cluster reliability

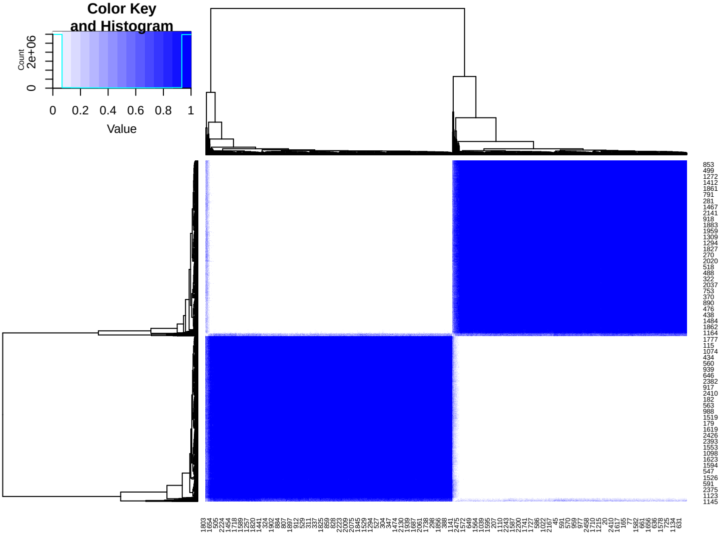
- Cannot assess cluster *accuracy* (correctness of cluster assignment) because true reproductive state is unknown
- Can use *bagging*[4] to assess cluster consistency, i.e. sensitivity to small changes in the input data

[4] BagClust2, Dudoit and Fridlyand (2003) *Bioinformatics*, **19**.

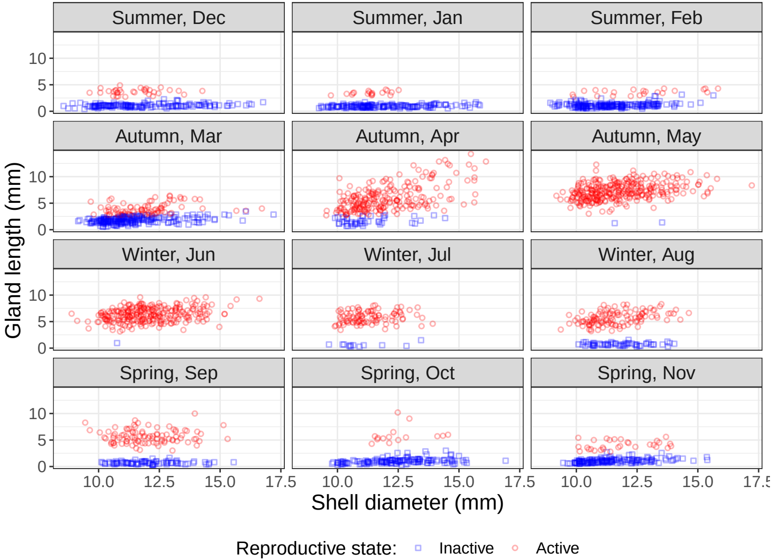
Bag clustering 2 algorithm

1. Initialise $A = [a_{ij}]$ and $M = [m_{ij}]$ matrices to zeros.
   - A records concurrences of observations in the same cluster across bootstrap samples
   - M records total occurrences of observations in the same bootstrap sample
2. Form the bth bootstrap sample $L_b = (x_{b1}, \ldots, x_{bn})$.
3. Apply clustering procedure P to L_b and obtain cluster labels $P(x_{bi}; L_b)$.
4. Update matrices A and M for each pair of observations based on cluster concurrence.
   - $a_{ij} \leftarrow a_{ij} + I[x_i \in L_b,\ x_j \in L_b,\ P(x_i; L_b) = P(x_j; L_b)]$
   - $m_{ij} \leftarrow m_{ij} + I[x_i \in L_b,\ x_j \in L_b]$
5. Repeat Steps 2–4 B times and compute dissimilarity matrix $D = [d_{ij}]$, where $d_{ij} = 1 - \frac{a_{ij}}{m_{ij}}$
6. Cluster the n original observations based on the dissimilarity matrix D

Cluster reliability

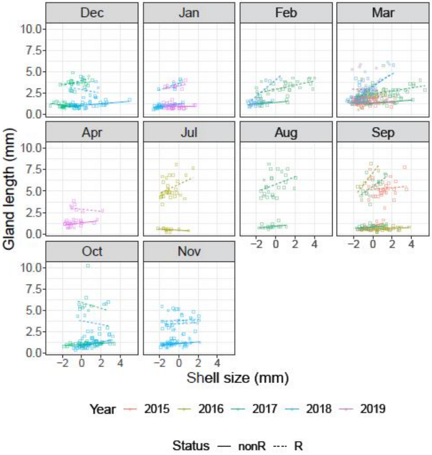


GMM clustering results



Next steps

- Predict gland length for "standard" sized snail by sample (month/year) and state



Next steps

- Correlate gland length with BOM and micro-climate data
  - Tree-based approach: binary (reproductive state) vs continuous (log(GL/SD)) response
- Single-step method, using a Bayesian approach?