Probabilistic Models for Understanding Ecological Data: Case studies in Seeds, Fish and Coral

Allan Tucker

Brunel University

London

Brunel UNIVERSITY LONDON

The Talk

- The Data Explosion and Ecology
- Case Studies:
  1. Data Driven Models for prediction: Seeds
  2. Integrating Knowledge and Data: Coral
  3. Dynamic Models and Latent Variables: Fish
- Conclusions

Brunel UNIVERSITY LONDON

Data historically...

- Preserve of handful of scientists:

Newton, 1600s

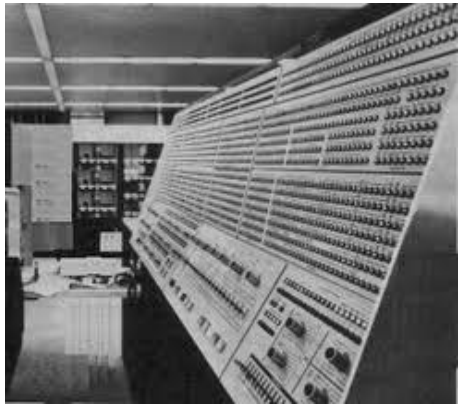
Darwin, 1800s

Galton, 1800s

Pearson, 1900s

Database Technology Timeline

- 1960s:
  - Data collection, database creation

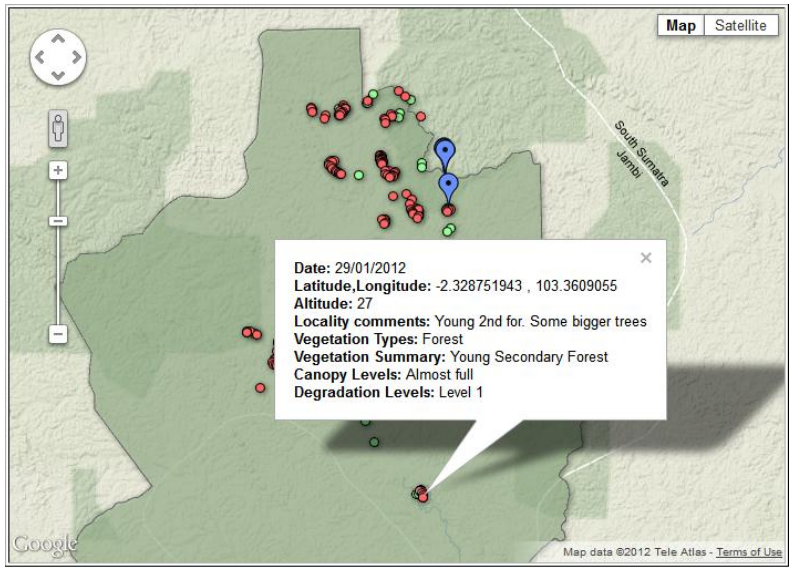


- 1970s:
  - Relational data model
  - Relational DBMS implementation
- 1980s:
  - Advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s—2000s:
  - Data Warehousing
  - Multimedia and Web databases
  - Distributed DW: The Cloud

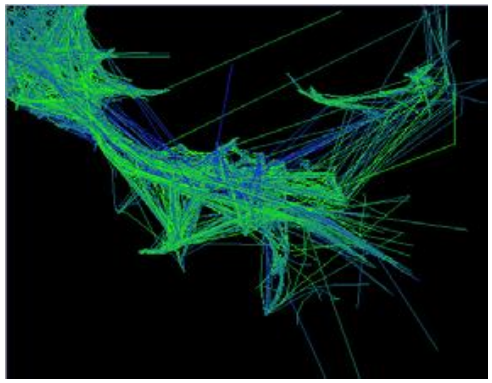
Brunel UNIVERSITY LONDON

Data Generation examples

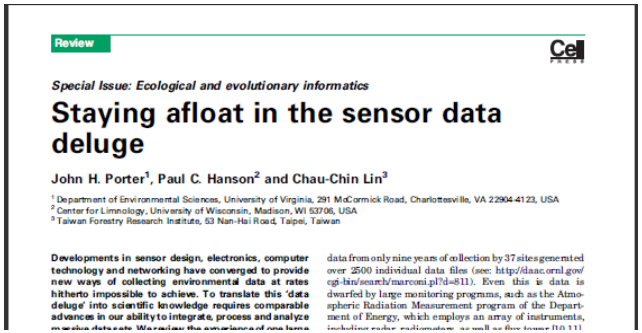
- Data collected from:
  - Online forms, Sensors, GIS, Mobile devices ...



Kew Gardens, Harapen Project



CASOS
Tech Report



Review

Cell

Special Issue: Ecological and evolutionary informatics

Staying afloat in the sensor data deluge

John H. Porter[1], Paul C. Hanson[2] and Chau-Chin Lin[3]

[1] Department of Environmental Sciences, University of Virginia, 291 McCormick Road, Charlottesville, VA 22904-4123, USA
[2] Center for Limnology, University of Wisconsin, Madison, WI 53706, USA
[3] Taiwan Forestry Research Institute, 53 Nan-Hai Road, Taipei, Taiwan

Developments in sensor design, electronics, computer technology and networking have converged to provide new ways of collecting environmental data at rates hitherto impossible to achieve. To translate this 'data deluge' into scientific knowledge requires comparable advances in our ability to integrate, process and analyze

data from only nine years of collection by 37 sites generated over 2500 individual data files (see: http://daac.ornl.gov/cgi-bin/search/marconi.pl?id=811). Even this is data is dwarfed by large monitoring programs, such as the Atmospheric Radiation Measurement program of the Department of Energy, which employs an array of instruments,

Brunel UNIVERSITY LONDON

Data Analysis

- Increasing ability to record & store
- So need to Analyse:
  - Data Mining,
  - Machine Learning,
  - Intelligent Data Analysis,
  - Knowledge Discovery in Databases
  - Bioinformatics
  - Ecoinformatics
  - Predictive Ecology

  ...
- Large overlap with statistics
  (and all the same caveats)

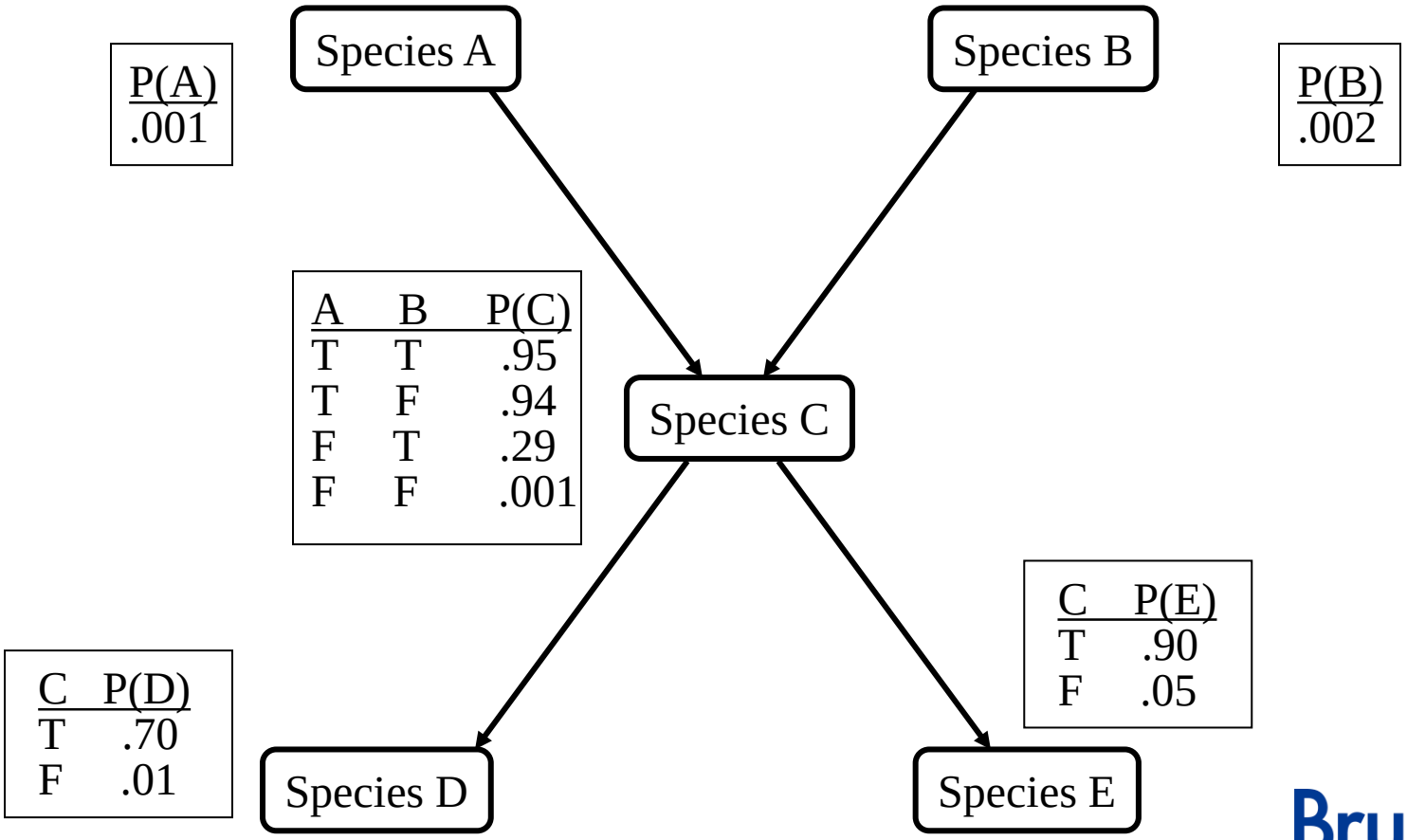
Brunel UNIVERSITY LONDON

Bayesian Networks for Data Mining

- Can be used to combine existing knowledge with data using informative priors
- Essentially use independence assumptions to model the *joint distribution* of a domain
- Independence represented by a graph: easily interpreted
- Inference algorithms to ask ‘What if?’ questions

Brunel UNIVERSITY LONDON

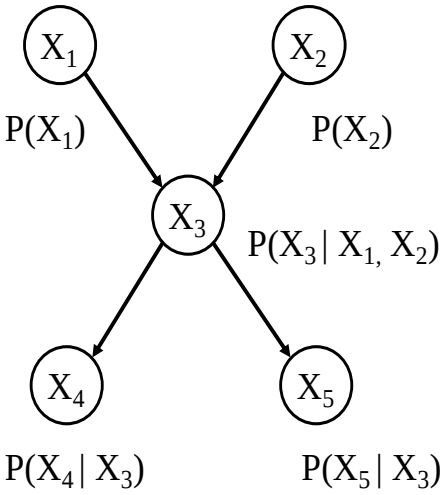
Example Bayesian Network

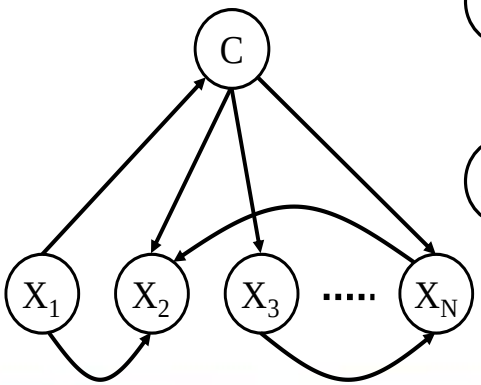


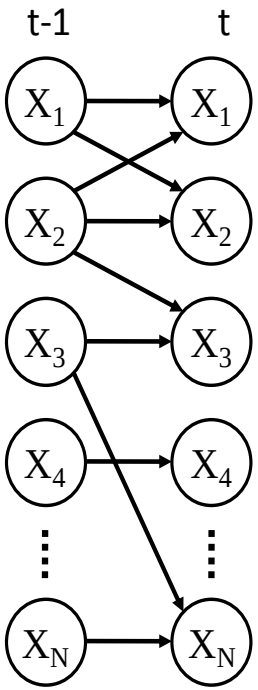
Brunel UNIVERSITY LONDON

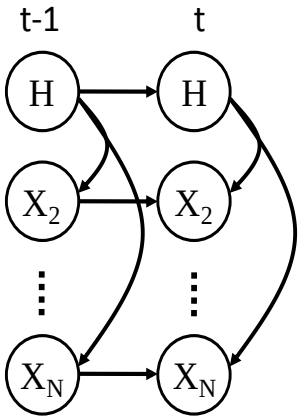
Bayesian Networks for Classification & Feature Selection & Forecasting

- Nodes that can represents class labels or variables at "points in time"
- Also latent variables via EM
- Feature Selection









Brunel UNIVERSITY LONDON

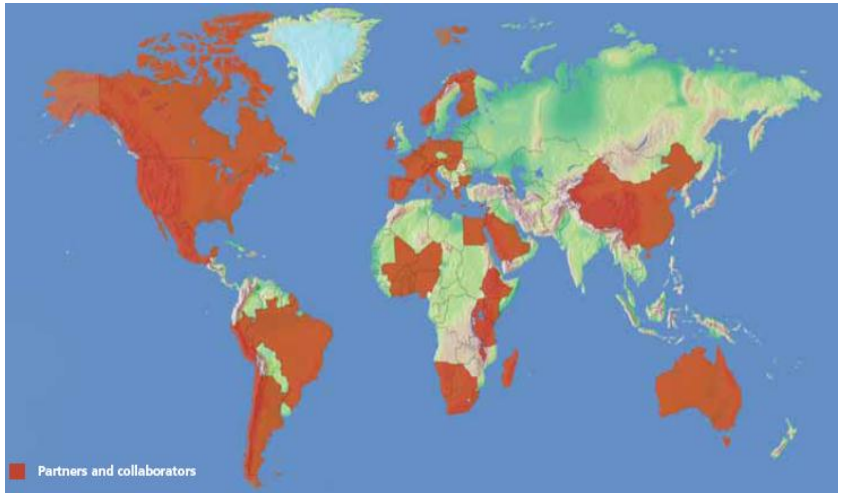
Predictive Ecology 1
Data Driven Models

- The Millennium SeedBank
- RBG, Kew banking seeds for 35 years
- MSB established for 12 years
- 152 partner institutions in 54 countries worldwide

Brunel UNIVERSITY LONDON

The Millennium SeedBank

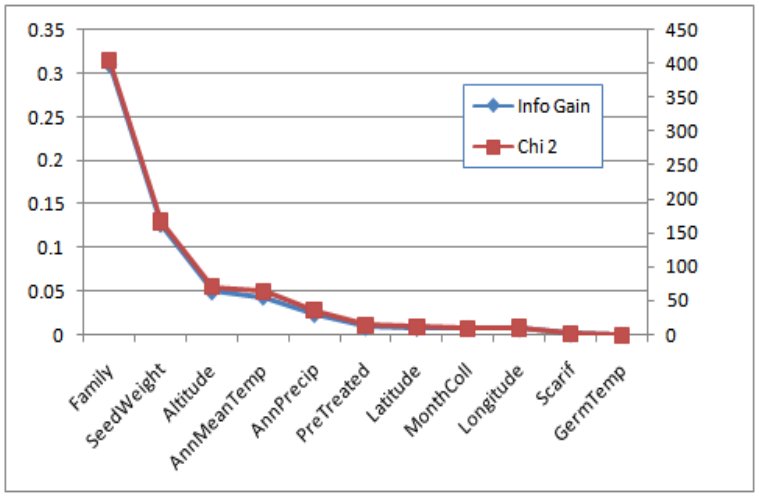
- Collected and stored >47,000 collections representing >24,000 species
- The Seedbank Database (SBD) - UK and worldwide
- GIS data (Detailed Climate)

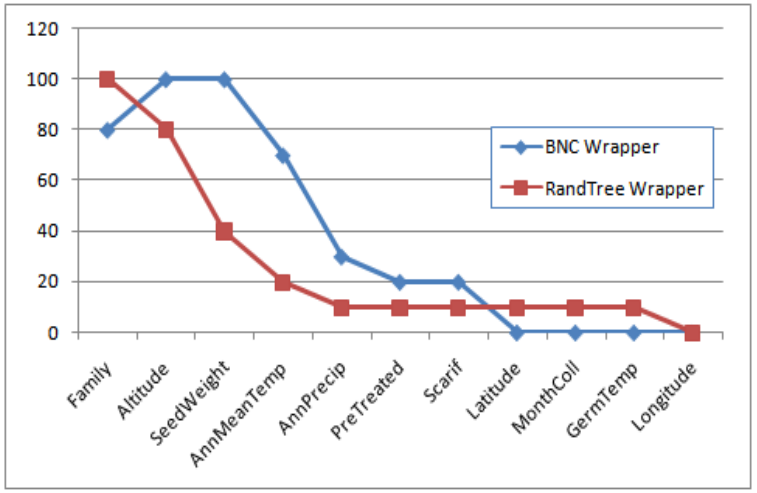


- Use this data to build predictive models for successful germination

Brunel UNIVERSITY LONDON

Results: Seedbank Data

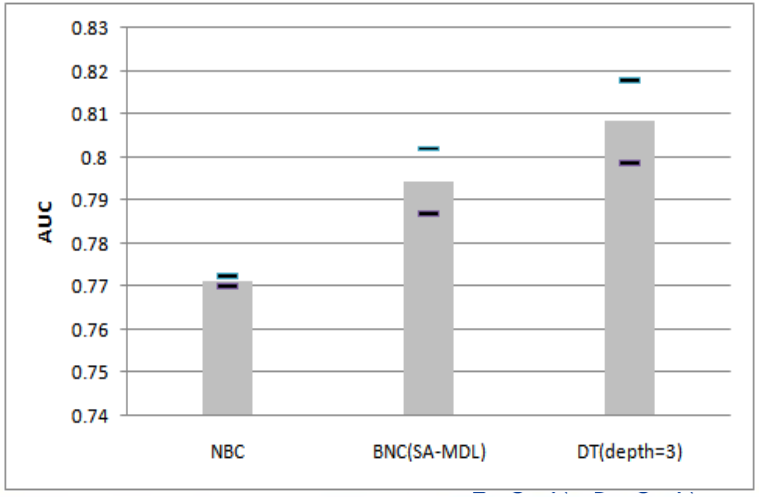




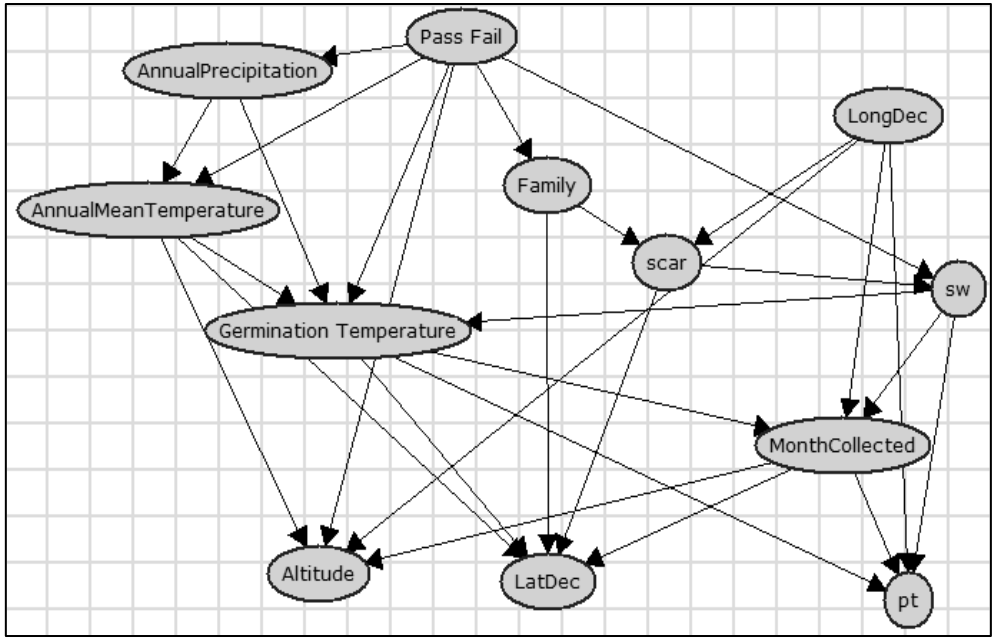
- Lots of similarity to filter method implying independence of features but some interaction (e.g. *scarification* and *latitude*)

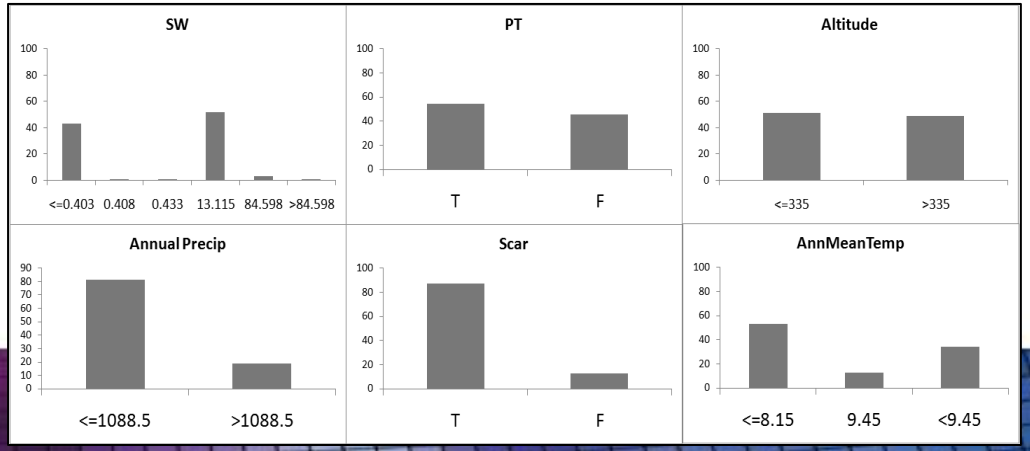
- Generally high predictive scores

- But explanation important

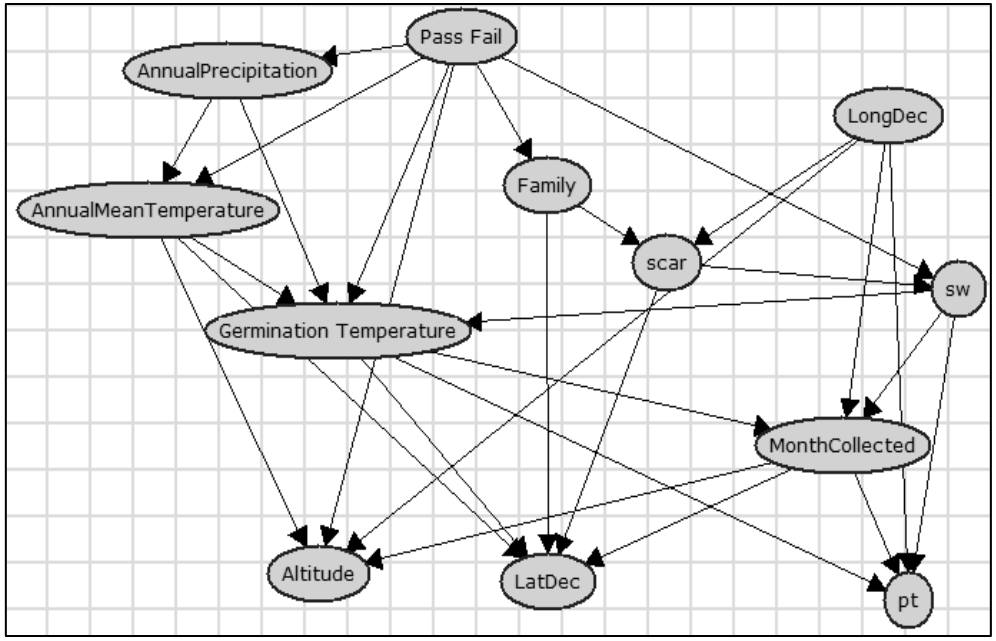


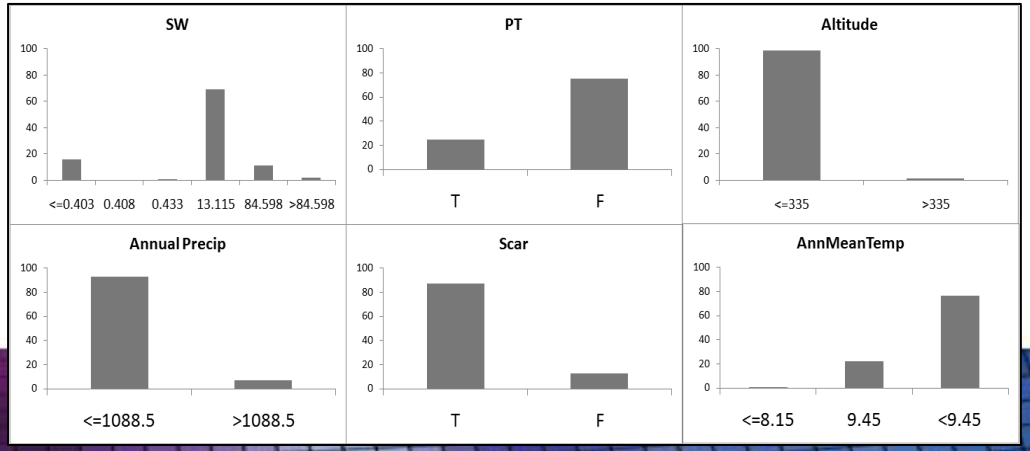
Results: Seedbank Data





Results: Seedbank Data





Results: Seedbank Data

- Markov Blanket includes all variables: all offer some improvement in prediction of germination success

- Exploit ‘what if’ queries by entering observations into model and applying inference:
  - Recognisable pattern emerging from Kew analysis that agrees with network:
  - Where pre-treatment is necessary, and it is applied, there is still relatively high probability of failure

Brunel UNIVERSITY LONDON

Summary

- Use of data mining / machine learning to
  - Utilise large scale data to *predict* and *explain* ecological phenomena
  - Explore data using ‘what if’ models
- Expanding this work to build models for predicting plant traits of ecosystems in different regions
  - Text mining of monographs
  - Large flora datasets
  - GIS, MSB, ...
- Predict what species likely to grow with others and what likely traits will be

Brunel UNIVERSITY LONDON

Predictive Ecology 2
Data and Knowledge Integration

- Modelling Coral Carbonate Budgets

Brunel UNIVERSITY LONDON

Coral Reefs

- Among the most complex and productive tropical marine ecosystems
- Made from calcium carbonate (**$CaCO_3$**) secreted by corals and other calcifying organisms
- Structure holds great variety of organisms and serves as breeding, spawning, nursery and foraging habitat

Brunel UNIVERSITY LONDON

Carbonate budget assessment

- Increasing climate variability and anthropogenic pressures driving reefs to deterioration and destruction
- Carbonate budget assessment
  - Management tool used to determine spatial and temporal variations of reef framework accretion ($CaCO_3$ deposition) and erosion ($CaCO_3$ removal)
  - BUT low reliability of this methodology for long term management actions due to limited temporal and spatial scales at which method can be used
- Can we exploit a combination of data sources in one framework to better manage reefs?

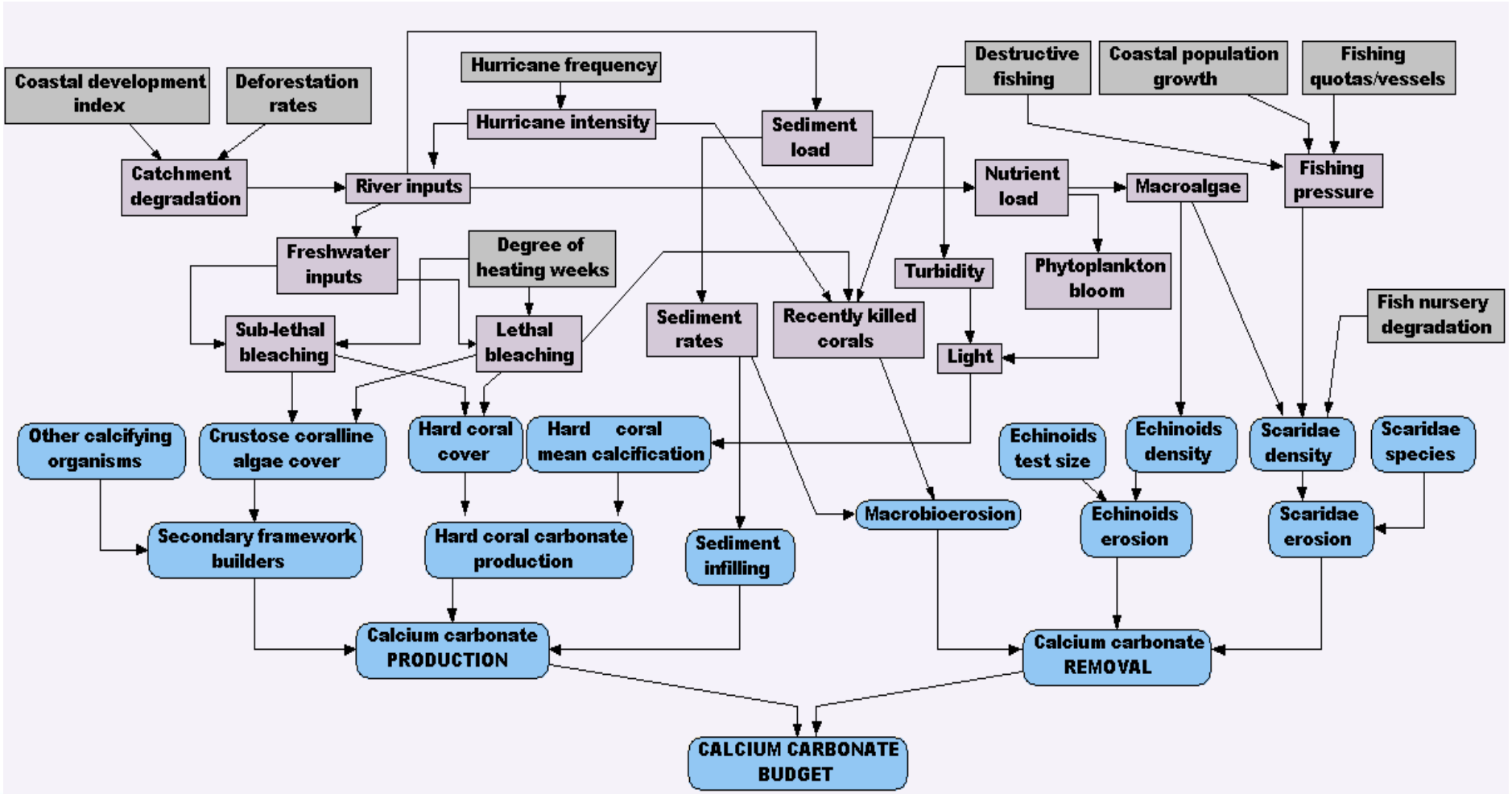
Brunel UNIVERSITY LONDON

Building the Model

- Initial structure constructed based on systematic review of *published literature* on carbonate budget (n= 11)
- Integrate with climatic and human disturbance nodes based on *international guidelines* for reef management and *expert knowledge* (parameters and structure)
- Indonesia data collected at three sites
  - Located across a gradient of sedimentation and turbidity
  - Continuous data discretised to two or three bins (severe/high, moderate/medium, low).
- *Data* used to update priors

Brunel UNIVERSITY LONDON

Bayesian Network for Carbonate Budget



Bayesian Network for Carbonate Budget

- Three subsets of nodes can be distinguished:
  - Nodes of the climatic and anthropogenic disturbances affecting coral reef framework accretive and erosive processes (grey-rectangular),
  - Nodes representing the direct effects of these disturbances on the framework processes (violet-rectangular)
  - Nodes closely related to $CaCO_3$ accretive and erosive processes (blue-oval)

Brunel UNIVERSITY LONDON

Results: Carbonate budget assessment

- Distinctive differences in the quantity of carbonate removed (CAR) at three sites
- Model was effective in detecting the quantitative differences in bioerosion (CAR) across environmental gradients BUT explanation was not clearcut
- Initial results proved ability of the model to inform which variables needed further investigation to assist future data collection (filtering out independent)

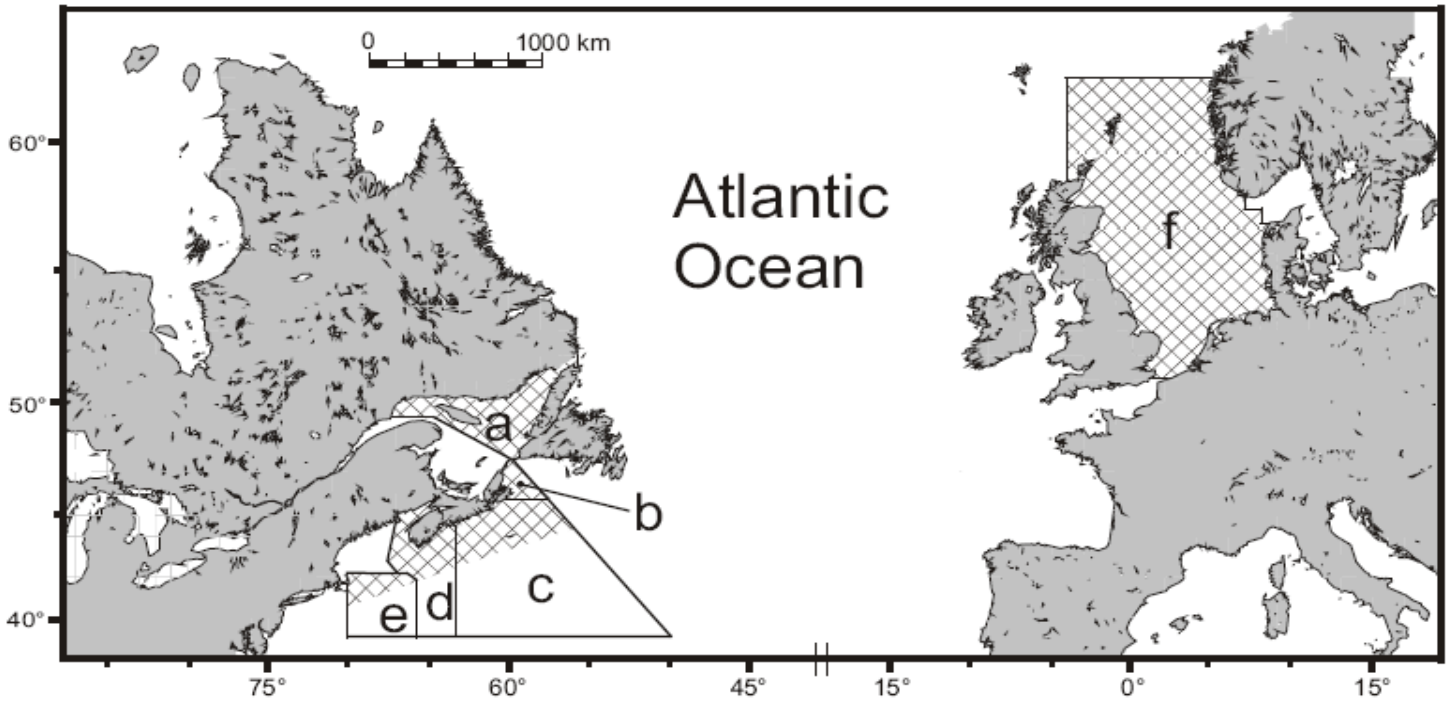
Brunel UNIVERSITY LONDON

Summary

- Can provide coral reef managers with tool that quantitatively assess rate of change of reef structure and inform which variables have driven changes the most
- Can provides managers with information on which reef components the data collection should be focused on in order to better understand reef ecosystem status
- Plan to extend this as a freely available tool to address questions for conservation by providing potential scenarios of reef status
- Plan to use data from different coral reef regions to provide reliable analysis of prediction (generalise between different regions – more on this later)

Brunel UNIVERSITY LONDON

Predictive Ecology 3
Dynamic Models with Latent Variables



Brunel
UNIVERSITY
LONDON

Fisheries Data

- George's Bank, East Scotian Shelf and North Sea
- Biomass data collected at different locations
- 100s of different species
- From 1960s until present day
- Massively complex foodwebs:
  - Predator / prey, cannibalism, competition …
- Foodwebs and catch data also available
- Lots of unmeasured variables

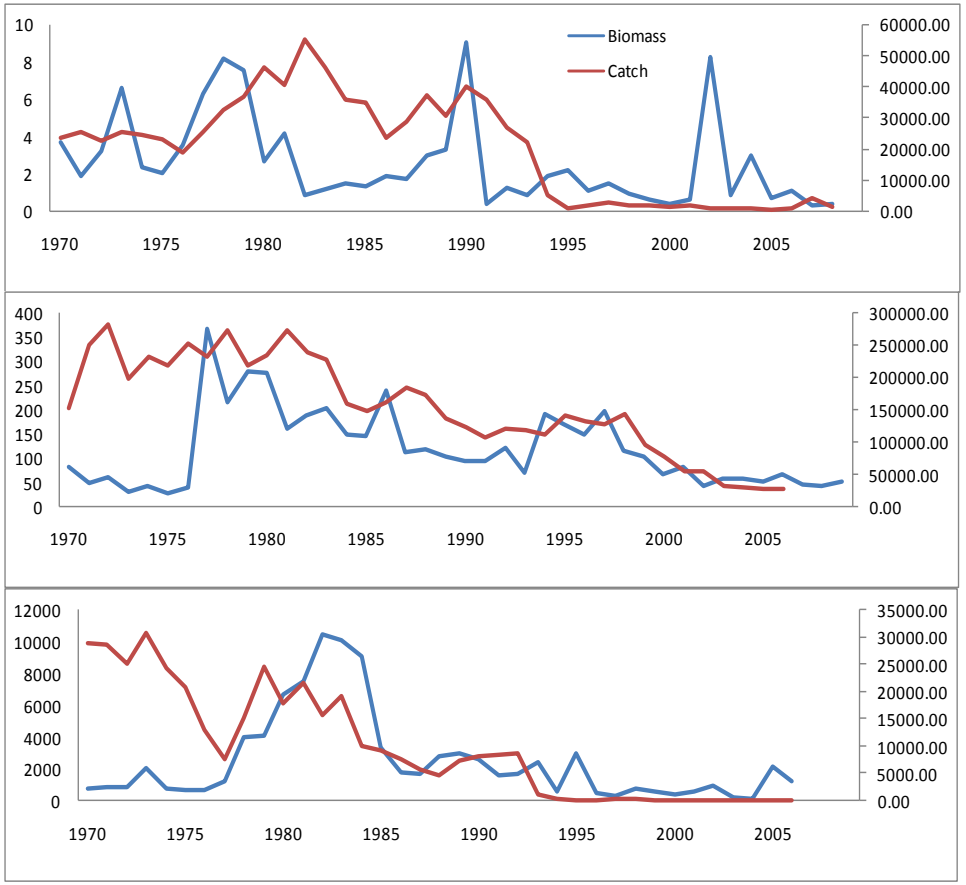
Brunel UNIVERSITY LONDON

Functional Collapse in G Bank, N Sea & ESS

George's Bank
Functional Collapse
in late '80s early '90s

North Sea
No Functional
Collapse

East Scotian Shelf
Functional Collapse
in late '80s early '90s

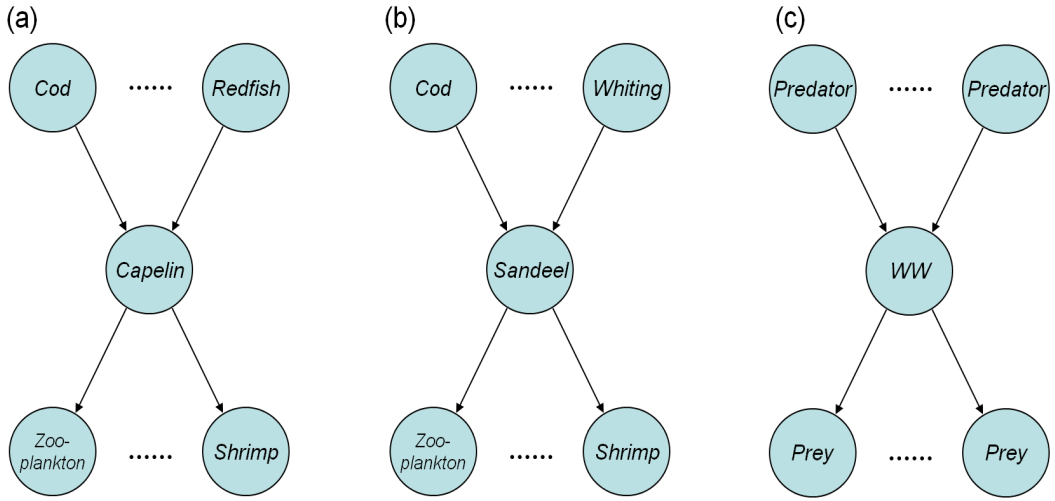


(Jaio, 2009)

Brunel UNIVERSITY LONDON

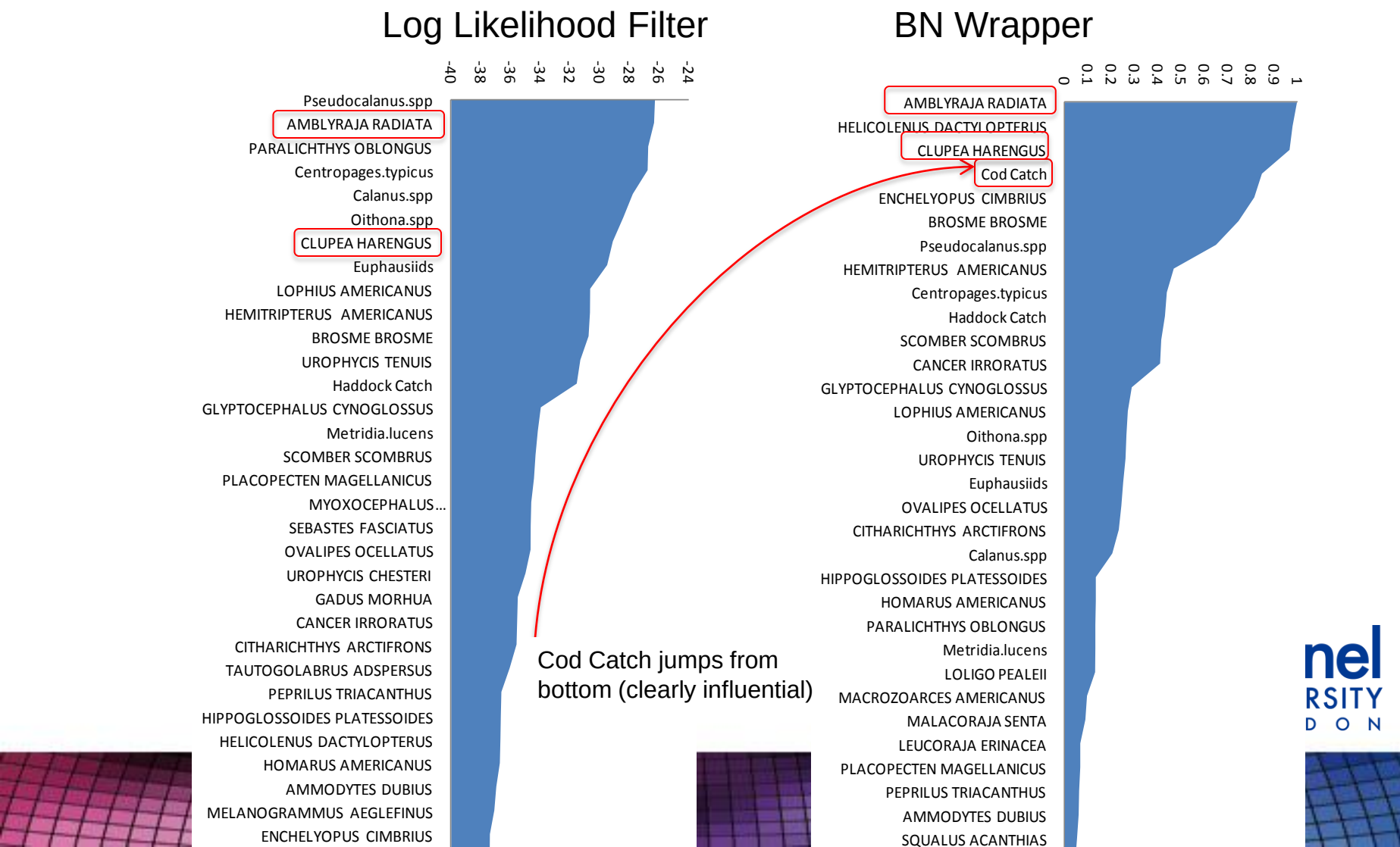
Questions

- Why do populations irrevocably collapse?
- What underlying ‘states’ dictate biomass?
- Can we generalise between regions?



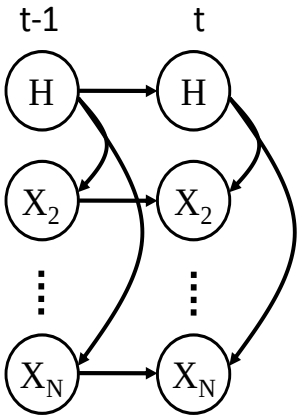
Brunel UNIVERSITY LONDON

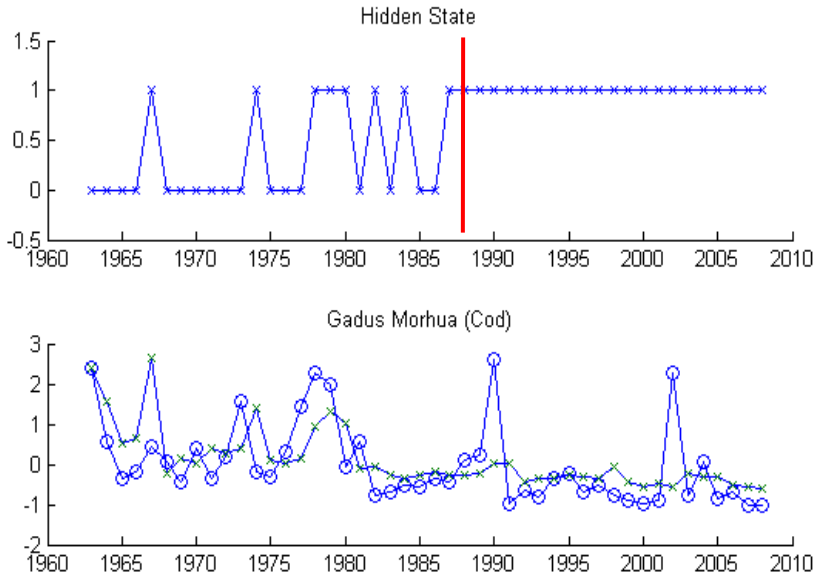
Results: Feature Selection to identify “cod collapse” in George’s Bank



Results: Fitting Dynamic Models & Identifying Functional Change

- Selecting species based on George's Bank foodweb, FS and cross correlation
- Learn DBNs with latent state variable



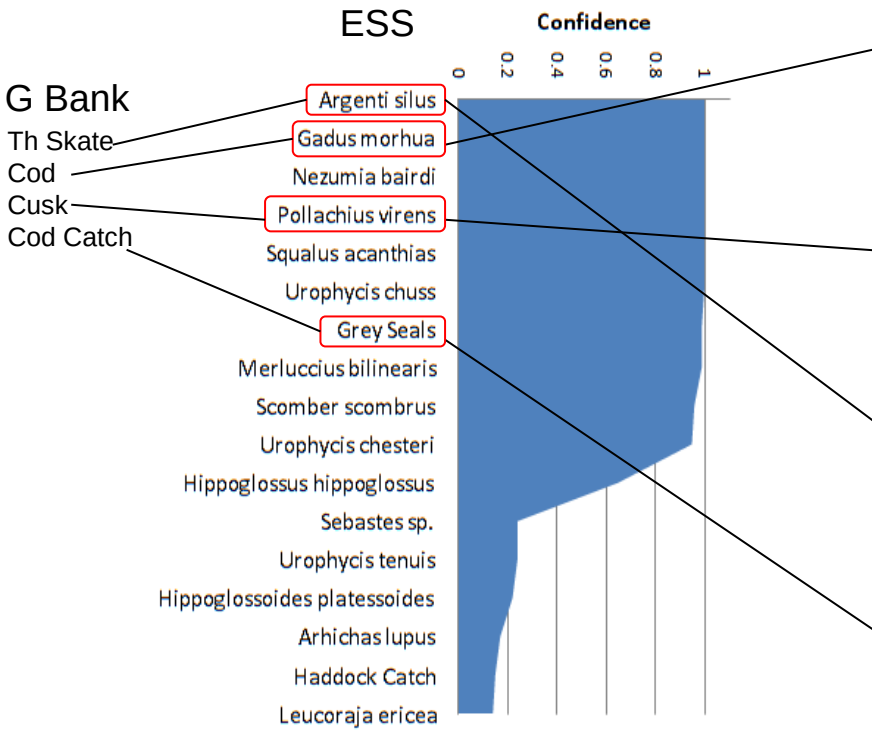


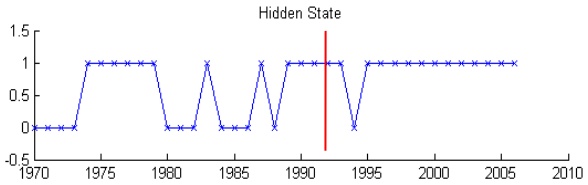
Brunel UNIVERSITY LONDON

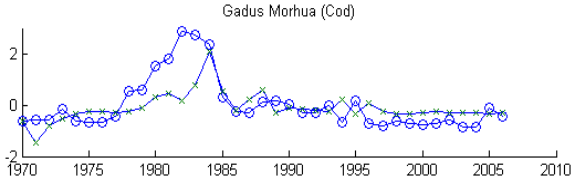
Results: Dynamic Functional Models

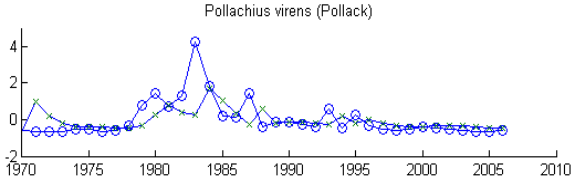
- Predicting ESS event &

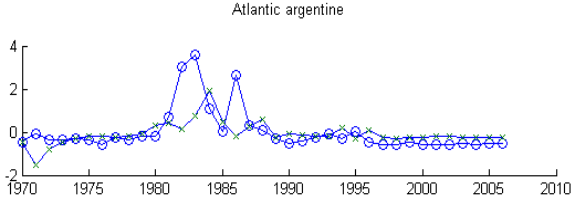
Cod biomass from G Bank

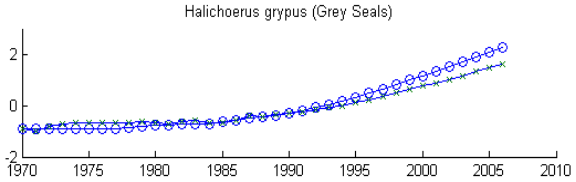










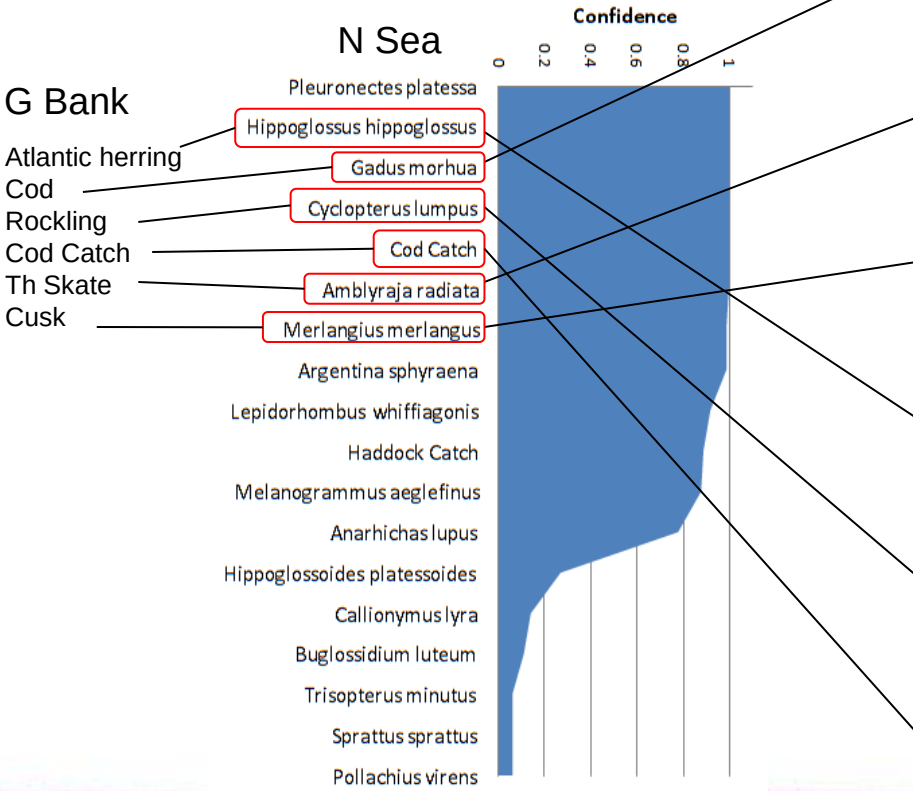


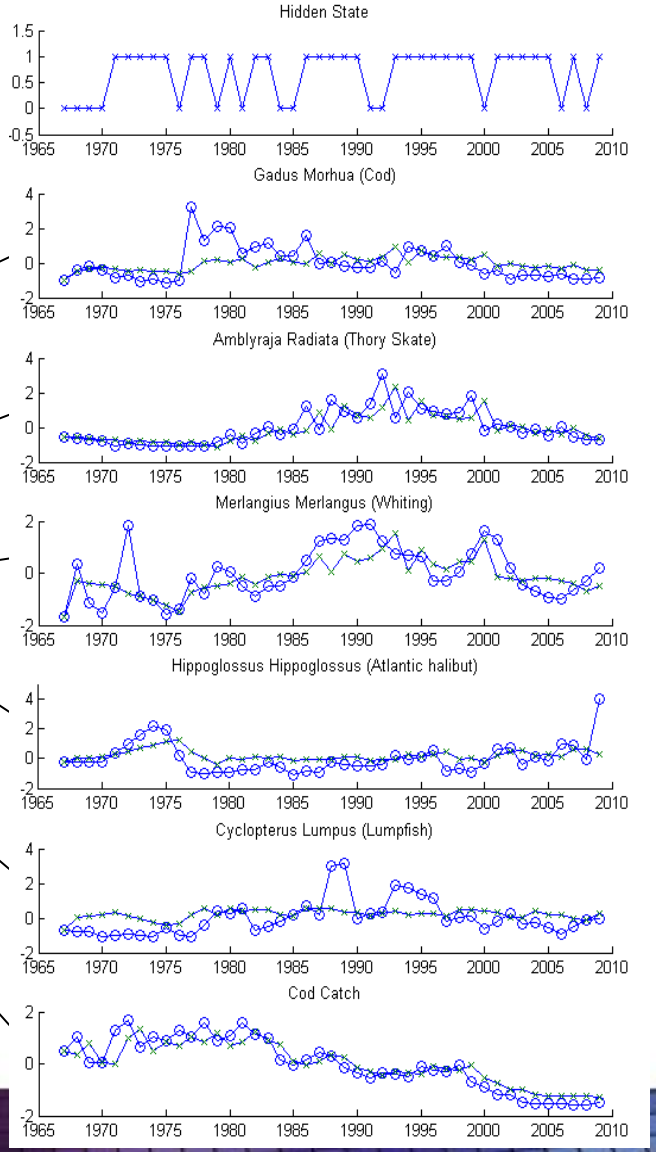
Brunel UNIVERSITY LONDON

Dynamic Functional Models

- Predicting N Sea event &

Cod biomass from G Bank





Brunel UNIVERSITY LONDON

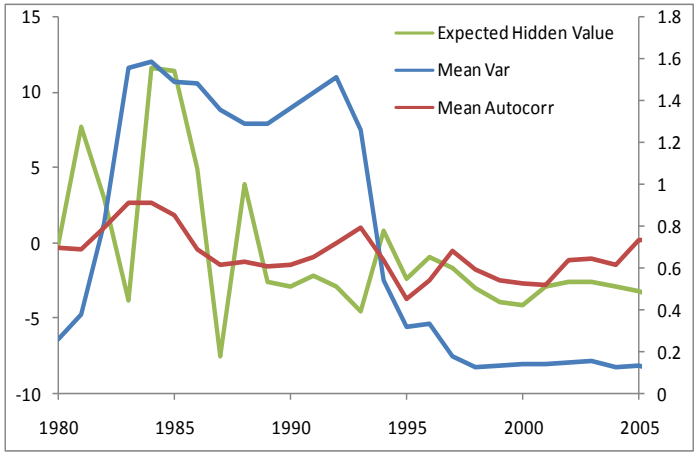
Summary

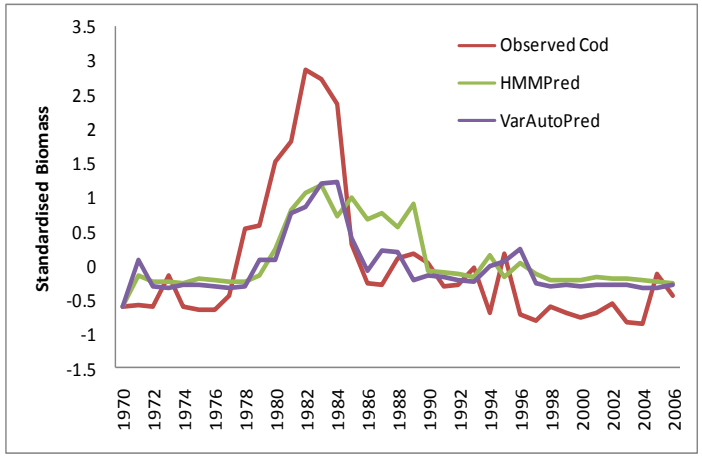
- Using Fisheries Data from several locations:
  - Identified functionally equivalent species in other locations
  - Used species in one location to build time-series models for prediction on species in other locations
  - Used latent variables to identify similar functional collapses (or not)

Brunel UNIVERSITY LONDON

Incorporating Variance and Autocorrelation metrics

- Prediction is improved when *regime shift metrics* are included (rather than relying on hidden states)
- A particular improvement in ESS: drop in large peak in 1982





Brunel UNIVERSITY LONDON

Conclusions

- Looked at 3 case studies in ecology
  - Data is noisy, complex, heterogeneous
- Bayesian network approaches to
  - Incorporate diverse data and expertise
  - Model latent variables and time
  - Perform, prediction, classification, forecasting & generalisation
  - Transparent: Can perform explanation
    - Structure and parameters are not black box
    - “What if” inference experiments

Brunel UNIVERSITY LONDON

Conclusions

- Prediction is

'*a property that sets the genuine sciences apart from those that arrogate to themselves the title without really earning it*',

Peter Medawar, nobel laureate, immunologist and philosopher of science

- *Predictive Ecology* is an important way to deal with modelling ecological phenomena:
  - Confidence in models
  - Deal with overfitting

- *Systems approach* also important

Brunel UNIVERSITY LONDON

Caveats

- Data Quality
  - Models only as good as the data that goes in
  - Exploitation of expert knowledge is key
  - Including the appropriate variables :
    - Human / Sociological factors
    - External factors to the system (latent variable analysis but expertise is better! E.g. regime shift metrics)
- Ecological events are often ‘novel situations’
  - Must be able to predict events outside of ‘normality’
  - If we have previous examples then must generalise to other regions
  - If not, must go beyond supervised learning (anomaly detection)
- Issues with Data Sharing and reproducibility

Brunel UNIVERSITY LONDON

Thanks to...

- **Chiara Franco & Liz Hepburn, Essex University, UK**
- **John Dickie & Don Kirkup, Royal Botanical Gardens, Kew, UK**
  - Kenwin Liu RBG, Kew
  - Robert Turner RBG Kew
- **Daniel Duplisea, Mont Joli Insitute, Canada**
  - Jerry Black DFO-BIO Halifax for assistance with the ESS survey data
  - Alida Bundy DFOBIO Halifax for the ESS food web,
  - ICES datras database for the North Sea IBTS data,
  - Bill Kramer NOAA–NMFS Woods Hole for the Georges Bank survey,
  - Jason Link NOASS–NMFS for the Georges Bank food web,
  - Jon Hare NOAA–NMFS for NE USA plankton data,
  - SAHFOS for North Sea plankton data
  - Mike Hammill for ESS grey seal data.

Brunel UNIVERSITY LONDON