An evaluation list as model selection aid

George A.K. van Voorn (Biometris, WUR)
P.W. Bogaart, H. Houweling, P. Janssen

PLANT RESEARCH INTERNATIONAL
WAGENINGEN UR

Overview

- Background
  - Problem
  - Akaike and others
- Tool development: evaluation list
  - Some important issues
- Current status
  - Testing
  - Prospects

Model complexity increase

- Past: Conceptual models for understanding
  - E.g. Fitz-Hugh Nagumo for spikes, Lorenz for weather, Verhulst-Pearl for populations, etc.

- Present: Numerical models and data bases for quantitative predictions

  → Increased model complexity

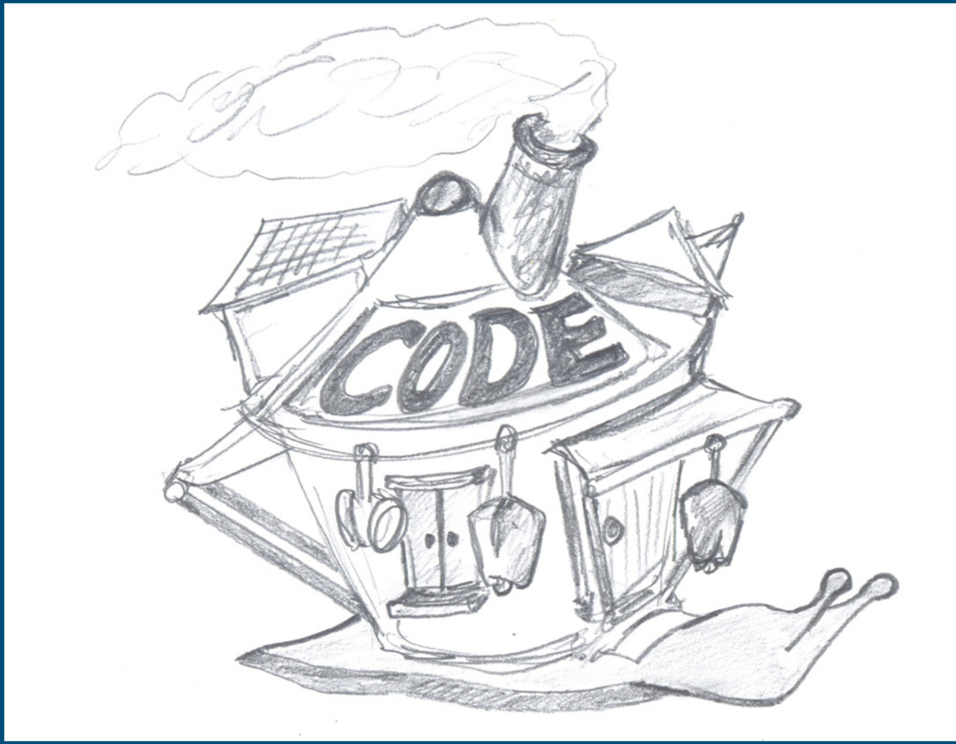
  → Increased role application

Problems?

- Some observations:
  - ‘Include all’ syndrome (Chwif et al., 2000)
    Fear of missing something
  - ‘Concorde’ effect (Van Nes & Scheffer, 2005)
    Investment > ➔ Conservation tendency >
  - ‘Swiss army knife’
    Multiple applications for the same model

Real dangers

- Models/data bases
  - Running time >>
  - Large data need
  - Difficult to comprehend
  - Large maintenance
  - Susceptible to management issues



Picture by GVV

Goal: equilibrium

- We want to find ‘balanced’ models/data bases:
  - Sufficiently complex for predictions within margins of uncertainty (governed by application(s))
  - Simple enough for understanding
  - Sufficient data supported for identification, etc.
- How?

Classical model selection

- Literature search:
- Selection of model based on data-fitting
  - Over-fitting (Myung, 2000)
- Complexity vs fit
  - AIC (Akaike, 1974), BIC, HQC, ...
- Also used for first principle models
  - But... Application-driven?

Demonstrative case

- Some small population growth models for budding yeast (*Saccharomyces cerevisiae*)
- Fit to biovolume measurements (Gause, 1932)
  - Global sensitivity analysis → initial guesses
  - Local minimum search → best fit (residual sum of sq.)
- Use results for AIC = 2 k + n ln(RSS)

Demonstrative case results

Model	k	AIC	HQC	BIC
$dx(t)/dt = r$	2	47.41	46.91	48.21
$dx(t)/dt = r\,x(t)\,(1 - x(t)/K)$	3	24.79	24.04	25.98
$dx(t)/dt = r\,x(t) + s$	3	21.23	20.48	22.42
$dx(t)/dt = r\,x(t) + s\,x(t)^3$	3	31.20	30.45	32.40
$dx(t)/dt = r\,x(t)\,(x(t)/z - 1)\,(1 - x(t)/K)$	4	25.65	24.65	27.24

Nonsense result

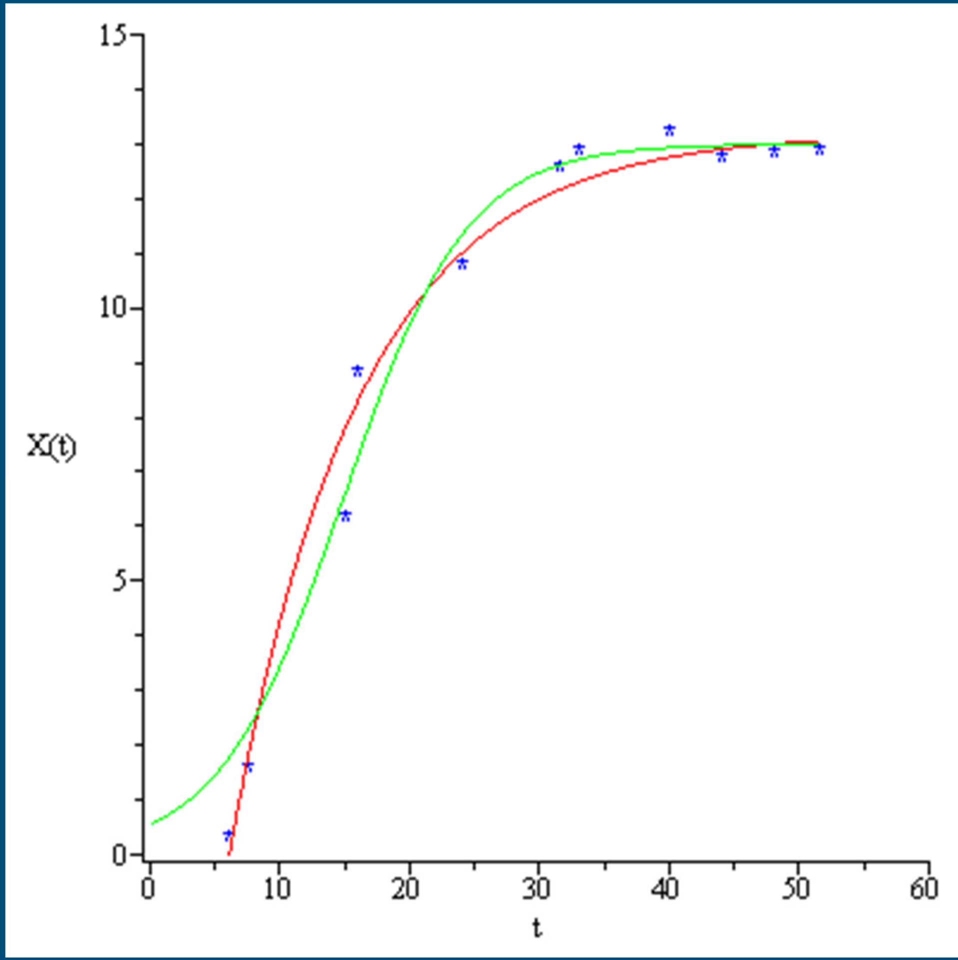
- 'Best' fit → Non-biological

model:

$dx(t)/dt = r\,x(t) + s$

Logistic growth:

$dx(t)/dt = r\,x(t)\,(1 - x(t)/K)$



Other approach

- AIC, BIC, etc.:
  - Warning: best ‘fit’ is not necessarily best model
  - Application-driven?
  - Guidelines for improvement
- What then? Approach:
  - Evaluation list model complexity (EMC)

Modelling question list

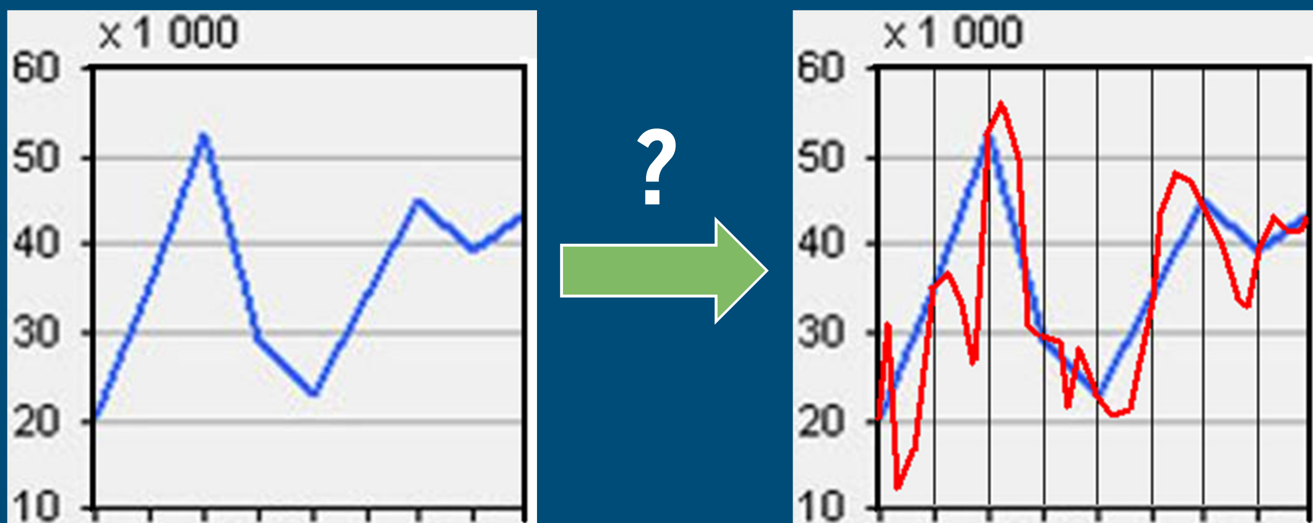
- EMC contains questions
  - Application-driven
- Follows modelling cycle
  - System analysis, Concept, Framework
  - Numerical implementation, Verification
  - Calibration, Validation, Uncertainty analysis
- Important issues:

(1) Complexity definition(s)

- Many different definitions (a.o. Holland, 2006)
  - Must include different model types/data bases
- Model: Parameters, variables, boundary conditions, non-linear interactions (Wagener et al, 2001)
- Data base: # cells, # possible cell states, # permutations of input

(2) Scale

- 3 elements of scale (Bierkens et al., 2000)
  - Extent (upper & lower boundaries)
  - Coverage, e.g. measurements every hour, but take 1 second ➔ coverage = 1/3600
  - Averaged over interval (Ljung, 1987)



(3) Requirements

- Model goal
  - Corresponds to (all) model application(s)?
- Each model application
  - Degree of overlap with other applications?
  - What output is required? vs what output is generated?
  - What data are required? vs what data are used?
  - What level of certainty?

(4) Identification costs

- Application determines modelling effort
  - ‘Cost of complexity’ (Hjalmarsson, 2009)
- Example: Water levels & dyke height
  - Application: Risk assessment flood
  - Need: High water level predictions very reliable
  - Unimportant when low(er)
  - Do not invest effort in getting the lower levels correct

Project status

- ECM version 0.1, evaluation
  - Expert review
  - Limited test cases
- Example: Budding yeast population (*Saccharomyces cerevisiae*)
  - Monitoring beer-making

Case beer-making

- Budding yeast model (Kurz et al., 2002)
  - Grows on wort. High sugar conc., low nutrients
  - Produces ethanol, CO_2, flavours

- Issues:
  - Limitation & inhibition effects (a.o. Crabtree-effect)
  - Common failures in temperature and oxygen flux
  - Fermenting takes 1-3 weeks

Case characteristics (1)

- Goal:
  - Metabolic model, include limitation/inhibition effects
  - 'Good' yeast for fast fermenting
  - Potential for process control

- Model:
  - r = A v, vectors: r substrate, v reaction rates
  - Stoichiometric, A 10-by-8 matrix
  - Non-linearity → Switches

Case characteristics (2)

- Implementation:
  - AQUASIM 2.0
- Data:
  - Clear literature sources for most parameter values
  - Glucose & biomass measurements at fixed T/O_2-supply
  - Coverage: 4 settings T, variations O_2, but not combined
  - Three case scenarios with temporal variations T/O_2

Evaluation case

- Main issues data & application:
  - 10 state variables, only 2 measured
  - No ethanol measurements (!?)
  - Match parameters literature and this model?
  - Oxygen decrease below extent of O_2 significant effects

- Verdict:
  - Increase data support for low O_2
  - Ethanol measurements
  - Few non-linear eq. probably better

Prospects

- Further testing
  - Superfluous questions
  - Unclear/complex questions
  - Impossible questions (relevant & clear but undoable)
  - Missing issues
- Future plans
  - Upgrade evaluation list
  - Apply list to model train

Thank you!

More information:

www.**biometris**.wur.nl/UK/Staff/George+van+Voorn

© Wageningen UR

PLANT RESEARCH INTERNATIONAL

WAGENINGEN UR