mathematical modeling and computational creativity

ljupčo todorovski
university of ljubljana, faculty of administration
ljubljana, slovenia

symposium on computational approaches to creativity in science (scacs)

talk outline

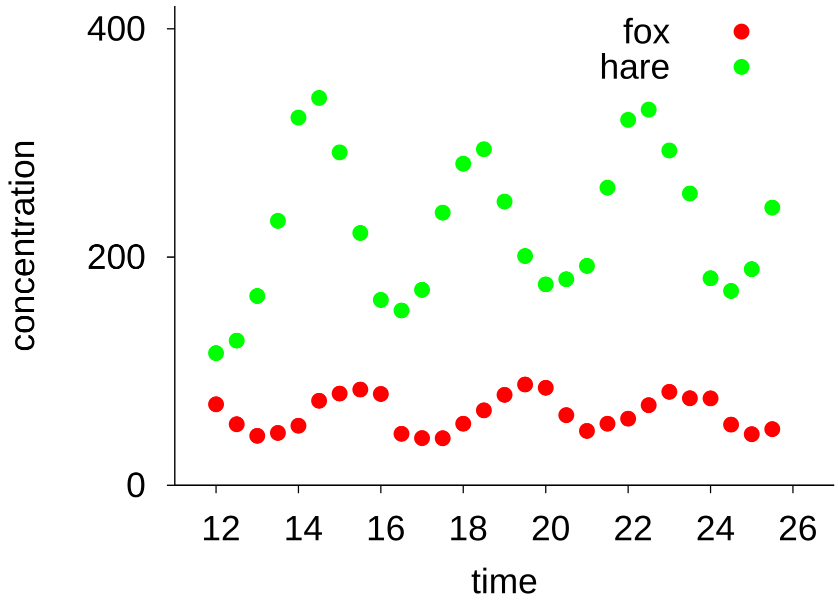
- task: mathematical modeling
- method: grammar-based approach to modeling
- example application: supporting creativity of scientists
- discovering creative solutions?

>

TASK
mathematical modeling

mathematical modeling of population dynamics: input

- GIVEN
  - system/state variables: concentration of species *fox* and *hare*
  - no exogenous variables: could be temperature, light, etc.
  - measurements of variables through time

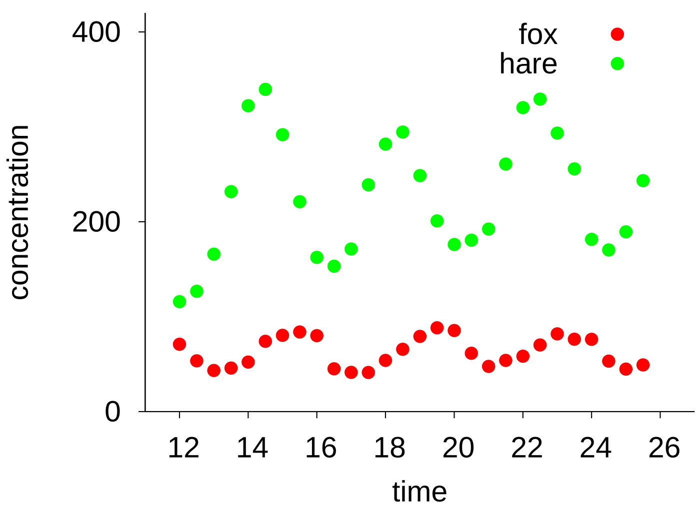


mathematical modeling of population dynamics: output

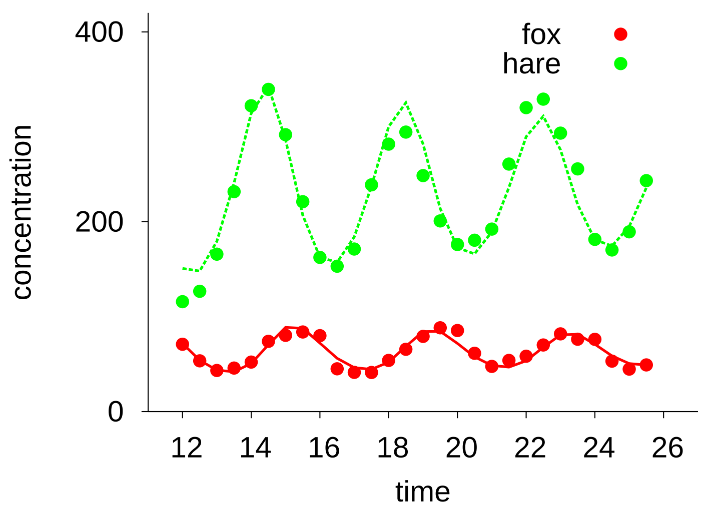
- FIND model, such that its simulation matches measurements

$$\frac{d}{dt} hare = 2.5 \cdot hare - 0.3 \cdot hare \cdot fox$$
$$\frac{d}{dt} fox = 0.1 \cdot 0.3 \cdot hare \cdot fox - 1.2 \cdot fox$$

measurements



measurements and model simulation



mathematical modeling: formal task definition

- GIVEN
  - system (state) variables $s_1, s_2, \ldots s_n$
  - exogenous (input/output) variables $e_1, e_2, \ldots e_m$
  - measurements of the variables in consecutive time points

time	s_1	$\ldots$	s_n	e_1	$\ldots$	e_m
t_0	$s_{1,0}$	$\ldots$	$s_{n,0}$	$e_{1,0}$	$\ldots$	$e_{m,0}$
t_1	$s_{1,1}$	$\ldots$	$s_{n,1}$	$e_{1,1}$	$\ldots$	$e_{m,1}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_N	$s_{1,1}$	$\ldots$	$s_{n,N}$	$e_{1,N}$	$\ldots$	$e_{m,N}$

- FIND mathematical model
  - casted as a set of differential equations

$$\frac{d}{dt} s_i = f_i(s_1, \ldots, s_n, e_1, \ldots, e_m), \quad i = 1 \ldots n$$

  - simulation of the model should match the measurements

motivation for mathematical modeling: important model properties

- prediction
  - a model can predict future behavior of the observed system
  - by running the model simulation beyond the observation period
  - population dynamics model can be used to estimate/predict the concentration of the two species in the next two weeks
- generalization
  - a model integrates large quantities of data in a single entity
  - population dynamics data: tens of measured values
  - population dynamics model: two equations and four parameters
- explanation
  - a model explains the observed behavior of the system
  - by revealing the processes that govern system behavior

population dynamics model: explanation

- terms in the model equations carry semantics

$$\frac{d}{dt} hare = 2.5 \cdot hare - 0.3 \cdot hare \cdot fox$$
$$\frac{d}{dt} fox = 0.1 \cdot 0.3 \cdot hare \cdot fox - 1.2 \cdot fox$$

- they correspond to three processes of
  - growth of the hare population
  - predator-prey interaction between the two populations
  - loss of the fox population

>

COMPUTATIONAL METHOD

lagramge: a grammar-based approach to automated modeling

lagramge: outline

- computational support for the task of mathematical modeling
- generate-and-test approach
  - generate model structure
  - evaluate its quality w.r.t. measurements
- design decisions to be made
  - specification of the search space: what model structures?
  - structure of the search space: ordering of the model structures
  - evaluation: how appropriate is the model structure?

relation between model properties and design decisions

- specification of the search space
  - relates to the explanation property: search space focuses on ”explanatory” model structures
- ordering of the search space
  - relates to the generalization property: ordering from the simplest to more complex model structures
- evaluation of the candidate model structure
  - relates mainly to the prediction property: quality is measured as degree-of-fit between the model simulation and measurements
  - might also relate to the generalization property: simple models are better (occam’s razor and minimal description length approaches)

population dynamics: from model to general schema

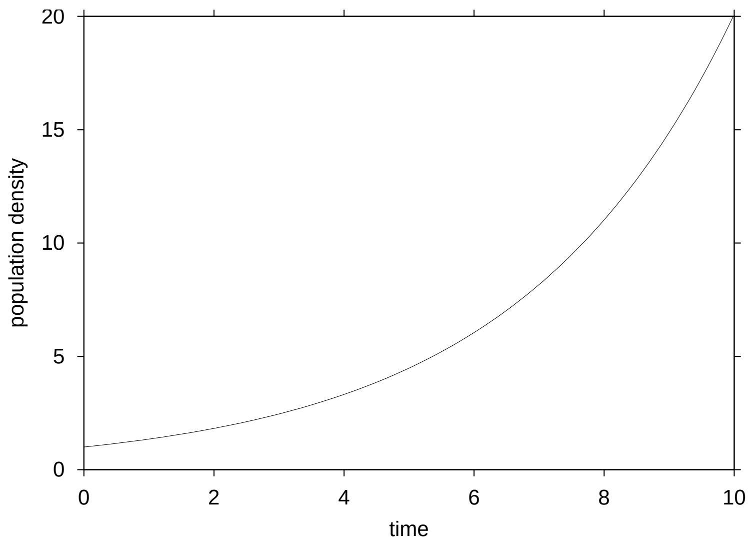
- following the semantics of the terms, we can generalize the model to

$$\frac{d}{dt}hare = \text{growth}(hare) - \text{interaction}(hare, fox)$$
$$\frac{d}{dt}fox = 0.1 \cdot \text{interaction}(hare, fox) - \text{loss}(fox)$$

- the basic Lotka-Volterra model from previous slides assumes
  - unlimited (exponential) hare growth: $\text{growth}(hare) = 2.5 \cdot hare$
  - unlimited (exponential) fox loss: $\text{loss}(fox) = 1.2 \cdot fox$
  - unlimited (unsaturated) predation capacity of fox: $\text{interaction}(hare, fox) = 0.3 \cdot hare \cdot fox$

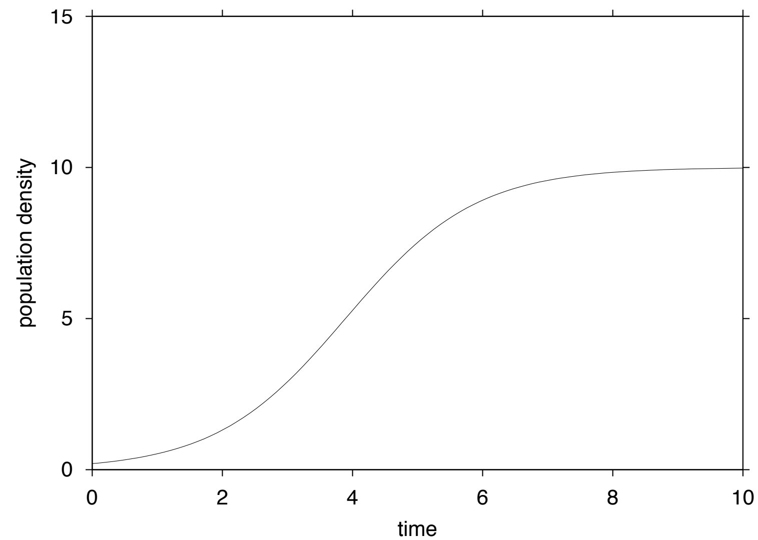
population dynamics: alternatives for modeling growth

exponential growth



$$growth\_rate \cdot P$$

logistic growth



$$growth\_rate \cdot P \cdot (1 - P / carrying\_capacity)$$

population dynamics: search space specification (and ordering)

$$
\begin{array}{rcl}
PPModel & \rightarrow & PreyChange,\ PredatorChange \\
\\
PreyChange & \rightarrow & PreyGrowth - Interaction \\
PredatorChange & \rightarrow & const * Interaction + PredatorLoss \\
\\
PreyGrowth & \rightarrow & const * V_{Prey} \\
PreyGrowth & \rightarrow & const * V_{Prey} * (1 - V_{Prey}/const) \\
\\
Interaction & \rightarrow & const * V_{Predator} * V_{Prey} \\
Interaction & \rightarrow & const * V_{Predator} * V_{Prey}/(V_{Prey} + const) \\
\\
PredatorLoss & \rightarrow & -const * V_{Predator} \\
\\
V_{Prey} & \rightarrow & hare \\
V_{Predator} & \rightarrow & fox
\end{array}
$$

evaluating the quality of a model structure

- for a given model structure
  - find values of the model parameters ($const$) that
  - minimize the discrepancy between simulation and measurements
- technical details
  - gradient descent method for nonlinear optimization
  - teacher forcing used for simulating models
- evaluation function: sum of squared errors

$$\sum_{i=1}^{n}\sum_{j=0}^{N}(s_{i,j} - \hat{s_{i,j}})^2$$

  - $s_{i,j}$ and $\hat{s_{i,j}}$ are the measured and simulated values of the system/state variable s_i in time point t_j

>

HOW DOES THE METHOD SUPPORT SCIENTIFIC CREATIVITY?
example application of the modeling approach

CASA: model revision task

- CASA model
  - developed at NASA-Ames [Potter & Klooster, 1997]
  - global production of biogenic trace gasses in the atmosphere
  - complex system of difference (and not differential) equations
- we focused on one part of CASA (CASA-NPPc)
  - predicts monthly net production of carbon (NPPc) at a site
  - system of algebraic equations
- data set used for revision consists of
  - 8 variables
  - measured at 303 sites/locations around the globe

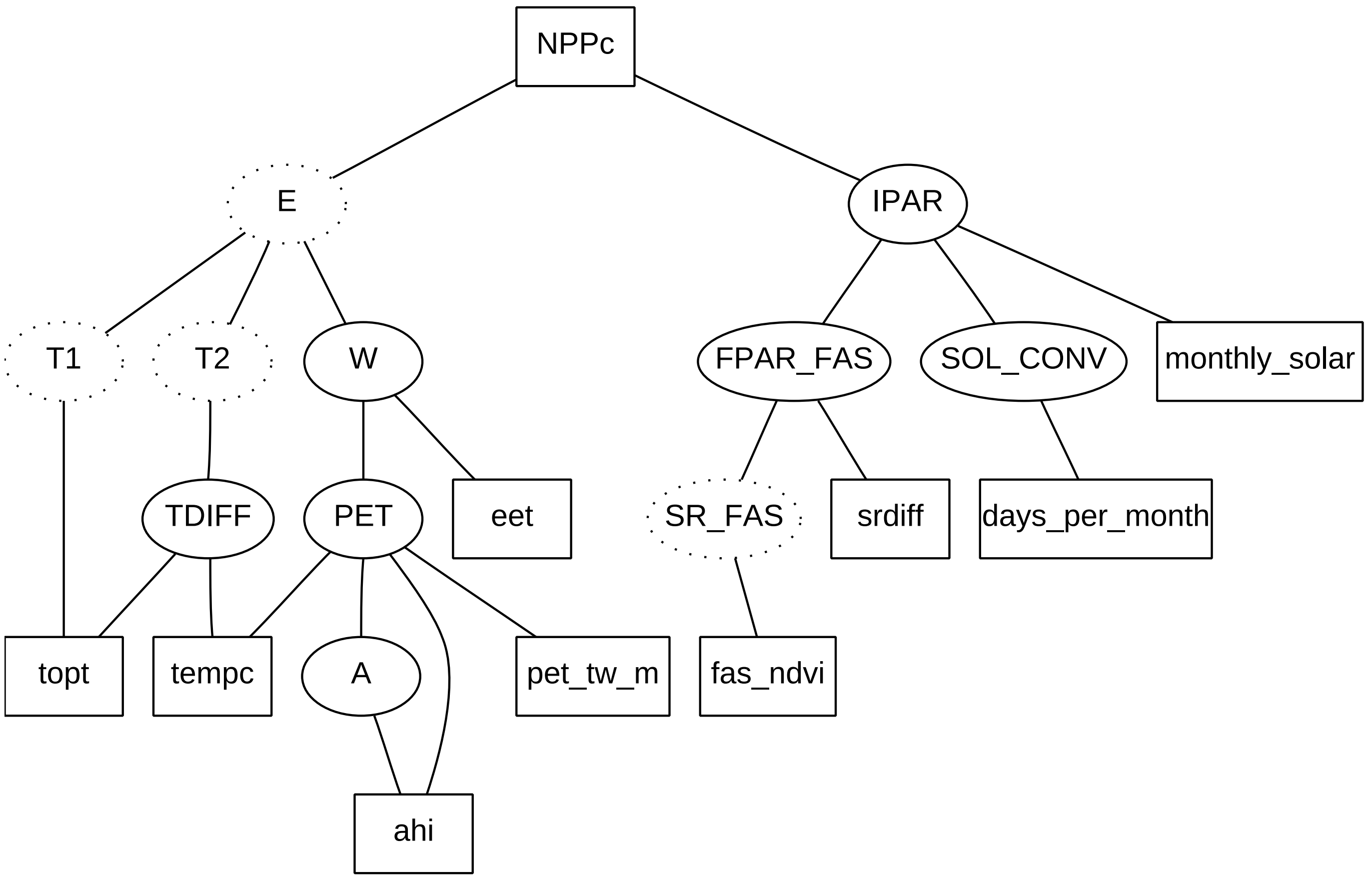
CASA: variables in the model (1)

- NPPc: net production of carbon at a site
- E: photosynthetic efficiency after factoring various sources of stress
- T1: temperature stress factor ($0<T1<1$) for cold weather
- T2: temperature stress factor($0<T2<1$) elsewhere
- W: water stress factor($0.5<W<1$)
- topt (observed): avg.temp. for the month at which fas-ndvi takes its max.value
- tempc (observed): avg.temp. for a given month
- eet (given): estimated evapotranspiration
- PET: potential evapotranspiration

CASA: variables in the model (2)

- pet-tw-m (observed): component of potential evapotranspiration related to the latitude, time of year, and days in month
- A: polynomial function of the annual heat index (ahi, observed)
- fas-ndvi: relative greenness as measured from space
- IPAR: energy intercepted from the sun after factoring in the time of year and days in month
- FPAR-FAS: the fraction of energy intercepted from the sun absorbed photosynthetically after factoring in vegetation type
- monthly-solar (observed): average radiation incoming for a given month
- SOL-CONVER: 0.0864 times the number of days in a month
- UMD-VEG: type of ground cover vegetation

CASA: dependencies between variables



CASA: the initial model

$$
\begin{aligned}
&NPPc = \max(0, E \cdot IPAR)\\
&\quad E = 0.389 \cdot T1 \cdot T2 \cdot W\\
&\qquad T1 = 0.8 + 0.02 \cdot topt - 0.0005 \cdot topt^2\\
&\qquad T2 = 1.1814/((1 + \exp(0.2 \cdot (TDIFF - 10))) \cdot (1 + \exp(0.3 \cdot (-TDIFF - 10))))\\
&\qquad\quad TDIFF = topt - tempc\\
&\qquad W = 0.5 + 0.5 \cdot eet/PET\\
&\qquad\quad PET = 1.6 \cdot (10 \cdot \max(tempc, 0)/ahi)^A \cdot pet\_tw\_m\\
&\qquad\qquad A = 0.000000675 \cdot ahi^3 - 0.0000771 \cdot ahi^2 + 0.01792 \cdot ahi + 0.49239\\
&\quad IPAR = FPAR\_FAS \cdot monthly\_solar \cdot SOL\_CONV \cdot 0.5\\
&\qquad FPAR\_FAS = \min((SR\_FAS - 1.08)/srdiff, 0.95)\\
&\qquad\quad SR\_FAS = (1 + fas\_ndvi/1000)/(1 - fas\_ndvi/1000)\\
&\qquad SOL\_CONV = 0.0864 \cdot days\_per\_month
\end{aligned}
$$

CASA: proposed alternatives

- experts identified four "weak" parts of the CASA-NPPc model
  - equations for E, $T1$, $T2$, and SR_FAS

- two alternatives for $E = 0.0389 \cdot T1 \cdot T2 \cdot W$
  1. revising parameter value: $E = \texttt{const} \cdot T1 \cdot T2 \cdot W$
  2. revising weights of terms: $E = \texttt{const} \cdot T1^{\texttt{const}} \cdot T2^{\texttt{const}} \cdot W^{\texttt{const}}$

- two alternatives for $T1 = 0.8 + 0.02 \cdot topt - 0.0005 \cdot topt^2$:
  1. revising parameter values: $T1 = \texttt{const} + \texttt{const} \cdot topt + \texttt{const} \cdot topt^2$
  2. revising structure: $T1 \rightarrow \texttt{const} \mid \texttt{const} + (T1) * topt$

CASA: revision results

- revised model structure
  - reveals that water stress factor (W) does not influence NPPc
  - initial model: $E = 0.0389 \cdot T1 \cdot T2 \cdot W$
  - revised model: $E = 0.402 \cdot T1^{0.624} \cdot T2^{0.215} \cdot W^{0}$
  - surprising result not expected when specifying the alternatives
- model performance improvement
  - relative error reduction of about 9% (cross-validated)
  - regarded non-trivial by scientists who developed the model
  - taken into account in further model development
  - reductions with single alternatives add up (orthogonal)

minimality of change principle

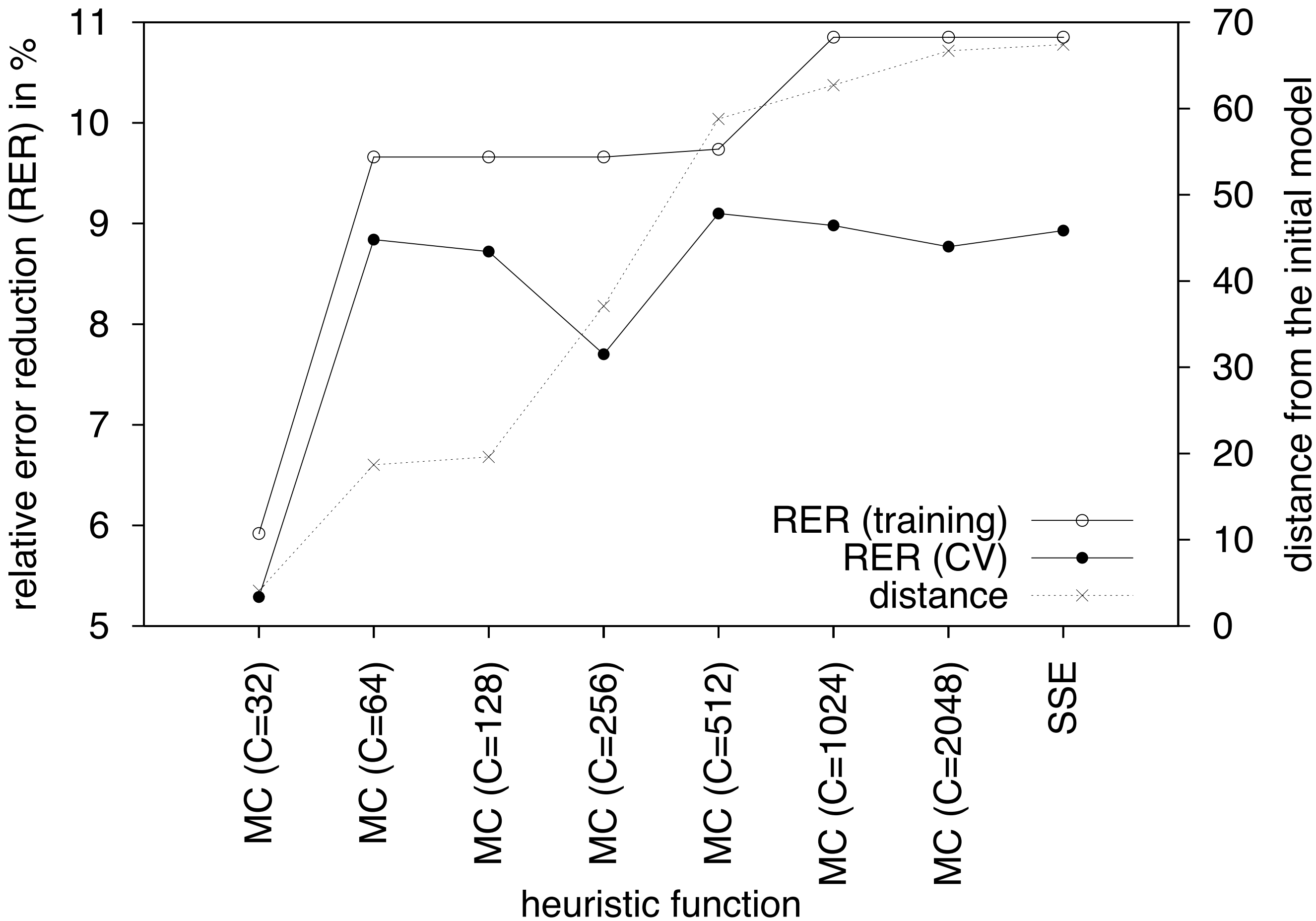
- heuristic function that takes into account distance

$$\mathrm{MC}(M) = \mathrm{SSE}(M) + \frac{\mathrm{distance}(M, M_0)}{C} \cdot \mathrm{SSE}(M_0)$$

  - M and M_0 are the candidate model revision and the initial model, respectively
  - measure of similarity/distance between models needed: editing distance between parse trees used [Richter, 1997]

- trade-off between performance and similarity to the initial model
  - large C: preference for accurate models
  - small C: preference for models similar to the initial one

CASA: the performance-similarity trade-off



supporting creativity: summary

- automated modeling scenario
  - scientists focus on identifying "weak" spots of the model and (optionally) suggesting alternatives for them
  - computational method systematically tests the alternatives and selects/suggests the best ones
  - scientists can trade-off performance gain vs. minimality of change
- in an alternative "manual revision" scenario
  - human expert could try few alternative model structures, perform parameter estimation against measurements for each, and select the best one
  - most of the human expert effort on performing repetitive tasks
- the obtained result surprised the scientists: computational creativity?

supporting creativity: other aspects

- different ways to structure the search space
  - an existing (fully specified) model
  - a partial model specification
  - processes in a particular modeling domain: process-based modeling (Will's talk tomorrow)
- interactive/graphical tools for building models
  - testing different modeling scenarios (set of alternatives)
  - supporting what-if analysis (conceptual simulations)

>

AND COMPUTATIONAL CREATIVITY?

an attempt to evaluate the method in terms of computational creativity

evaluation frameworks considered

- [Ritchie 2001, 2007]: analysis of the relation between
  - conceptual space of artifacts considered by the method
  - inspiring set of artifacts used by the method developer
  - set of artifacts (results) obtained using the method
- [Colton et al. 2001, Pease et al. 2001]: also take into account
  - knowledge used to fine-tune the method
  - (search) process performed by the method
- [Boden 1998]: exploratory vs. transformational creativity

conceptual space of models

- membership (MS)
  - any model structure, i.e., arithmetical expression on the right-hand side of the model equations
- typicality space (TS)
  - includes typical models wrt to the particular modeling domain
  - lagramge expects this space to be encoded in the search space (SS) definition
- quality space (QS)
  - includes good models as judged by a human expert
  - wrt to both measurements and the domain at hand
- result set (RS): model structures that our method produces

inspiring set (IS) of models

- different interpretations possible
  - only models that were used to produce the grammar specification of the search space: in CASA example it is the initial model, in the population dynamics example the initial Lotka-Volterra model (IS0)
  - all model structures that the input grammar generates (ISA)

- or we can distinguish between
  - inspiring set IS0
  - knowledge used to define the search space

most of the results seem negative

- the fact TS = SS
  - leaves no space for novelty and (therefore) creativity [Ritchie 2007]
  - the same holds for the fact: $RS \subseteq ISA$, i.e., $RS \setminus ISA = \emptyset$

- if we distinguish between inspiring set and knowledge
  - the fact $RS \setminus IS0 \neq \emptyset$
  - leaves space for novel models to be obtained

- are they? – yes; human expert considered the CASA revised model
  - to be a surprising result, but confirmed its relevance/quality (this implies $RS \cap QS \neq \emptyset$, an indication of computational creativity)
  - therefore, we might interpret the surprise as novelty

exploratory vs. transformational creativity

- even if we confirm the novelty in a proper empirical study
  - we can speak about novelty within the search space
  - i.e., exploratory creativity
  - namely, lagramge does not include any (meta-level) heuristics for extending the grammar with new alternatives
  - thus, we can not obtain results outside SS
- transformational creativity require meta-level heuristics
  - such as, mutation of existing alternatives to new ones
  - or inventing new processes
  - that would be effective in case of failure to find good model
- more on meta-level extensions of our approach: Will's talk tomorrow

computational creativity: summary and a question

- is our approach computationally creative: probably no!
- is mathematical modeling a creative task: slim chance!
  - "art of mathematical modeling": 2,750 google hits
  - as opposed to 558,000 for the "art of computer programming"
- a (serious) question for discussion: are the paradigms of
  - providing computational support for creative tasks (meta-level focuses on shrinking the search space) and
  - building computational creative methods (meta-level would focus on extending the search space)

  symbiotic or mutually exclusive?

thanks to collaborators, fund providers, and

- YOU FOR THE ATTENTION

- Sašo Džeroski, Jožef Stefan Institute, Slovenia
- Pat Langley & Will Bridewell, Stanford University
  - method development and applications
- Nataša Atanasova & Boris Kompare, University of Ljubljana, Faculty of Civil and Geodetic Engineering, Slovenia
  - modeling population dynamics and aquatic ecosystems
- Christopher Potter & Stephen Clooster, NASA-Ames research center
  - revising the CASA model

- Slovenian Research Agency and National Science Foundation (USA)