Validation and Choosing the Right Model — Robert Marks

Simulating has some advantages and some disadvantages:

1. **Simulation and model building can provide Sufficiency (if you build this model, then this is the behaviour), but not in general Necessity (this is the *only* model that generates this behaviour: the cause).**

2. **But simulation is not restricted to a representative agent; indeed it relishes heterogeneity, diversity.**

3. **And simulation encourages stochastic, probabilistic model building.**

NetLogo is well suited to 2. and 3. And, with BehaviorSpace, to exploring interactions.

1. Sufficiency and Necessity

Simulations demonstrate: existence and sufficiency,

but not necessity.

Simulations can demonstrate the untruth of a proposition,

but not provide proofs or theorems.

Simulations cannot provide generality.

What, never?

Well, hardly ever.

Does this matter?

Formal Simulation

Mathematical "model A" comprises the conjunction $(a_1 \wedge a_2 \wedge a_3 \cdots \wedge a_n)$, where $\wedge$ means "AND", and the a_i denote the elements (equations, parameters, initial conditions, etc) that constitute the model.

***Sufficiency*: If model A exhibits the desired target behaviour B, then model A is *sufficient* to obtain exhibited behaviour B:**
$A \Rightarrow B$

Thus, any model that exhibits the desired behaviour is sufficient, and demonstrates one conjunction of conditions (or model, or solution) under which the behaviour can be simulated. It's an *existence proof.*

But if there are several such models, how can we choose among them? And what is the necessary set $\mathcal{N}$ of all such conjunctions (models)?

Necessity

Necessity: **Only those models A belonging to the set of necessary models $\mathcal{N}$ exhibit target behaviour B.**

That is, $(A \in \mathcal{N}) \Rightarrow B$, and $(D \notin \mathcal{N}) \not\Rightarrow B$.

A difficult challenge: determine the set of necessary models, $\mathcal{N}$.

Since each model is not simple: $A = (a_1 \wedge a_2 \wedge a_3 \cdots \wedge a_n)$, searching for the set $\mathcal{N}$ of necessary models means searching in a high-dimensional space, with no guarantee of continuity, and a possible large number of non-linear interactions among elements.

Lack of Necessity Means ...

For instance, if $D \not\Rightarrow B$, it does not mean that *all* elements a_i of model D are invalid or wrong, only their conjunction, that is, model D.

It might be only a single element a_k that precludes model D exhibiting behaviour B.

But determining whether this is so and which is the offending element a_k is a costly exercise, in general, for the simulator.

Without clear knowledge of the boundaries of the set $\mathcal{N}$ of necessary models, it is difficult to generalise from simulations.

Simulation Can Sometimes Demonstrate Necessity . . .

Only when the set $\mathcal{N}$ of necessary models is known to be small (such as in the case of DNA structure by the time Watson & Crick were searching for it) is it relatively easy to use simulation to derive necessity.

Watson & Crick had much information about the properties of DNA (from others):

when they hit on the simulation we know as the "double helix", they knew it was right.

But still "*A* structure ...", not "*The* structure" in the title of their 1953 *Nature* paper.

(And Kepler's 1605 ellipses?)

Formalisation of Validation

Let set P be the possible range of observed outputs of the real-world system.

Let set M be the exhibited outputs of the model (in any week).

Let set H be the specific, historical output of the real-world system (in any week).

Let set Q be the intersection, if any, between the set M and the set H, $Q \equiv M \cap H$.

We can characterise the model output in several cases. (Mankin et al. 1977).

Five Cases for Validation

a. **no intersection between M and H ($Q = \varnothing$), then the model is *useless.***

b. **intersection Q is not null, then the model is *useful,* to some degree: will correctly exhibit some real-world system behaviours, will not exhibit other behaviours, and will exhibit some behaviours that do not historically occur. Both incomplete and inaccurate.**

c. **If M is a proper subset of H ($M \subset H$) then all the model's behaviours are correct (match historical behaviours), but the model doesn't exhibit all behaviour that historically occurs: accurate but *incomplete.***

d. **If H is a proper subset of M ($H \subset M$) then all historical behaviour is exhibited, but will exhibit some behaviours that do not historically occur: complete but *inaccurate.***

e. **If the set M is equivalent to the set H ($M \Leftrightarrow H$), then (in your dreams!) the model is complete and accurate, but might be overfitted.**

Or Graphically ...

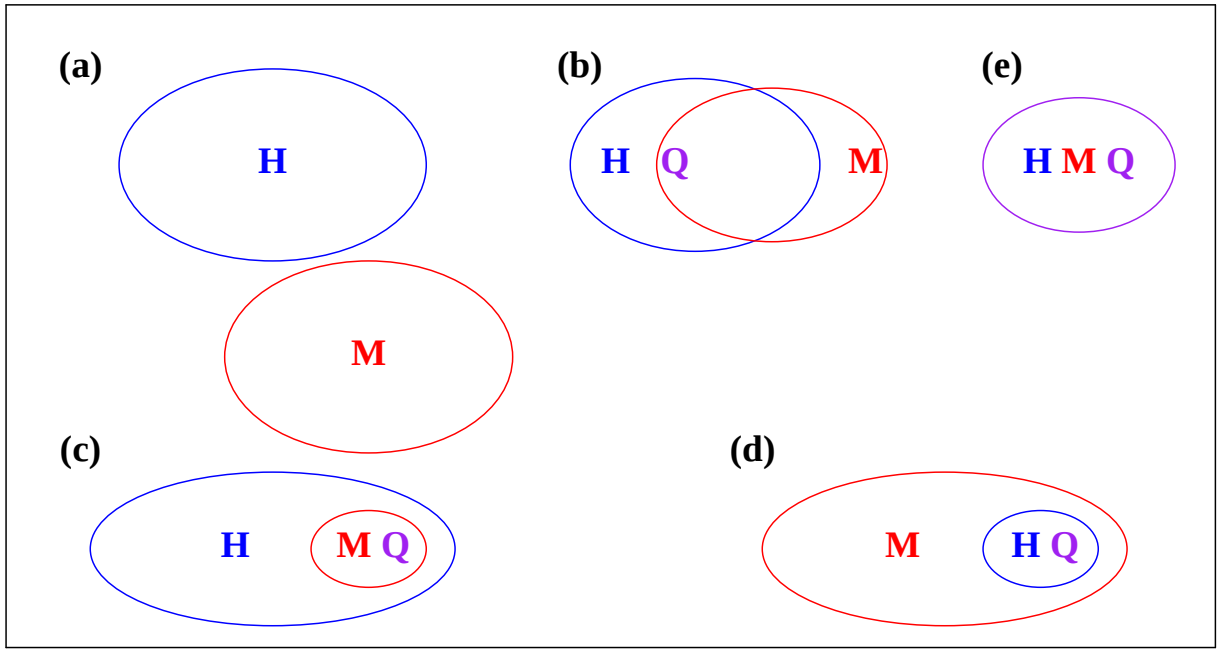


Figure 1: Validity relationships (after Haefner (2005)).

a. **useless**
b. **useful, but incomplete and inaccurate**
c. **accurate but incomplete**
d. **complete but inaccurate ← possibly the best to aim for**
e. **complete and accurate**

Modelling Goals

One goal: to construct and calibrate the model so that

$M \approx Q \approx H$: there are very few historically observed behaviours that the model does not exhibit,

and there are very few exhibited behaviours that do not occur historically.

The model is close to being both complete and accurate.

But in practice, a modeller might be happier to achieve case d., where the model is complete (and hence provides sufficiency for all observed historical phenomena), but not accurate.

Not least to accommodate later real-world observations.

Measures of Validity

A measure of validity which balances the Type I error of inaccuracy with the Type II error of incompleteness.

Define a metric $m()$ (a ratio scale) on the sets.

Define *inaccuracy* α as

$$\alpha \equiv 1 - \frac{m(Q)}{m(M)}\,, \tag{1}$$

and *incompleteness* γ as

$$\gamma \equiv 1 - \frac{m(Q)}{m(H)}\,. \tag{2}$$

Or: $m(|H - M|)$, as in the SSM (see below).

Continued ...

A measure of degree of validation V: a weighted average of inaccuracy α and incompleteness γ:

$$V \equiv \beta(1-\alpha) + (1-\beta)(1-\gamma)$$

$$\therefore V = \beta \frac{m(Q)}{m(M)} + (1-\beta) \frac{m(Q)}{m(H)}$$

$$\therefore V = m(Q)\left(\frac{\beta}{m(M)} + \frac{1-\beta}{m(H)} \right) \tag{3}$$

The value of the weight β, $0 \leq \beta \leq 1$, reflects the tradeoff between accuracy and completeness.

Trade-offs

Possible to reduce incompleteness by generalising the model and so expanding the domain of set M until H is a proper subset of M, as in case d.

Or by narrowing the scope of the historical behaviour to be modelled, so reducing the domain of H (or P).

Also be possible to reduce inaccuracy by restricting the model through use of narrower assumptions and so contracting the domain of M.

If M is sufficiently small to be a proper subset of H, as in case c., then the model will never exhibit anhistorical behaviour.

But not guaranteed to maintain a non-null intersection Q, and it is possible that the process results in case a., with no intersection.

Fagiolo et al. on Validation of AB Models

AB models can be characterized as:

- **bottom-up models (unlike e.g. Systems Dynamics simulation models, or closed-form models)**
- **heterogeneous agents (endowments, properties, memory, rationality, etc.)**
- **boundedly rational, usually with adaptive expectations**
- **networked direct interactions.**

Closer to dynamic, decentralized markets and economies than traditional models.

See: Fagiolo G., Moneta A., & Windrum P. (2007),

Fagiolo 2

And yet reluctance to use AB models. Why? Four key problems:

1. **no common set of the heterogeneous AB models previously developed**
2. **(hence) lack of comparability across these models with high degrees of freedom, hence a wide range of outputs, together with lack of necessity.**
3. **lack of standard techniques for constructing and analyzing AB models**
4. **the "problematic" relationship between AB models and empirical data ← this is *validation*.**

Fagiolo 3

Some questions:

- **Is "realist" methodology appropriate?**
- **Should empirical validation be the primary basis for accepting/rejecting a model?**
- **Are there other tests apart from generating stylized facts?**
- **How should we calibrate the parameters, initial conditions, stochastic variability to historicla data?**
- **How dependable are the micro and macro stylized facts, anyway?**
- **What if the "stylized facts" shed no light on the dynamics of the generating stochastic processes?**
- **What if the "stylized facts" are too general to distinguish among models?**

Fagiolo 4 — Issues with Empirical Validation

Comparing historical data with generated model outputs.

1. **The world is complex: a trade-off between complexity in modelling ("concretization") and reductionism ("isolation"): where to draw the line in modelling? Realism v. tractability, again.**
2. **Friedman (1953) argued that realism was not necessary so long as the output allowed accurate prediction ("instrumentalism"), but others seek realism in the model and its assumptions as well as accuracy.**
3. **How wedded should the modeller be to a priori assumptions (about the goals of agents, say)? Or should all aspects of the model be available ("pluralism")?**
4. **Importantly: how to choose which of several models is best (the "identification" or "under-determination" problem).**

Fagiolo 5

Three alternative approaches:

1. **the indirect calibration approach**
2. **the Werker-Brenner appraoch**
3. **the history-friendly approach, and**

4. **the State Similarity Measure (of mine, below).**

Monte Carlo — Probabilistic Sampling

1. **Run two simulation models for n runs ($n > 50$).**
2. **Save the relevant output data from each MC run. This provides a distribution for each (stochastic) model, A and B.**
3. **Other models generate (via MC) other distributions, or we could have historical data to compare with. e.g. cumulative winnings for agents with degrees of risk preference choosing lotteries.**
4. **Null hypothesis: the two means of the outputs come from the same distribution: H_0: $\mu_A - \mu_B = 0$.**
5. **Statistical testing for the null hypothesis. Use confidence intervals: the chance that the difference between the two means is *not* random.**

Case 1: Using a GA in NetLogo to Search for Risk Profile

What is the best risk profile for an agent to have in a risky world?

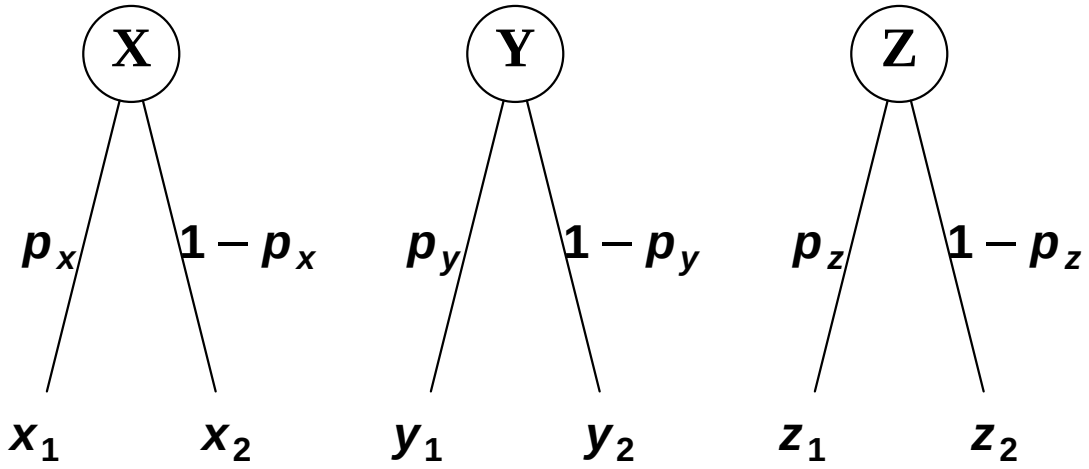
"Best" means that the agent's average net winnings from choosing successive lotteries is highest: this is risky decision making.

By "risky" is meant that both the possible outcomes and probabilities (which might be subjective) are known.

Informally, it is widely held that in a risky world, with the possibility of the discontinuity of bankruptcy, the most prudent risk profile is *risk aversion.*

A Risky Decision

Three two-prize lotteries with random prizes and probability:



Calculate the three expected utilities:

$$U(X) = p_x U(x_1) + (1-p_x)U(x_2)$$
$$U(Y) = p_y U(y_1) + (1-p_y)U(y_2)$$
$$U(Z) = p_z U(z_1) + (1-p_z)U(z_2)$$

Choose the lottery I with the highest expected utility. Win whichever prize (i_1 or i_2) is realised in that lottery, based in the lottery's probability p_i.

Methodology

We use three kinds of utility function:

1. **the wealth-independent exponential function, or CARA;**
2. **a wealth-dependent function, the CRRA; and**
3. **the DRP from Prospect Theory.**

These reflect actual human decision making.

Then run computer experiments in which each agent successively chooses among three lotteries, and is then awarded with the outcome of the chosen lottery k.

Repetition of this choice by many agents allows us use the Genetic Algorithm (Holland 1992) to search for the best risk profile, where "best" means the highest average payoff when chosing among lotteries.

CARA Utility Functions

The exponential CARA utility function is

$$U(x) = 1 - e^{-\gamma x}, \tag{4}$$

where $U(0) = 0$ and $U(\infty) = 1$, and
where γ is the *risk aversion coefficient*:

$$\gamma \equiv -\frac{U''(x)}{U'(x)}. \tag{5}$$

Sign of γ	Risk profile	Curvature
$\gamma = 0$	risk neutral	$U''(x) = 0$
$\gamma > 0$	risk averse	$U''(x) < 0$
$\gamma < 0$	risk preferring	$U''(x) > 0$

CRRA Utility Functions

The Constant Elasticity of Substitution (CES) CRRA utility function:

$$U(w) = \frac{w^{1-\rho}}{1-\rho}, \quad w > 0, \tag{6}$$

where w is agent's wealth, and ρ is the Arrow-Pratt measure of relative risk aversion (RRA):

$$\rho(w) = -w\,\frac{U''(w)}{U'(w)} = w\gamma \tag{7}$$

This introduces wealth w into the agent's risk preferences, so that lower wealth can be associated with higher risk aversion. The coefficient γ is as in (5).

As $\rho \to 1$, (6) becomes logarithmic: $u(w) = \ln(w)$, risk averse.

With $w > 0$, $\rho > 0$ is equivalent to *risk averse*, while $\rho < 0$ is equivalent to *risk preferring*; $\rho = 0$: *risk neutral*.

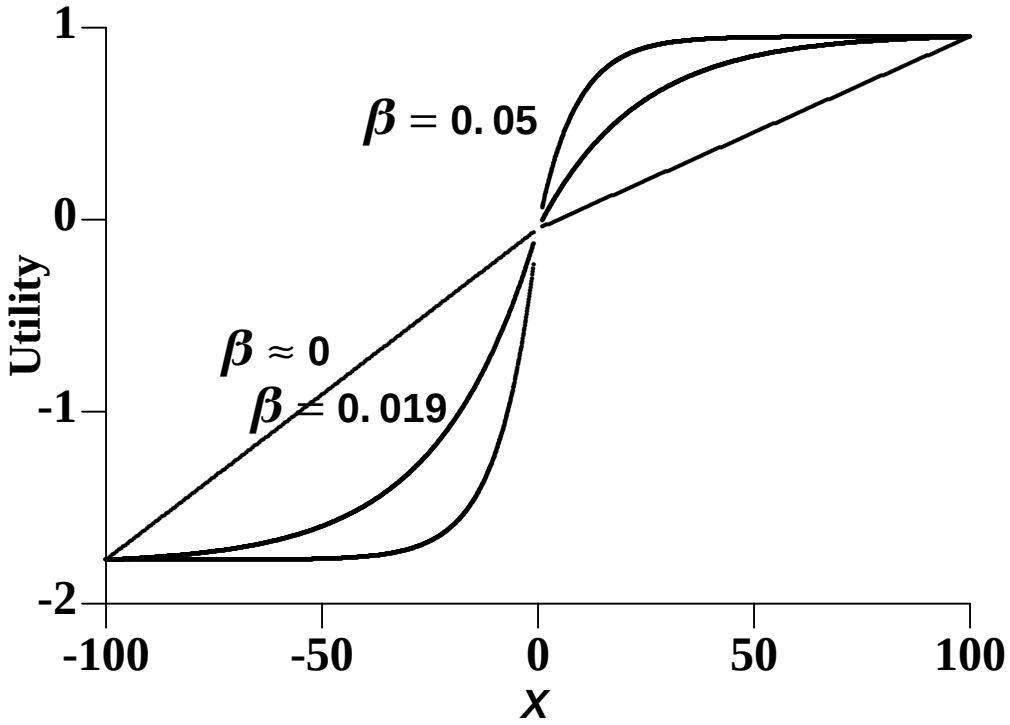
The DRP function

This function is inspired by Prospect Theory (Kahnemann & Tversky 1979):

$$U(X) = \frac{1 - e^{-\beta X}}{1 - e^{-100\beta}}, \quad 0 \le X \le 100 \tag{8}$$

$$U(X) = -\delta \frac{1 - e^{\beta X}}{1 - e^{-100\beta}}, \quad -100 \le X \le 0. \tag{9}$$

$\beta > 0$ models the curvature of the function, and $\delta \ge 1$, the asymmetry associated with losses. The DRP function is not wealth-independent. DRP exhibits the S-shaped asymmetric function of Prospect Theory. It exhibits risk seeking (loss aversion) when X is negative with respect to the reference point, $X = 0$, and risk aversion when X is positive.



A DRP Function (δ = 1.75).

As $\delta \to 1$ and $\beta \to 0$, the value function asymptotes to a linear, risk-neutral function.

4. The Simulations

Each lottery is randomly constructed: the two payoffs ("prizes") are randomly chosen in the interval [−$100, +$100], and the probability is also chosen randomly.

Each agent calculates the expected utility of each of the three lotteries, using its utility function (a function of its γ or ρ/w or (β, δ)), and chooses the lottery k with the highest expected utility. To do this, agents know the prizes and probabilities of all three lotteries.

Then the actual (simulated) outcome of the chosen lottery k is randomly realised, using its probability. The winnings of agent is incremented accordingly. Each agent chooses 1000 lotteries.

Use an implementation of the GA (Gilbert 2005) in NetLogo to search for the best risk profile.

`http://www.agsm.edu.au/bobm/teaching/SimSS/NetLogo4-models/DRA-CRRA-EU-GM-revH-3l2p.html`

The Results

Mean results across *n* Monte Carlo runs:

Model	Variable/s	Mean	*p*-value	Fitness	*n* MC runs
CARA	γ	0.0064	≈ 0	\$37.65	55
CRRA	ρ	19.21	0.30	\$29.40	109
DRP(a)	δ	1.466	≈ 0	\$37.67	55
DRP(b)	β	0.0047	≈ 0	\$38.81	102
DRP(c)	δ &	1.256	≈ 0	\$37.72	54
	β	0.0072	≈ 0		

In DRP(a) $\beta = 0$ and search for δ.
In DRP(b) $\delta = 1$ and search for β.
In DRP(c) search for β and δ jointly.

In all cases the null hypothesis is risk neutrality ($\gamma = \rho = \beta = 0$ and $\delta = 1$).

CRRA is the only case here where the null is not rejected.

But the Fitness of DRP(b) > DRP(c) (\$38.81 > \$37.72): the GA has not converged!

This suggests two further models:

4. **Linear, risk-neutral**
5. **Clairvoyant**

Model DRP(c) is unconstrained, and so can find $\delta = 1$ (with the higher Fitness) but has not yet done so: the GA has not (yet) converged.

Strongly suggests that risk neutral (with $\delta = 1$ and $\beta = 0$) will have a higher Fitness than $38.81.

Short cut: introduce the Linear, risk-neutral model.

The Clairvoyant model has the highest possible Fitness, but requires clairvoyance or perfect foresight.

All Models Compared

Model	Variable/s	Mean	p-value	Fitness	n MC runs
CARA	γ	0.0064	≈ 0	\$37.65	55
CRRA	ρ	19.21	0.30	\$29.40	109
DRP(a)	δ	1.466	≈ 0	\$37.67	55
DRP(b)	β	0.0047	≈ 0	\$38.81	102
DRP(c)	δ &	1.256	≈ 0		
	β	0.0072	≈ 0	\$37.72	54
Linear				\$39.19	102
Clairvoyant				\$50.25	50

So the hunch from the GA results is correct: Linear has the highest Fitness.

To confirm: the data from DRP(b) and Linear are statistically distinct, using a Wilcoxen Two-Sample Test.

Case 2: Comparing Sets of Time-Series

Q: how can we measure the degree of similarity of two sets of time-series?

***One*: the historical record of the rivalrous dance among the sellers in an oligopoly, while**

***The other*: the output from a (agent-based) simulation model of the market, where each seller agent prices this week as a function of the state of the market last week (or earlier).**

Q: how can we *output validate* our model against history?

Or: how can we derive a *degree of confidence* in the model output?

The Issue: Heterogenous Agents and Time-series Prices

Two reasons to compare such model output against history:

1. **To choose better parameter values, to "calibrate" or (more formally) "estimate" the model against the historical record.**
2. **To measure how closely the output reflects history, to *validate* the model.**

We are interested in the second, having used machine learning (the Genetic Algorithm) to derive the model parameters in order to improve each agent's weekly profits (instead of fitting to history) in our agent-based model.

Figure 1 shows historical data from a U.S. supermarket chain's sales of (heterogeneous) brands of sealed, ground coffee, by week in one city (Midgley et al. 1997).

This simulation (which also uses the Genetic Algorithm) was written in C, not NetLogo, because we started this research before NetLogo was available.

Historical Data: Market Prices and Volumes

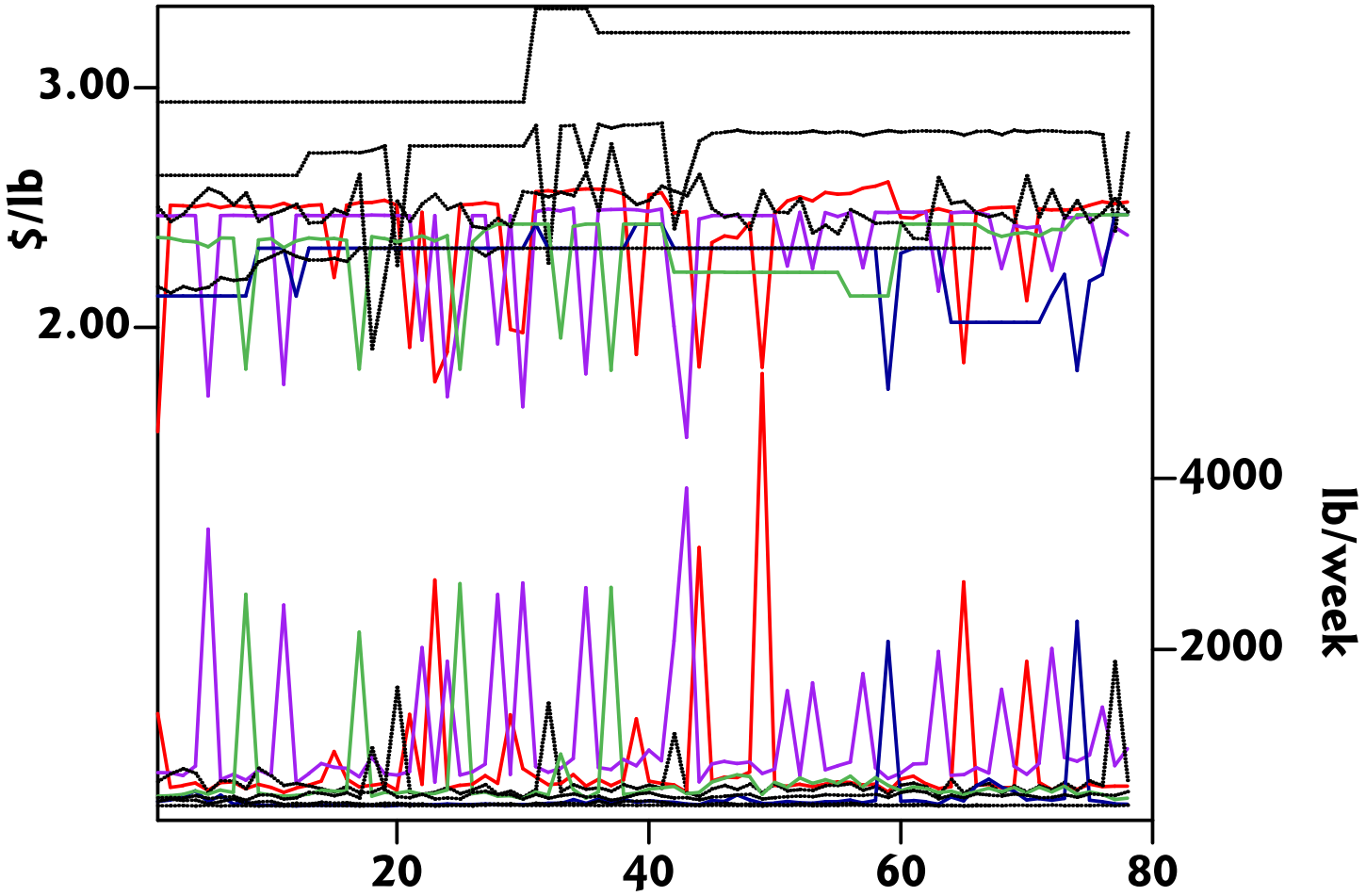


Figure 1: **Weekly Sales and Prices (Source: Midgley et al. 1997)**

Stylised Facts of the Market Behaviour

- **Much movement in prices and quantities of four brands — a rivalrous dance.**
- **Pattern: high price (and low quantity) punctuated by low price (and high quantity).**
- **Another four brands: stable prices and quantities**

Questions:

What is the cause of these patterns?

- **— shifts in brand demand?**
- **— reactions by brands?**
- **— actions by the supermarket chain?**
- **— unobserved marketing actions?**

Explanations?

Interactions of profit-maximising agents, plus external or internal factors $\rightarrow$ via a model $\rightarrow$ behaviour

Similar (qualitatively or quantitatively) to the brands' behaviours of pricing and sales.

Note: assuming profit-maximising (or purposeful) agents means that we are not simply curve-fitting or description using D.E.s. Going beyond the rivalrous dance.

Dichotomous Price Partitioning of the Historical Data

To handle the curse of dimensionality.

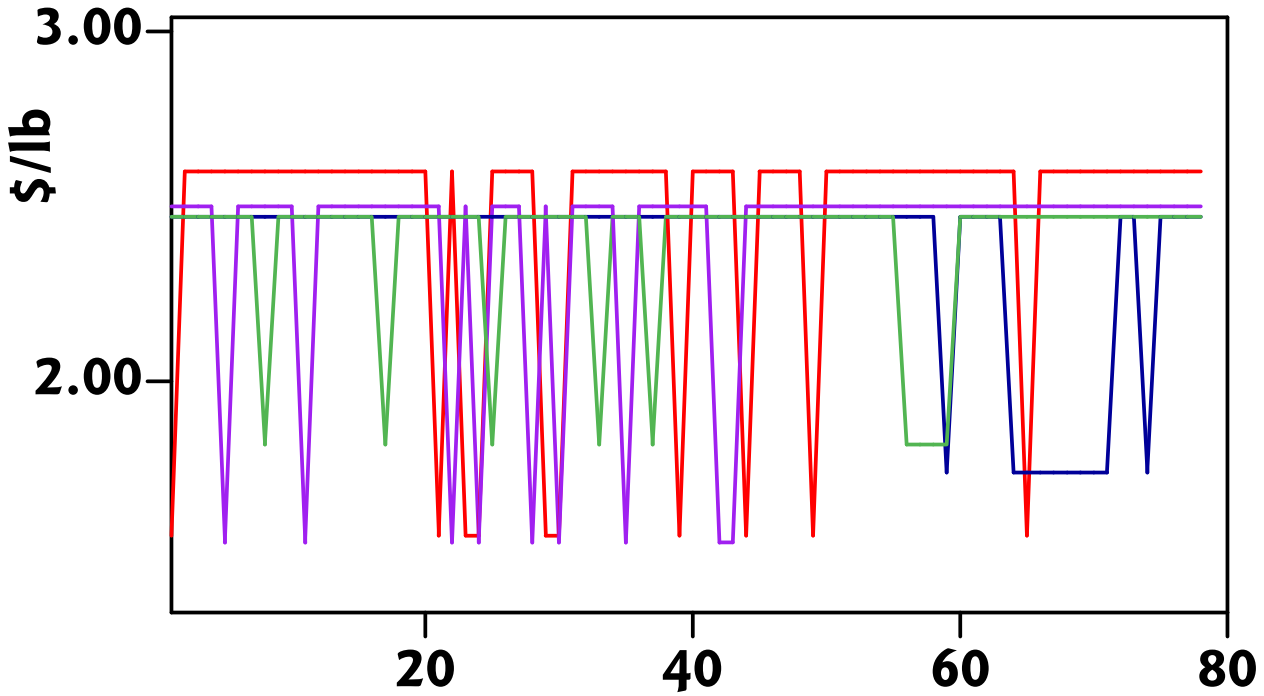


Figure 2: **Partitioned Historical Weekly Prices of the Four Brands**

A Model of Strategic Interaction

We assume that the price P_{bw} of brand b in week w is a function of the state of the market M_w at week w, where M_w in turn is the product of the weekly prices S_w of all brands over several weeks:

$$P_{bw} = f_b(M_w) = f_b(S_w \times S_{w-1} \times S_{w-2} \cdots)$$

Earlier in the research program undertaken with David Midgley et al., we used the Genetic Algorithm to search for "better" (i.e. more profitable) brand-specific mappings, f_b, from market state to pricing action.

And derived the parameters of the model, and derived its simulated behaviour, as time-series patterns (below).

The State Similarity Measure (SSM)

The SSM (Marks 2013) derives the distance between two sets of time-series, by calculating the sum of absolute differences in observed window states between the two set, so what?

First, the greater the sum, the more distant the two sets of time-series.

Second, we can calculate the maximum size of the summed difference: zero intersection between the two sets (no states in common) implies a measure of $2 \times S$ where S is the number of possible window states, from the data.

Third, we can derive some statistics to show that any pair of sets in not likely to include random series (below).

Example of a Simulated Oligopoly (Marks et al. 1995)

Simulating rivalry between the three asymmetric brands: 1, 2, and 5, Folgers, Maxwell House, and Chock Full O Nuts.

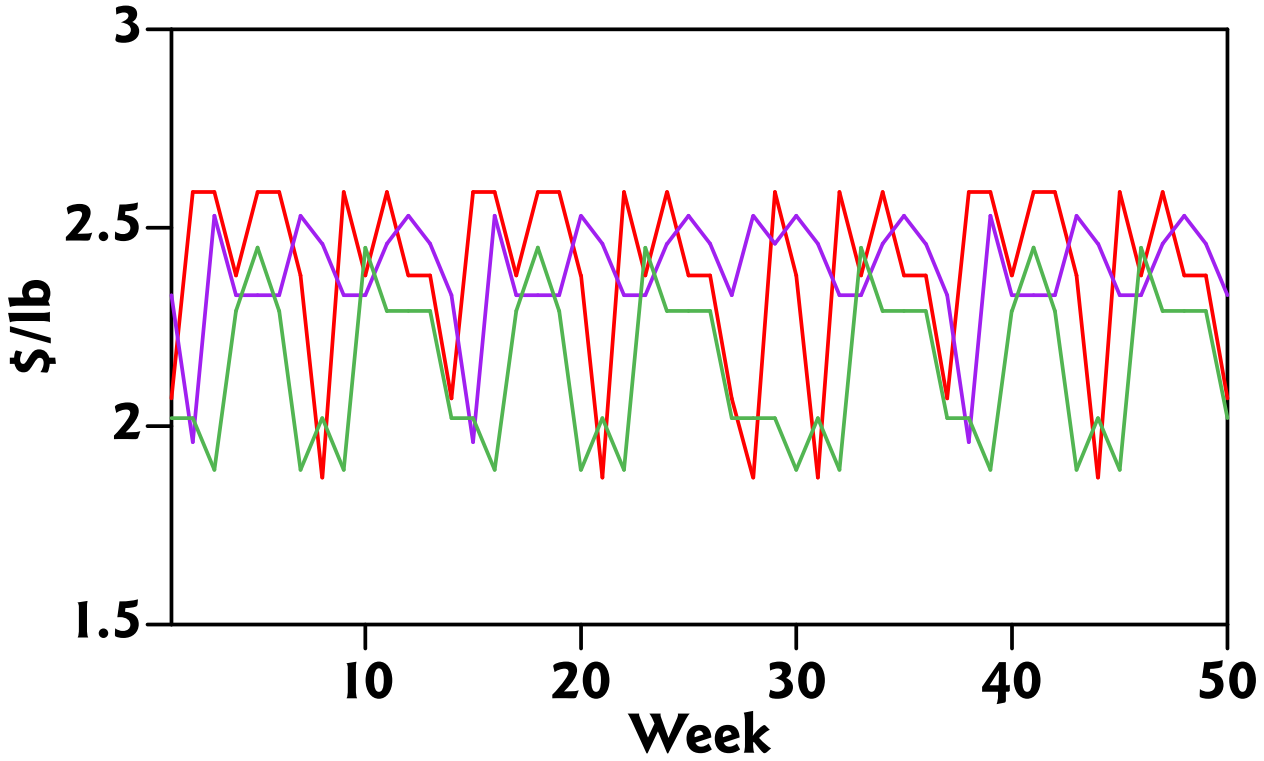


Figure 3: **Example of a Simulated Oligopoly (Marks et al. 1995)**

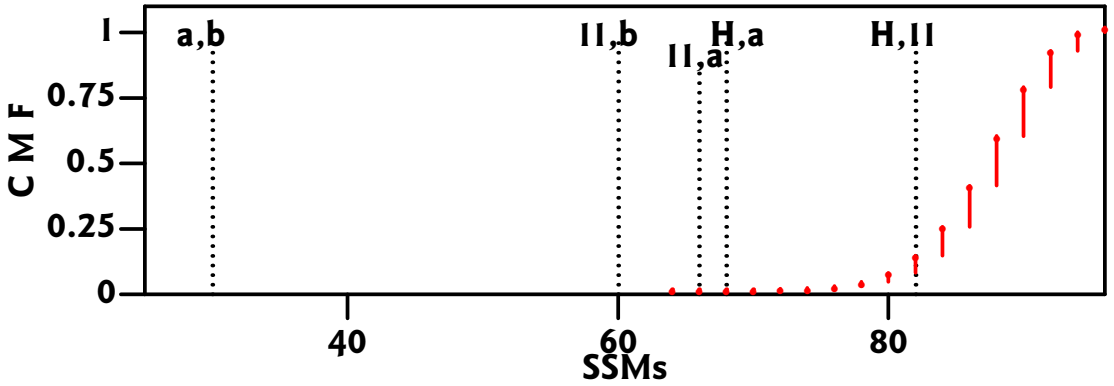
Distances Between History and Three Runs (Brands 1, 2, 5)

	History	Run 11	Run 26a	Run 26b
History	0	82*	68	68
Run 11	82*	0	66	60
Run 26a	68	66	0	30
Run 26b	68	60	30	0

Table: Distances Between History and Three Runs (Brands 1, 2, 5)
(* : cannot reject the null at the 5% level)

Here, S, the maximum number of states = 48, so the maximum distance apart is 96. The three Runs are closer to each other than to the Historical Data; Runs 26a and 26b are very close, only 30/96 = 31.25% apart.

Testing for Randomness



The red lines are the CMF of pairs of sets of random series (3 series, 48 observations) from 100,000 Monte Carlo parameter bootstraps.

The one-sided confidence interval at 1% corresponds to a SSM of 76, and at 5% 80.

Cannot reject the null hypothesis (random sets) for Historical data and Run 11; reject the null (random) hypothesis for all other pairs.

Conclusions — the State Similarity Measure

This measure, *the State Similarity Measure* (SSM), is sufficient to allow us to put a number on the degree of similarity between two sets of time-series which embody dynamic responses.

Such a metric is necessary for scoring the distance between any two such sets, which previously was unavailable.

Here, the SSM has been developed to allow us to measure the extent to which a simulation model that has been chosen on some other criterion (e.g. weekly profitability) is similar to historical sets of time-series.

The SSM will also allow us to measure the distance between any two sets of time-series and so to estimate the parameters, or to help calibrate a model against history.

References

[1] Fagiolo G., Moneta A., & Windrum P. (2007), "A critical guide to empirical validation of agent-based models in economics: methodologies, procedures, and open problems," *Computational Economics*, 30(3): 195–226.

[2] Friedman M. (1953), *Essays in Positive Economics*, University of Chicago Press.

[3] Gilbert, N. (2005) Axelrod's Iterated prisoners' dilemma tournament, from Chapter 10 of *Simulation for the Social Scientist,* by N. Gilbert and K.G. Troitzsch, 2nd ed., Open University Press, 2005.
http://cress.soc.surrey.ac.uk/s4ss/code/NetLogo/axelrod-ipd-ga.html

[4] Haefner J.W. (2005), *Modeling Biological Systems: Principles and Applications*, New York: Springer, 2nd ed.

[5] Holland J.H. (1992), *Adaptation in Natural and Artifical Systems,* MIT Press.

[6] K. L. Judd, Computationally Intensive Analyses in Economics, *Handbook of Computational Economics, Volume 2: Agent-Based Modeling*, ed. by Leigh Tesfatsion & Kenneth L. Judd, Amsterdam: Elsevier Science, 2006, Ch. 2.

[7] Kahneman, Daniel, Tversky, Amos (1979), "Prospect Theory: An Analysis of Decision under Risk", *Econometrica* 47 (2): 263.

[8] I. Lorscheid, B.-O. Heine, M. Meyer, Opening the "black box" of simulations: increased transparency and effective communication through the systematic design of experiments, *Comput Math Organ Theory* (2012) 18: 22–62. DOI 10.1007/s10588-011-9097-3

[9] Mankin, J.B., O'Neill, R.V., Shugart, H.H., and Rust, B.W. (1977), The importance of validation in ecosystem analysis, in G.S. Innis, ed. *New Directions in the Analysis of Ecological Systems, Part 1, Simulation Council Proceedings Series*, Simulation Councils, La Jolla, California, 5: 63–71. Reprinted in H.H. Shugart and R.V. O'Neill, eds. *Systems ecology*, Dowden, Hutchinson and Ross, Stroudsburg, Pennsylvania, 1979, pp. 309–317.

[10] Marks, R.E. (2007), "Validating simulation models: a general framework and four applied examples", *Computational Economics*, 30(3): 265–290. http://www.agsm.edu.au/bobm/papers/s1.pdf

[11] Marks, R.E. (2013), "Validation and model selection: three measures compared," *Complexity Economics*, 2: 41–61, May. http://www.agsm.edu.au/bobm/papers/parispaper.pdf

[12] **Marks R.E. (2014) "Monte Carlo," in *The Palgrave Encyclopædia of Strategic Management*, edited by David Teece and Mie Augier, (London: Palgrave), forthcoming.**
http://www.agsm.edu.au/bobm/papers/MonteCarloPESM-longer.pdf

[13] **Marks R.E. (2014) "Risk Neutral is Best for Risky Decision Making," http://www.agsm.edu.au/bobm/papers/singapore14.pdf**

[14] **Marks R.E., Midgley D.F., and Cooper L.G. (1995) "Adaptive behavior in an oligopoly," *Evolutionary Algorithms in Management Applications,* ed. by J. Biethahn and V. Nissen, (Berlin: Springer-Verlag), pp.225–239.**

[15] **Midgley, D.F., Marks, R.E., and Cooper, L.G. (1997) "Breeding competitive strategies," *Management Science*, 43(3): 257–275.**
http://www.santafe.edu/media/workingpapers/95-06-052.pdf

[16] **D.F. Midgley, Marks R.E., and Kunchamwar D. (2007) "The Building and Assurance of Agent-Based Models: An Example and Challenge to the Field," *Journal of Business Research*, Special Issue: Complexities in Markets, 60(8): 884–893.**

[17] **Watson, J. D., & Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* 171, 737–738 (1953)**