Validation and Model Selection: Three Measures

Global Systems Dynamics & Policy
Agent-Based Modeling Workshop
Salle Marc Bloch et Cavailles,
Sorbonne, Paris
September 8–10, 2011

Robert Marks,
Economics, UNSW, Sydney, and
the Melbourne Business School

bobm@agsm.edu.au

Outline

1. Introduction — Toy or Tool?
2. Validation and Model Selection
3. One Issue: Heterogeneous Agents, Sets of Time-series of Prices, Defining the States of the Market
4. Three Measures of the Distance Between Sets of Time-Series
   a. Kullback-Leibler (K-L) Information Loss
   b. The *State Similarity Measure* (SSM)
   c. The Generalized Hartley Measure (GHM)
5. Conclusions

1. Introduction — Toy or Tool?

Toys?

We all remember the "Sim" line of computer games — an early example of ABM.

Toys like this are serious business:

although not an ABM, last year *Call of Duty: Black Ops* took in over USD$650 million of sales in the game's first five days, which set a five-day global record for a movie, book or videogame.

Tools?

Despite the general limitation of simulations in general (and ABM in particular) to provide sufficiency only, but simulations can help explain simple phenomena in principle, and complex historical phenomena — although this latter requires validation.

One use of CAS modelling is to find *sufficient conditions* for the model to exhibit such characteristics.

See the discovery of the structure of DNA by Crick and Watson in 1953 through simulations of physical models, constrained by known properties of the molecule. ("A Structure ...")

And one may derive suggestive trade-offs and hypotheses, to be tested.

Two Kinds of ABM

I suggest that we can think of two kinds of ABM:

1. ***demonstrative* ABM models**

   These models demonstrate principles, rather than tracking historical phenomena. A demonstrative ABM is an existence proof.
   Examples: Schelling's Segregation Game, my Boys and Girls NetLogo model, my Emergence of Risk Neutrality, and others

2. ***descriptive* ABM models.**

   These models attempt to derive sufficient conditions to match historical phenomena, as reflected in historical data. This requires validation (model choice).
   Examples: Midgley et al. modelling brand rivalry, poli sci models (Earnest).

2. Validation and Model Selection

"All models are wrong, but some are useful" — Box (1976).

Previously, validation (as the name suggests): *validating* a model as "close" to historical reality, in some way — models are approximations, so the comparison dimension must be a function of the model's purpose.

But Anderson & Burnham (2002) make a strong case for validation as *model selection*: for the researcher generating a selection of models, and choosing the model which loses the least *information* compared to reality (the historical data).

How to select a "best approximating model" from the set? Anderson & Burnham review and use Akaike's information criterion (AIC).

Akaike's Information Criterion (AIC)

Considering AIC to be an extension of R. A. Fisher's likelihood theory, Akaike (1973) found a simple relationship between Kullback-Leibler "distance" or "information" and Fisher's maximized log-likelihood function.

This leads to very general methodology for selecting a parsimonious approximating model.

Can think of modelling as being directed towards finding a good approximating model of the information encoded in the empirical, historical data.

Information about the process under study exists in the data. Want to express this information in a model: more compact, and understandable.

The role of a good model is to filter the historical data so as to separate information from noise.

3. One Issue: Heterogenous Agents, Time-series Price, Defining the States of the Market

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 choose the "best" model from a selection of possible models (different structures, parameter values, etc)**

We are interested in the second, having used machine learning (the GA) 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).

Historical Data: Prices and Volumes in Chain 1

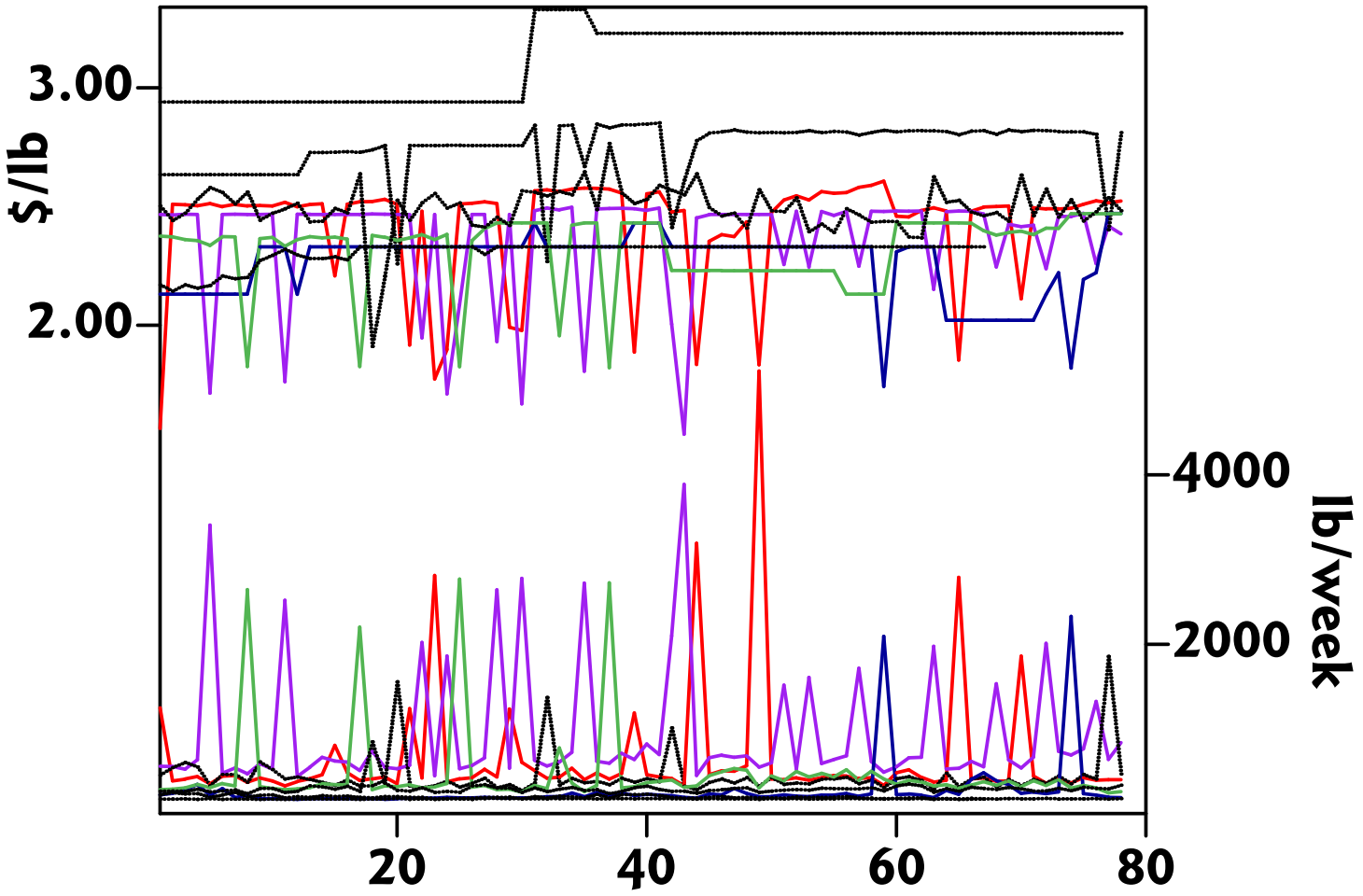


Figure 1: **Weekly Sales and Prices (Source: Midgley et al. 1997)**

Stylised facts of the historical data:

1. **Much movement in the prices and volumes of four strategic brands,**

2. **For these four (coloured) brands, high prices (and low volumes) are punctuated by a low price (and a high volume).**

3. **Another five (non-strategic) brands exhibit stable (high) prices and (low) volumes.**

In addition, the competition is not open slather: the supermarket chain imposes some restrictions on the timing and identity of the discounting 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-1} at week $w-1$, where M_{w-1} in turn might be a product of the weekly prices S_{w-j} of all brands over several weeks:

$$P_{bw} = f_b(M_{w-1}) = f_b(S_{w-1} \times S_{w-2} \times S_{w-3} \cdots)$$

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 models, and derived their simulated behaviour, as time-series patterns (below).

Partitioning the Data

A curse of dimensionality: each brand can price anywhere between \$1.50 and \$3.40 per pound: 190 price points. Consider three stragic brands only.

The first coarsening:
Marks (1998) explores partitioning while maximising information (using an entropy measure). Finds that dichotomous partition is sufficient.

Here: use symmetric dichotomous partitioning: a brand's price is labelled 0 if above its midpoint, else 1 below.

The second coarsening:
Consider three depths of memory:
with 1-week memory, three brands, each pricing Low or High: $2^3 = 8$ possible states;
with 2-week memory: $8^2 = 64$ possible state;
with 3-week memory: $64^2 = 512$ possible states.

Dichotomous Symmetric Price Partitioning of Chain 1

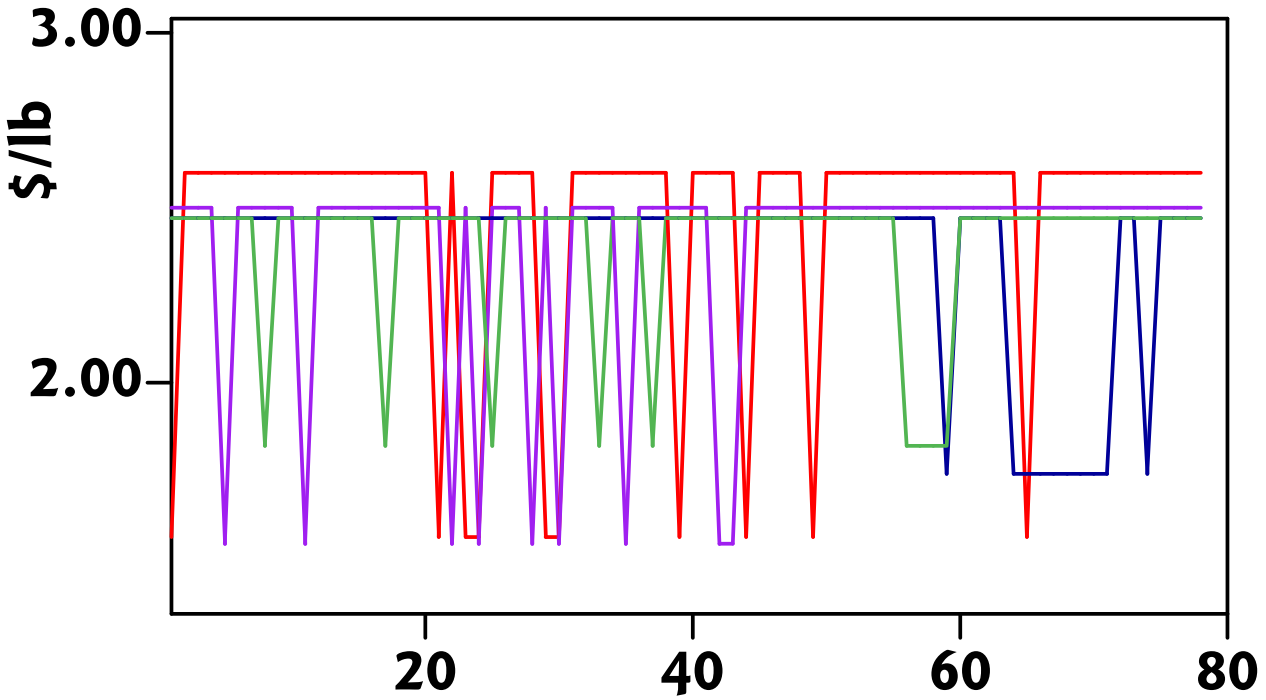


Figure 2: **Partitioned Weekly Prices of the Four Chain-One Brands**

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.

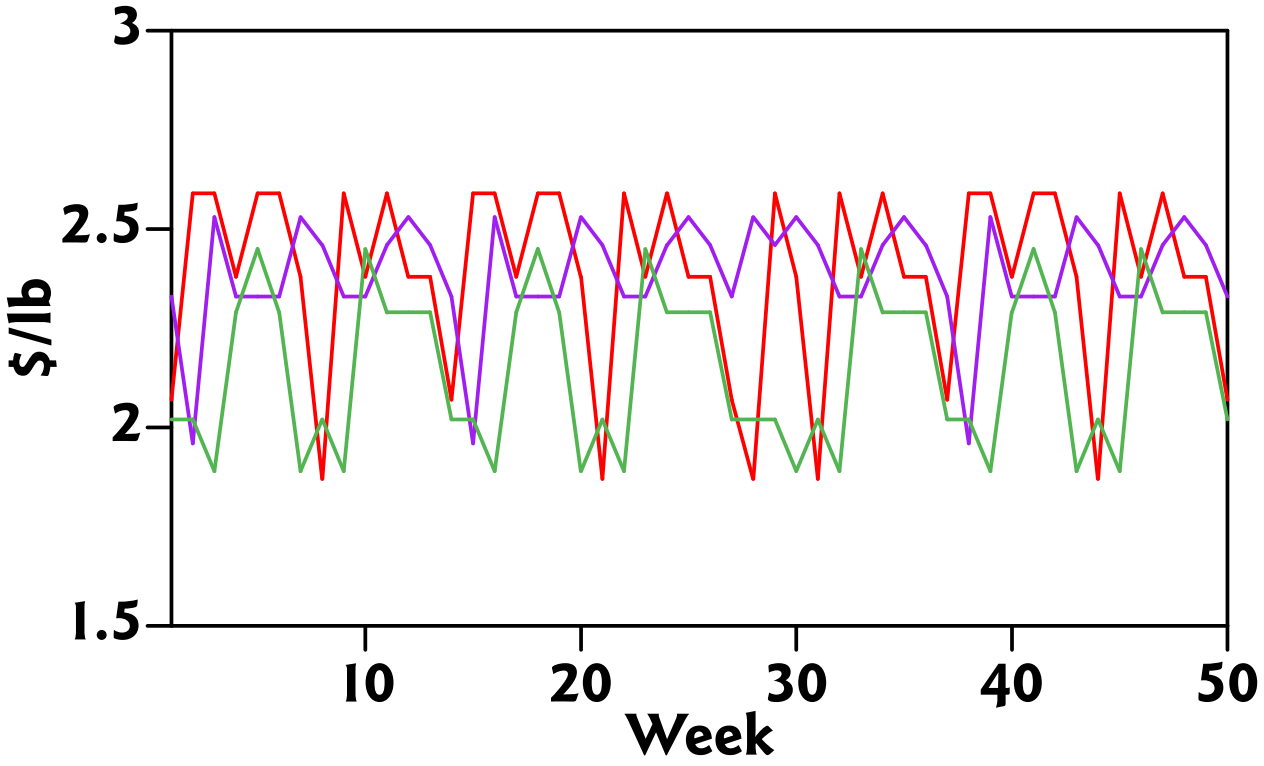


Figure 3: **Example of a Simulated Oligopoly (Marks et al. 1995)**

Three brands, one-week memory, 50 weeks observed

Table 1 shows the observed distribution of states in the historical Chain 1, and in the three models: Run 11, Run 26a, and Run 26b.

State	Chain 1	Run 11	Run 26a	Run 26b
000	32	0	30	20
001	2	18	11	10
010	6	15	3	7
011	1	0	0	0
100	7	16	5	12
101	0	0	0	0
110	2	0	1	1
111	0	1	0	0
Total	50	50	50	50

So: how close are the three models to reality (Chain 1)?

Table 1: The observed frequencies of the 8 states over 50 weeks.

4. Three Measures of the Distance Between Sets of Time Series 4.a. Kullback-Leibler (K-L) Information Loss

Based on Shannon (1948) entropy:

$$SE(p(x) \mid x \in X) = -\sum p(x) \log_2(p(x))$$

The K-L information loss provides a measure of the information lost when model g is used to approximate full reality f:

$$I(f, g) = \sum_{i=1}^{k} p_i \times \log_2\left(\frac{p_i}{\pi_i}\right) \qquad (1)$$

with full-reality f distribution $0 < p_i < 1$, and model g distribution $0 < \pi_i < 1$, with $\sum p_i = \sum \pi_i = 1$.

Two shortcomings: 1. in our data often $\pi_i \times p_i = 0$ because one or both is zero — both must be positive for K-L.

2. the K-L measure is not a true metric: it is not symmetrical and doesn't satisfy the triangle inequality.

4.b. The State Similarity Measure (SSM)

The SSM method reduces the dimensionality of the historical behaviour (and sometimes the model output too) by partitioning the price line in order to derive a measure of similarity or distance beween two sets.

Calculating the SSM:

1. **For each set, partition the time-series $\{P_{bw}\}$ of price P_{bw} of brand b in week w into {0,1}, where 0 corresponds to "high" price and 1 corresponds to "low" price to obtain time-series $\{P_{bw}'\}$;**
2. **For the set of 3- or 4-brand time-series of brands' partitioned prices $\{P_{bw}'\}$, calculate the time-series of the state of the market each week $\{S_w\}$;**
3. **For each set, calculate the time-series of the state of the 3- or 4-week moving window of partitioned prices $\{M_w\}$, from the per-week states $\{S_w\}$;**

4. **Count the numbers of each state observed for the set of time-series over the given time period; convey this by an $n \times 1$ vector c, where $c[s]$ = the number of observations of window state s over the period;**

5. **Subtract the number of observations in set A of time-series from the number observed in set B, across all n possible states; $d^{AB} = c^A - c^B$;**

6. **Sum the absolute values of the differences across all possible states:**

$$D_{ad}^{AB} = \mathbf{1}' \times |d^{AB}| \qquad (2)$$

This number D_{ad}^{AB} is *the distance between two time-series sets A and B.*

SSM for three memory depths.

We have now (re)calculated the six pairs of SSMs between the three runs and the historical data (from Chain 1), using 50-week data series: Table 2.

Pair	1-week memory	2-week memory	3-week memory
Chain 1, Run 11	70	88	92
Chain 1, Run 26a	18	36	54
Chain 1, Run 26b	28	48	68
Run 11, Run 26a	62	76	88
Run 11, Run 26b	42	60	80
Run 26a, Run 26b	22	42	60

Remember: an SSM of zero means that the two sets of time series are identical; larger SSMs imply less similarity. The maximum SSM occurs when the intersection between the states of the two sets of time series is null: here, this would

be seen with an SSM of 100 (given that there are 50 observations per set of time series).

As the partitioning becomes finer (with deeper memory of past actions), the SSMs increase as the two sets of time series become less similar. This should not surprise us. We also note that with these four sets of time series, the rankings do not change with the depth of memory: (from closer to more distant) (Chain 1, Run 26a), (Run 26a, Run 26b), (Chain 1, Run 26b), (Run 11, Run 26b), (Run 11, Run 26a), and (Chain 1, Run 11).

Asking which of the three models is closest to the historical data of Chain 1, the SSM tells us that Run 26a is best, followed by Run 26b, with Run 11 bringing up the rear.

We might ask whether measures from GHM (Klir 2006) have the same ranking.

4.c. The Generalized Hartley Measure (GHM)

Ralph Hartley (1928) showed that the only meaningful way to measure the amount of uncertainty associated with a finite set E of possibilities from the larger (finite) set X is to use a functional of the form $c \log_b \sum_{x \in X} |E|$, where $b \neq 1$. Specifically,

$$H(r_E) = \log_2 |E| = \log_2 \sum_{x \in X} r_E(x)$$

for measurement of H in bits, where the basic possibility function $r_E \in \{0, 1\}$.

Notes: (1) $0 \leq H(E) \leq \log_2 |X|$, for any $E \in$ the power set $P(X)$.

(2) If a given set of possible alternatives, E, is reduced by the outcome of an action to a smaller set $E' \subset E$, then the amount of information $I_{(A:E \to E')}$ generated by the action $A: E \to E'$ is measure by the difference $H(E) - H(E')$.

The Generalized Hartley Measure

Relax the "either/or" restriction on r_E: allow $r : X \to [0, 1]$.

Klir (2006) provides proofs and properties of the GHM, and notes that "possibility theory is based on *similarity*".

Start with $X = \{x_1, x_2, \cdots, x_n\}$, where r_i denotes for each $i \in N_n$ the *possibility* of x_i.

Sort the elements of X so that the possibility profile $\mathbf{r} = < r_1, r_2, \cdots, r_n >$ is ordered so that $1 = r_1 \geq r_2 \geq \cdots \geq r_n$, and $r_{n+1} = 0$ by convention.

Then the GHM is given by:

$$GHM(\mathbf{r}) = \sum_{i=2}^{n} (r_i - r_{i+1}) \log_2 i = \sum_{i=2}^{n} r_i \log_2 \left(\frac{i}{i-1} \right) \qquad (3)$$

The Results using the GHM

From Table 1 (one-week memory), we can reorder the possibilities (observed frequencies) of the three runs and the historical data, to get the four reordered possibility profiles:

$$\mathbf{r} = < r_1, r_2, \cdots, r_n >$$

Chain 1:	32	7	6	2	2	1	0	0
Run 11:	18	16	15	1	0	0	0	0
Run 26a:	30	11	5	3	1	0	0	0
Run 26b:	20	12	10	7	1	0	0	0

Table 3: Four possibility profiles (non-normalized *)

The GHMs for the three runs and Chain 1 have been calculated for the three memories of 1 week, 2 week, and 3 week.

(* Normalization here means $r_1 = 1$, not $\sum r_i = 1$.)

GHMs for History (Chain 1) and 3 Models (Table 4)

Process	1-week memory	2-week memory	3-week memory
History (Chain 1)	0.386	0.495	0.782
Run 11	1.399	2.179	2.787
Run 26a	0.516	0.679	1.085
Run 26b	1.054	1.657	2.542

These are true metrics (they satisfy the triangle inequality, unlike the K-L information loss), and so we can compare the differences between the four measures.

We readily see that Run 26a (0.516) is closest to the historical data of Chain 1 (0.386); next is Run 26b (1.054), with Run 11 (1.399) furthest from the historical data.

Moreover, we see that Run 26a is closer to the historical Chain 1 data than it is to Run 26b.

Results using the SSM

Having derived the distance between two sets of time-series using the *State Similarity Measure*, 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).

SSM Distances Between Historical Chain 1 and Three Models

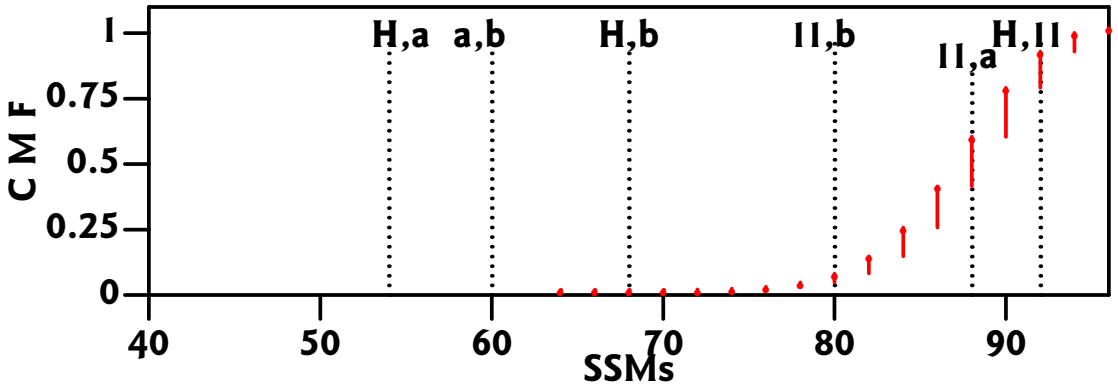
	Chain 1	Run 11	Run 26a	Run 26b
Chain 1	0	92*	54	68
Run 11	92*	0	88*	80*
Run 26a	54	88*	0	60
Run 26b	68	80*	60	0

Table 5: Distances Between Chain 1 and Three Models (with 3 Brands, 3-week memory)
(* : cannot reject the null at the 5% level)

Here, S, the maximum number of states = 48, so the maximum distance apart is 96. We see that Run 26a is closest to historical Chain 1, closer than it is to Run 26b; we also see that Run 11 is very distant from Chain 1, possibly into randomness.

So this SSM result accords with the GHM: Run 26a is closest.

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 Chain 1 and Run 11; reject the null (random) hypothesis for all other pairs.

Conclusions — the SSM and the GHM

The 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. So is the GHM.

Such metrics are 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.

It remains to explore the theory of SSM as others have explored GHM.

[1] Akaike H. (1973), "Information theory as an extension of the maximum likelihood principle," in B.N. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267–281.

[2] Burnham K.P. and Anderson D.R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd. ed., New York: Springe

[3] Fisher R.A. (1922), "On the mathematical foundations of theoretical statistics." Royal Society of London. *Philosophical Transactions* (Series A) 222: 309–368.

[4] Hartley, R. V. L. (1928), "Transmission of information." *The Bell System Technical Journal*, 7(3): 535–563.

[5] Klir G. J. (2006), *Uncertainty and Information: Foundations of Generalized Information Theory,* New York: Wiley.

[6] Kullback J.L. and Leibler R.A. (1951), "On information and sufficiency," *Annals of Mathematical Statistics*, 22: 79–86.

[7] R.E. Marks (1998) Evolved perception and behaviour in oligopolies, *Journal of Economic Dynamics and Control*, 22(8–9): 1209–1233, July.

[8] Marks R.E. (2007), "Validating simulation models: a general framework and four applied examples," *Computational Economics,* 30(3): 265–290, October

[9] Marks R.E. (2010), "Comparing Two Sets of Time-Series: The State Similarity Measure," In *2010 Joint Statistical Meetings Proceedings – Statistics: A Key to Innovation in a Data-centric World*, Statistical Computing Section. Alexandria, VA: American Statistical Association, pp. 539–551.
http://www.agsm.edu.au/bobm/papers/asap.pdf

[10] R.E. Marks, D.F. Midgley, and L.G. Cooper (1995) Adaptive behavior in an oligopoly, *Evolutionary Algorithms in Management Applications*, ed. by J. Biethahn and V. Nissen, (Berlin: Springer-Verlag), pp. 225–239.
http://www.agsm.edu.au/bobm/papers/marks-midgley-cooper-95.pdf

[11] D.F. Midgley, R.E. Marks, and L.G. Cooper (1997), Breeding competitive strategies, *Management Science*, 43: 257–275.

[12] Ramer, A. (1987), "Uniqueness of information measure in the theory of evidence," *Fuzzy Sets and Systems,* 24(2): 183–196.

[13] Rényi, A. (1970), *Probability Theory,* Amsterdam: North-Holland (Chapter 9, "Introduction to information theory," pp. 540–616).