

Validation and Model Selection

**Robert Marks,
Economics, UNSW, Sydney, and
the University of Melbourne
robert.marks@gmail.com**

Outline

1. **Validation and Model Selection**
2. **One Issue: Heterogeneous Agents, Sets of Time-series of Prices, Defining the States of the Market**
3. **A Measure of the Distance Between Sets of Time-Series:
*The State Similarity Measure (SSM)***
4. **The Results**
5. **Conclusions**

(See Marks R.E., "Validation and Model Selection: Three Similarity Measures Compared," *Complexity Economics*, 2013, forthcoming.)

Two Kinds of ABM

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, a life models, etc

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.

A new technique: SSM

We outline a new technique, the State Similarity Measure, for tackling the fourth core issue of Fagiolo et al. (2007): validating agent-based models using historical data.

Compared to their three methods (indirect calibration, the Werker-Brenner approach, and the history-friendly approach), SMM focusses on:

- the micro-level output**
- an empirical comparison of model output v. history**
- leading to a choice of model that best fits, to resolve any identification (or under-determination) issues.**

The SSM includes here a Monte Carlo simulation to eliminate random observations.

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 I

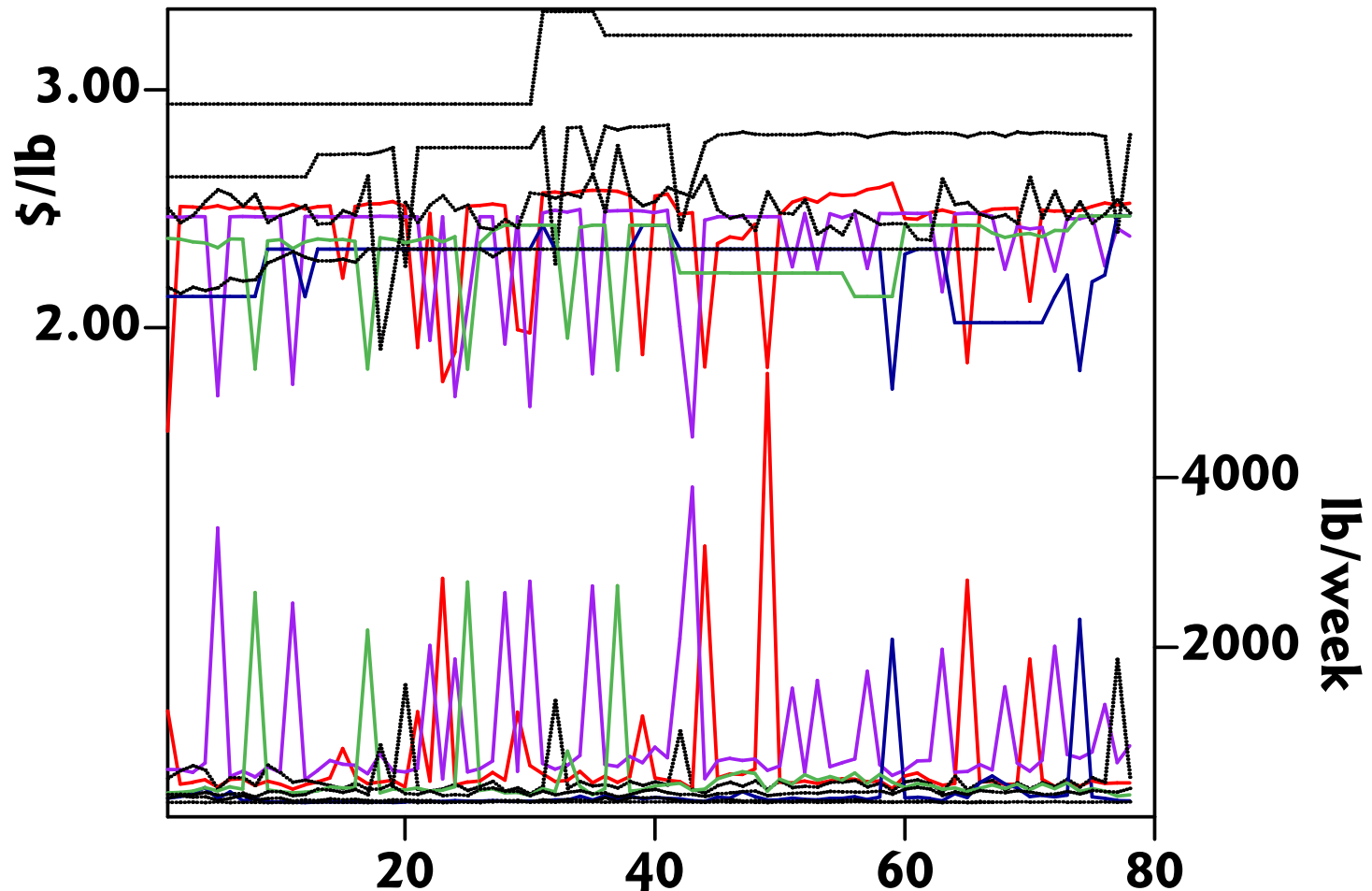


Figure 1: Weekly Sales and Prices (Source: Midgley et al. 1997)

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 strategic 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 History

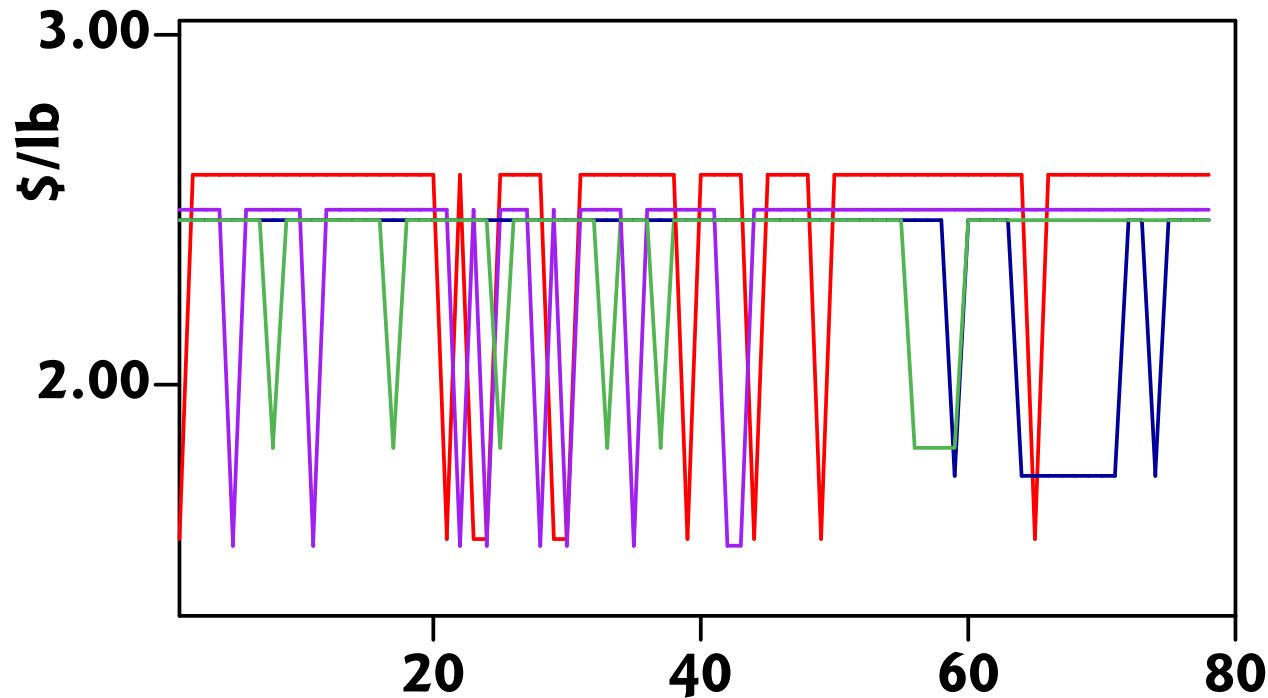


Figure 2: Partitioned Weekly Prices of the Four Chain-One Brands

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.

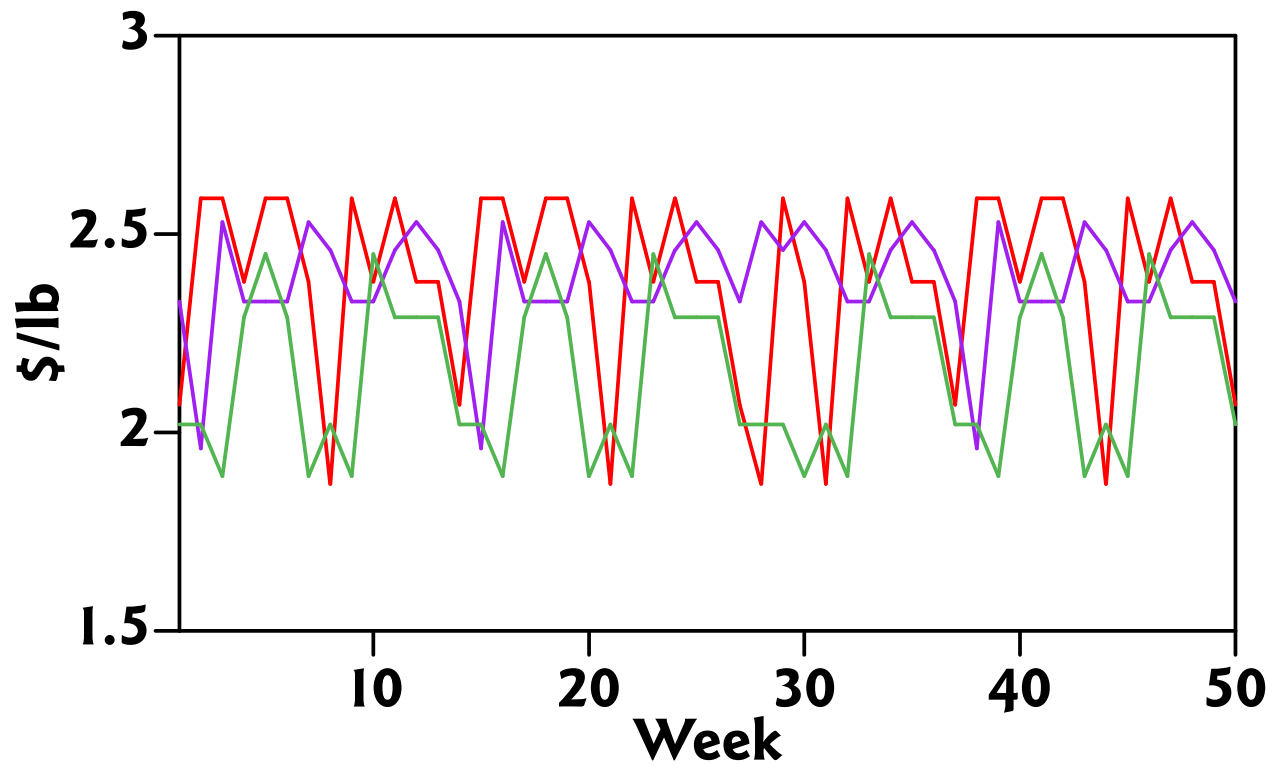


Figure 3: Example of a Simulated Oligopoly (Marks et al. 1995)

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: 11, 26a, and 26b.

State	History	Model 11	Model 26a	Model 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 History?

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

4. 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 between two sets.

Calculating the SSM:

1. Calculate the weekly states of the market: 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 $\{S_w\}$, where

$$S_w = P'_{1,w} \times P'_{2,w} \cdots;$$

For a 3-brand time series,

$$S_w = 4 \times P'_{1,w} + 2 \times P'_{2,w} + P'_{3,w}. \text{ Then construct the}$$

windowed states of the market. (Table 2.)

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 $\{S_w\}$, where $M_w = S_w \times S_{w-1} \times S_{w-2} \cdots$. For a 3-week window, $M_{3w} = 64 \times S_w + 8 \times S_{w-1} + S_{w-2}$. (The powers of 8 are because, with three brands, there are 8 possible states of the market S_w each week.) For a 3-week memory, there are $8^3 = 512$ possible states.**
4. **Count the numbers of each state M_w observed for the set of time-series over the given time period; convey this by an $n \times 1$ vector p , where $p_s \geq 0$ is the number of observations of window state s over the period; With T longitudinal observations the maximum SSM distance apart of two sets of time series is $2 \times (T - w + 1)$, where w is the number of weeks remembered. (This would happen when the two sets are disjoint.)**

Week	B r a n d ($P'_{b,w}$)			1-Week $\therefore S_w$	2-Week $\therefore M_{2w}$	3-Week $\therefore M_{3w}$
	Red	Purple	Green			
18	0	0	0	0		
19	0	0	0	0	0	
20	0	0	0	0	0	0
21	1	0	0	4	32	256
22	0	1	0	2	20	160
23	1	0	0	4	34	276
24	1	1	0	6	52	418
25	0	0	1	1	14	116
26	0	0	0	0	1	14
27	0	0	0	0	0	1
28	0	1	0	2	16	128
29	1	0	0	4	34	272
30	1	1	0	6	52	418

Table 2: An example: three brands, 1-, 2-, and 3-week windows

5. Subtract the number of observations in set P of time-series from the number observed in set Q, across all n possible states; $D^{PQ} = p - q$;

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

$$d_1^{PQ} = \sum |p_i - q_i|. \quad (2)$$

This number d_1^{AB} is the distance between two time-series sets P and Q , the SSM.

SSM for three memory depths.

We have now calculated the six pairs of SSMs between the three models and the Historical data (from History), using 50-week data series: Table 3.

Pair	1-week memory	2-week memory	3-week memory
History, Model 11	70	88	92
History, Model 26a	18	36	54
History, Model 26b	28	48	68
Model 11, Model 26a	62	76	88
Model 11, Model 26b	42	60	80
Model 26a, Model 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) (History, Model 26a), (Model 26a, Model 26b), (History, Model 26b), (Model 11, Model 26b), (Model 11, Model 26a), and (History, Model 11).

Asking which of the three models is closest to the Historical data of History, the SSM tells us that Model 26a is best, followed by Model 26b, with Model 11 bringing up the rear.

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. (In the Figure, * : cannot reject the null at the 5% level.)

SSM Distances Between Historical Chain I and Three Models

	History	Model 11	Model 26a	Model 26b
History	0	92*	54	68
Model 11	92*	0	88*	80*
Model 26a	54	88*	0	60
Model 26b	68	80*	60	0

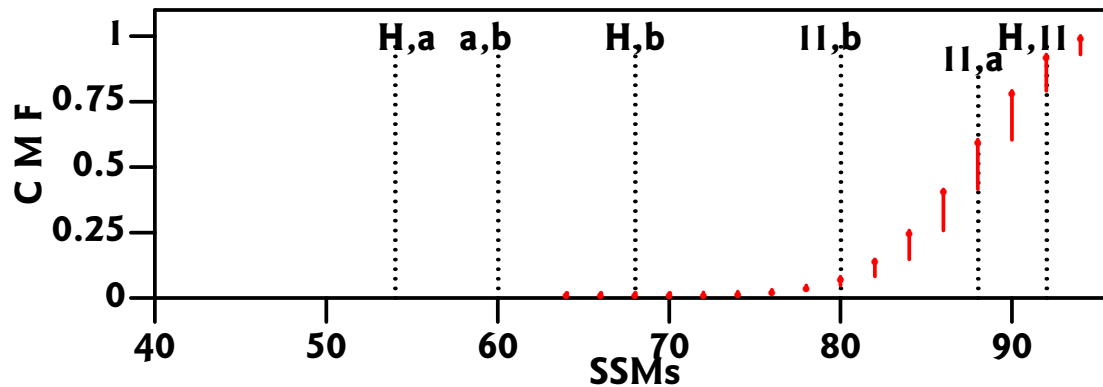
Table 5: Distances Between History and Three Models (with 3 Brands, 3-week memory)

Here, S , the maximum number of states = 48, so the maximum distance apart is 96.

We see that Model 26a is closest to Historical Chain I, closer than it is to Model 26b; we also see that Model 11 is very distant from History, possibly into randomness.

Null Hypothesis: each of the two sets of time series is random.

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 History and Model II; reject the null (random) hypothesis for all other pairs.

Conclusions – the SSM

The SSM is a true metric sufficient to allow us measure the degree of similarity between two sets of time-series which embody dynamic responses.

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 allow us to identify which of several models' outputs is closest to history, to determine which model has captured most information of the historical time series.

Validation can be used to select the best simulation model.

References

- [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: Springer
 - [3] Fagiolo G., Moneta A., and 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.
 - [4] Fisher R.A. (1922), "On the mathematical foundations of theoretical statistics." *Royal Society of London. Philosophical Transactions (Series A)* 222: 309–368.
 - [5] Hartley, R. V. L. (1928), "Transmission of information." *The Bell System Technical Journal*, 7(3): 535–563.
 - [6] Klir G. J. (2006), *Uncertainty and Information: Foundations of Generalized Information Theory*, New York: Wiley.
 - [7] Krause E. F. (1986), *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*, New York: Dover.
 - [8] Kullback J.L. and Leibler R.A. (1951), "On information and sufficiency," *Annals of Mathematical Statistics*, 22: 79–86.
 - [9] R.E. Marks (1998) Evolved perception and behaviour in oligopolies, *Journal of Economic Dynamics and Control*, 22(8–9): 1209–1233, July.
 - [10] Marks R.E. (2007), "Validating simulation models: a general framework and four applied examples," *Computational Economics*, 30(3): 265–290, October
 - [11] Marks R.E., "Validation and Model Selection: Three Similarity Measures Compared," *Complexity Economics*, 2013, forthcoming.
<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] 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>
- [14] D.F. Midgley, R.E. Marks, and L.G. Cooper (1997), Breeding competitive strategies, *Management Science*, 43: 257–275.
- [15] Ramer, A. (1987), "Uniqueness of information measure in the theory of evidence," *Fuzzy Sets and Systems*, 24(2): 183–196.
- [16] Rényi, A. (1970), *Probability Theory*, Amsterdam: North-Holland (Chapter 9, "Introduction to information theory," pp. 540–616).
- [17] Schelling, T.C. (1971), Dynamic models of segregation, *Journal of Mathematical Psychology*, 1: 143–186.
- [18] Shannon, C.E. (1948). "A mathematical theory of communication," *Bell System Technical Journal*, 27: 379–423, 623–656, July, October.

4. Three Measures of the Distance Between Sets of Time Series

4.a. Kullback-Leibler (K-L) Information Loss

Based on Shannon (1948) entropy:

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

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

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

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.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 $I_{(A:E \rightarrow E')}$ generated by the action $A: E \rightarrow E'$ is measure by the difference $H(E) - H(E')$.

The Generalized Hartley Measure

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

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

Start with $X = \{x_1, x_2, \dots, x_n\}$, 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 $r = \langle r_1, r_2, \dots, r_n \rangle$ is ordered so that $1 = r_1 \geq r_2 \geq \dots \geq r_n$, and $r_{n+1} = 0$ by convention.

Then the GHM is given by:

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

The Results using the GHM

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

$$\mathbf{r} = \langle r_1, r_2, \dots, r_n \rangle$$

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

Table 4: Four possibility profiles (non-normalized *)

The GHMs for the three models and History 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 5)

Process	1-week memory	2-week memory	3-week memory
History (Chain 1)	0.386	0.495	0.782
Model 11	1.399	2.179	2.787
Model 26a	0.516	0.679	1.085
Model 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 Model 26a (0.516) is closest to the Historical data of Chain 1 (0.386); next is Model 26b (1.054), with Model 11 (1.399) furthest from the historical data.

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