Comparing the Observed and Model-generated Behavior Patterns to Validate System Dynamics Models

Yaman Barlas
Georgia Institute of Technology, Atlanta

ABSTRACT

Model behavior evaluation is an important component of System Dynamics (SD) model validation. SD methodology has often been criticized for its lack of quantitative/formal behavior evaluation tools. System Dynamicists have responded by stating the relative, subjective, qualitative nature of model validation. We argue that using formal quantitative behavior tools is not inconsistent with a relativist, holistic philosophy of model validation. We suggest a multi-step, quantitative behavior evaluation procedure which focuses on individual pattern components of a composite behavior pattern. The procedure is relatively easy to apply and to interpret. We then test the performance of the procedure through a series of simulation experiments. The experimental results suggest that the multi-step procedure is appropriate for SD model behavior evaluation. The experiments also give us an idea of what the expected values and the variations of the suggested quantitative tools are.

I- INTRODUCTION

Comparing the observed and model-generated behavior patterns (which we call 'behavior evaluation') is an important component of SD model validation. In other methodologies (such as econometrics, autoregressive modeling, discrete-event simulation), behavior evaluation is a quite formal and quantitative process whereas behavior evaluation in SD is in general quite informal and qualitative. Thus, SD has been criticized for its lack of quantitative, formal behavior evaluation tools (see for example Ansoff and Slevin 1968). Sterman (1984) notes that the reluctance to use formal quantitative tools creates "an impression of sloppiness and unprofessionalism". System Dynamicists have usually responded to such criticisms by arguing that model validation is inherently relative and judgemental, hence largely informal and nonquantitative (see for instance Forrester 1968). The relativist and holistic (as opposed to absolutist and reductionist) validation philosophy of SD is well established (see Forrester 1961 and 1968 Forrester and Senge 1980). According to this philosophy -shared by most System Dynamicists including this author- model validation is ultimately judging the usefulness of a model. Thus, validity is always relative to a purpose because usefulness can not be established without specifying a purpose. Absolute validity is theoretically and practically impossible. Validity can not be proven, but it can be agreed upon. Validation is inevitably relative and informal because validity is situation-dependent. The analyst (user) must usually chose between alternative models, some of which are informal, mental models. Validation is ultimately a matter of social conversation, judgement and agreement. But this relativist/holistic philosophy of validation should not lead to denying the role of formal quantitative tools in behavior evaluation. Making full use of various formal quantitative tools is not inconsistent with that philosophy. To say that the overall model validation is inevitably relative, informal and judgemental is not to say that formal quantitative tools are useless in behavior evaluation. On the contrary, formal quantitative tools are most

useful when they are used with the relativist philosophy. Accordingly, formal tools can not turn the overall validation problem into a purely formal, objective process. But these tools are very useful and effective ways of organizing, summarizing and communicating information. The relativist philosophy should provide the proper perspective: The outcome of a formal quantitative procedure, by itself, can not determine whether a model is valid or not. But it can provide valuable information in judging and then communicating the usefulness of a model. To sum up, we need quantitative formal methods of behavior evaluation for the same reason why we need quantitative formal models of social systems.

In the SD literature, R.D. Wright (1972) was first to evaluate the applicability of several quantitative techniques to model validation. Peter M. Senge (1977) used simulation experiments to evaluate the accuracy of least squares estimation techniques as applied to SD models. Although his work focused on 'parameter estimation', his results have important implications for behavior evaluation. D.W. Peterson (1980) developed an estimation technique based on "Kalman filtering", which again has implications for behavior evaluation. Finally, Sterman (1984) recommended the use of "Theil's inequality proportions" in SD behavior evaluation.

SD models attempt to reproduce and predict broad patterns, rather than the individual data points. The major pattern components are: trends, periods of cycles, means, amplitude variations and phase angles. We suggest a set quantitative tools to measure and compare the five pattern components of the observed and model-generated behavior patterns. Then, we test the performances of these tools through a series of simulation experiments.

II- THE RESEARCH PROCEDURE

We develop a multi-step quantitative procedure to compare the observed and model-generated behavior patterns. To test the performance of this procedure we take a 'synthetic' experimental approach similar to Senge's (1977). We build a model with structure R and call it the 'synthetic real system'. R generates the performance patterns P . We then build a model of R and call it the 'model' M. M generates its performance patterns P . Since we have defined the synthetic system, we have a perfect control over its structure and its parameters. This makes it possible to investigate under what conditions, which quantitative tools are more useful and reliable. In this paper we assume that we have an idealized model with 'perfect' structure so that M=R, except that the exact noise sequences in the two are different (accounting for certain factors which are imposible to estimate parfectly). Our purpose is to see what type and degree of accuracy can be expected from SD models in the limit (M=R). The other purpose is to test the performances of the suggested quantitative measures in this limiting case and to examine the effects of non-structural errors (input error, parameter error, observation error) on these measures. (In the second phase of the research, we introduce various structural errors so that $M \neq R$, and explore ways of detecting structurally inadequate models which may exhibit 'false' behavior accuracy. The results of this second phase will be presented in another paper).

III-THE QUANTITATIVE TOOLS

Statistics and Time-series literature offers a large variety of quantitative

measures. In selecting a set of quantitative tools appropriate for SD behavior evaluation, two criteria are most important: First, the selected tools must be pattern oriented rather than point oriented. SD models are built to reproduce and predict broad patterns observed in real systems. Such models are not designed to reproduce the observed behavior on a point-by-point basis. Secondly, the selected tools must be relatively easy to implement and interpret. It is unlikely that an extremely expensive and complicated tool will be widely used by the practitioners. Taking these two criteria into consideration, we select and test the following quantitative tools:

1- Trend Comparison and Removal. If there is an indication of a trend component, the latter can be estimated by fitting a regression line of the form $\hat{y}=\hat{b}_o+\hat{b}_1 t$ to the observed and model generated behavior patterns. One can compare the trend components by comparing the corresponding coefficients $\hat{b}$ of the two regression equations. In this experimental work, the trend component is not strongly exponential so that the regression equation takes the simple linear form of $\hat{y}=\hat{b}_o+\hat{b}_1 t$. The slope coefficient $\hat{b}_1$ yields an estimate of the trend involved. (One can test the equality of the coefficients $\hat{b}_1$ in a rigorous way by making use of their standard errors. But to be accurate, such a hypothesis requires that the observations y_i are independent, an assumption bound to be violated by virtually any SD behavior pattern). Almost all summary statistics require at least stationarity in the means. Therefore, after deciding that there is no significant error in the trend components, the latter must be removed from the observed and model-generated responses by using $z_i=y_i-\hat{b}_1 t$.

2- Comparing the Periods. To compare the periods involved, we design a procedure based on the 'sample autocorrelation functions'. The sample 'autocovariance function' of X_i is defined as:

$$Cov(k)=\frac{1}{N}\sum_{i=1}^{N-k}(X_i-\bar{X})(X_{i+k}-\bar{X}) \quad , \text{ for lag } k=0,1,2,3...$$

Then, the sample 'autocorrelation function' is obtained by dividing Cov(k) by Cov(0) (which is the variance of X):

$$r(k)=\frac{Cov(k)}{Cov(0)}=\frac{Cov(k)}{Var(X)} \quad , \text{ for } k=0,1,2,3,...$$

This function gives a measure of how the successive observations are interrelated. We are interested in plotting r(k) only for positive lags 1,2,3... because r(k)=r(-k). The function starts at 1 at lag 0 and lies always between +1 and -1. As k becomes larger, r(k) approaches 0. For random (or 'almost random') series, r(k) quickly drops to 0 for $k\geq 1$. For moderately correlated series, r(k) approaches 0 in a negative exponential fashion. If the series involves trend, r(k) dies down very slowly; if the series is cyclic, r(k) is also cyclic.

To use the autocorrelation function in behavior comparison, we must compute the autocorrelation functions of the observed and the model-generated series and then compare the two autocorrelation functions. This comparison is compelling because it is a way of comparing the patterns rather then the individual data points. To be able to carry out a formal test, we need the distribution of r(k) which is unknown except for some extremely restricted cases. As an approximate test, we can compute the standard errors (se) of r(k) and construct a 2se confidence band. A number of variance estimates are available for r(k).

Our theoretical and empirical analysis of the available variance formula suggested that the best one was the finite-sample approximation provided by O.D. Anderson (1982):

$$\mathrm{Var}(r(k)) = \frac{1}{N(N+2)} \sum_{i=1}^{N-1} (N-i)\,(r(k-1)+r(k+i)-2r(k)r(i))^2$$

Now we can devise an approximate test of hypothesis: If $r_S(k)$ is estimated from the simulated response and $r_A(k)$ from the actual one, then the null hypothesis is:

H_o: $r_S(1)=r_A(1)$, $r_S(2)=r_A(2)$, ... $r_S(m)=r_A(m)$

and H_1: $r_S(k)\neq r_A(k)$ for at least one k.

Consider the difference $d_1 = r_S(1)-r_A(1)$. The standard error of d_1 is:

$$s_{d_1} = \sqrt{\mathrm{Var}(r_S(1)) + \mathrm{Var}(r_A(1))}$$

Since under H_o $d_1=0$, to test H_o, we construct the interval $\{-2s_{d_1}, +2s_{d_1}\}$, compute d_1 and reject H_o if d_1 falls outside the interval. We repeat the same procedure for every d_k, k=1,2,...m. Under the normality assumption, this constitutes a test of hypothesis at about $\alpha=0.05$. A major source of approximation is due to the fact that H_o is a multiple test and succesive r(k)'s are not independent random variables. For even slightly autocorrelated data, the autocorrelation estimates are highly correlated statistics. Hence, the computation of the 'actual significance level' of this multiple hypothesis -which requires knowledge of the covariance matrix of r(k)- represents an exceedingly complex problem in practice. But we believe that other errors inherent in any SD validity testing are of much larger significance than the imprecision of α. We therefore do not seek to improve further the precision of this test.

3- Comparing the Means. The sample mean is given by $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. To compare the means of the simulated (S) and actual (A) behavior patterns, we define the 'percent error in the means' E1:

$$E1 = \frac{|\bar{X}_S - \bar{X}_A|}{|\bar{X}_A|}$$

4- Comparing the Amplitude Variations. We use the standard deviations and the corresponding 'percent error in the variations' E2:

$$s = \sqrt{\frac{1}{N}\sum (X_i - \bar{X})^2} \quad \text{and} \quad E2 = \frac{|s_S - s_A|}{s_A}$$

In this step, it may also be necessary to compare graphical measures of the amplitudes of the cycles involved in the two time histories.

5- Testing for Phase Lag. We suggest using the crosscorrelation function which shows how two series are correlated at different time lags k. The crosscorrelation function between the series S and A is given by:

$$C_{SA}(k) = \frac{\frac{1}{N}\sum_{i=k}^{N} (S_i - \bar{S})\,(A_{i-k} - \bar{A})}{s_S\, s_A} \qquad \text{for } k=0,1,2,3,...$$

and

$$C_{SA}(k) = \frac{\frac{1}{N}\sum_{i=-k}^{N} (A_i - \bar{A})\,(S_{i+k} - \bar{S})}{s_S\, s_A} \qquad \text{for } k=0,-1,-2,-3,...$$

The crosscorrelation function is not necessarily symmetrical. But note that

$C_{SA}(k)=C_{AS}(-k)$. Thus, the function is fully known when $C_{SA}(k)$ and $C_{AS}(k)$ are known or when only one of them is known for both positive and negative lags. Note also that $C_{SA}(0)=C_{AS}(0)$ which is the simple correlation coefficient between the two series. The crosscorrelation function provides a measure of the phase relationship between two series.

6- A 'Discrepancy coefficient' U. Henry Theil (1958) proposed the 'inequality coefficient' to evaluate the accuracy of forecasts:

$$U_o = \frac{\sqrt{\sum(S_i - A_i)^2}}{\sqrt{\sum S_i^2} + \sqrt{\sum A_i^2}}$$

The disadvantage of this coefficient, as noted by Theil himself, is that its value is not uniquely determined by the 'sum of squared errors' (SSE). If we have two different forecasts resulting exactly in the same SSE, the forecast with larger S_i would yield smaller U_o because this latter is discounted by S_i. It is even possible to have two sets of forecasts such that $SSE_1 > SSE_2$ but $U_1 < U_2$. Thus, overestimating becomes a 'safe way' of obtaining smaller U values. As a way of eliminating this problem, Griner et al (1978) suggested to center the data by subtracting from S_i and A_i their respective means. This yields:

$$U = \frac{\sqrt{\sum(S_i - \bar{S} - A_i + \bar{A})^2}}{\sqrt{\sum(A_i - \bar{A})^2} + \sqrt{\sum(S_i - \bar{S})^2}} = \frac{\sqrt{\sum(E_i - \bar{E})^2}}{\sqrt{\sum(A_i - \bar{A})^2} + \sqrt{\sum(S_i - \bar{S})^2}} = \frac{s_E}{s_A + s_S}$$

We adapt this 'discrepancy coefficient' which is insensitive to the additive constants and preserves its nice property of being betwen 0 and 1. But note that U does not anymore reflect errors in the means.

The quantitative tools described in this section, taken together, form a multi-step behavior evaluation procedure focusing on individual pattern components. These pattern components are : trends, periods of cycles, means, amplitude variations and phase relations. Through a set of simulation experiments, we attempt to assess the performances of these quantitative tools.

IV- THE SYNTHETIC SYSTEM

As the 'synthetic reality', we specify a set of relationships believed to be a realistic description of certain epidemic dynamics. The system can be verbally described as follows: The total population consists of a healthy population and a sick population. The healthy population has two subgroups: susceptibles and immunes. When a susceptible person contacts a contagious person, he is infected with some probability. After a person is infected, he becomes contagious for some period of time during which he can transmit disease. At the end of this period, he starts showing symptoms and is recognized as contagious. From this time on, he is isolated (or people avoid contacting him) until he recovers or dies. Once he recovers, he becomes immune for a period of time, at the end of which he completes the cycle by joining the susceptible population. The other source of replenishment of susceptibles is the population growth. Every sub-population contributes to the total conception rate except the 'contagious population recognizable' which is assumed to exercise some sort of contraception. When infants are born, they are immune for a certain period before they become susceptible. The flow diagram of the system is shown in figure 1. We see that the system has a number of negative and positive loops coupled together. A most significant loop is the 'contagious population - contact rate - infection rate - population infected -

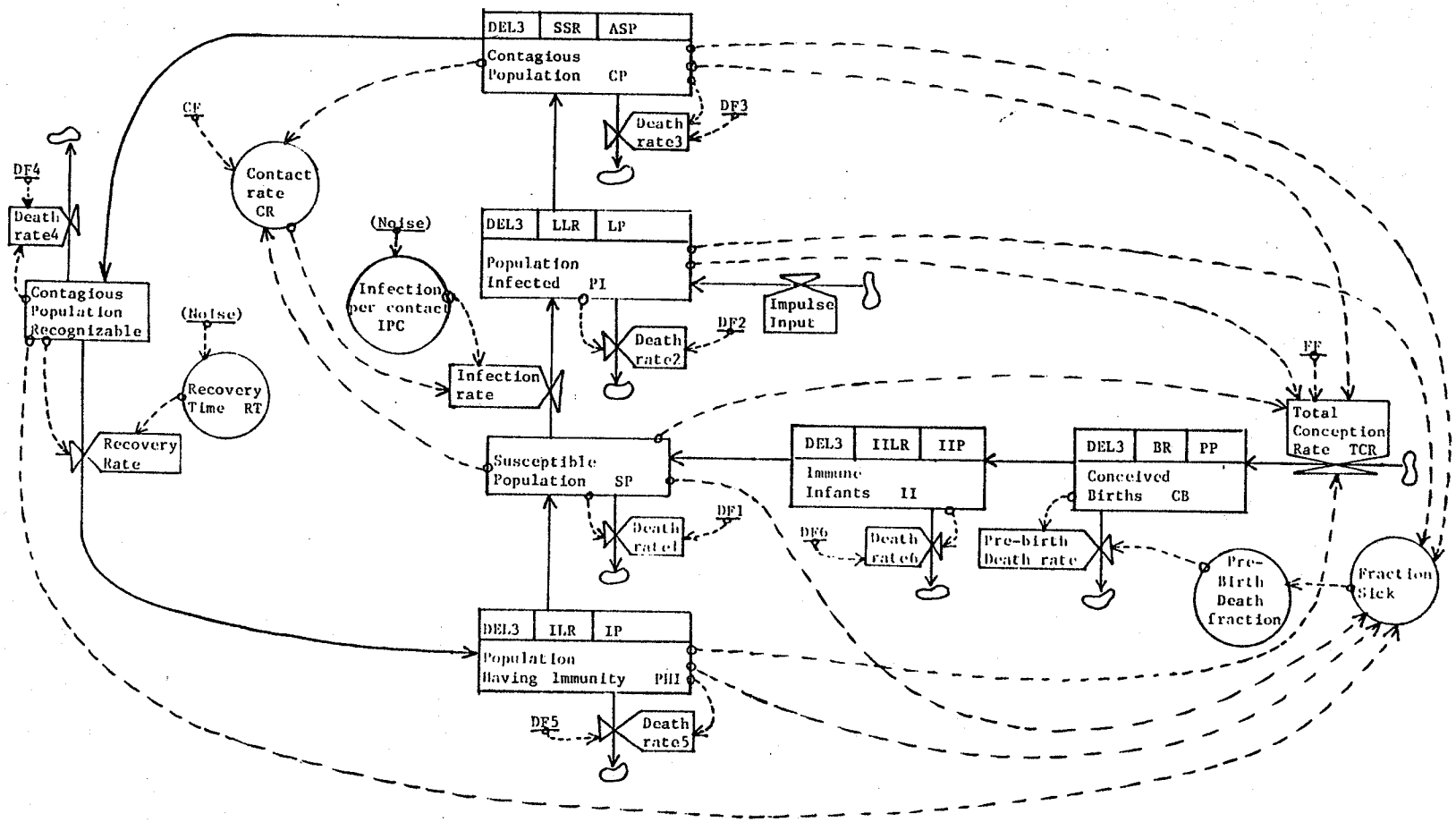


Figure 1. Flow Diagram for the Synthetic System.

latency leaving rate' loop. This positive loop says that the epidemic spreads by contact, that the larger the contagious population, the greater is the number of infective contacts. Thus, the contagious population reinforces its own growth and would exhibit indefinite growth in the absence of other influences. But there are other influences such as the susceptible population. Since disease is transmitted by contact, the susceptible population determines the infection rate just as the contagious population does. The larger the susceptible population, the larger the infection rate. But the larger the infection rate, the smaller the susceptible population, which results in a negative feedback loop. The coupling of this negative loop to the positive loop described above yields a growth followed by a decline. As the susceptible population is replenished, the growth-then-decline pattern repeats itself, yielding the observed epidemic cycles.

The complete list of the model equations is given in the Appendix. Here, we describe how and where noise is introduced to the system. The first 'noisy' equation is the 'infection rate' IR:

```
   R IR.KL=IPC.K*CR.K   people/month
   IPC: Infections per contact
   CR: Contact rate
```

IPC is a normal random variable sampled every INT1 time units (months):

```
   A IPC.K=0.10+SAMPLE(NOIS1.K,INT1,0.0)   People/contacts
   A NOIS1.K=NORMRN(0.0,STDV1)   Normal random variate
   C INT1=3.0   Sampling interval (months)
   C STDV1=0.005   Standard deviation (5% of the average IPC)
```

and the contact rate CR is given by:

```
   A CR.K=CF*SP.K*CP.K   Number of contacts/month.
```

Where,

```
   CF=0.004 Contact fraction (per month)
   SP: Susceptible Population
   CP: Contagious Population
```

The second noisy equation is the 'recovery rate' RR:

```
   R RR.KL=CPR.K/RT.K
   CPR: Contagious Population recognizable
```

RT is the recovery time, and chosen to be a normal random variable with mean 2 (months), standard deviation 0.2, sampled at every 0.5 months:

```
   A RT.K=2+SAMPLE(NOIS2.K,INT2,0.0)
   C INT2=0.5   Sampling Interval
   A NOIS2.K=NORMRN(0.0,STDV2)
   C STDV2=0.2   Standard deviation (10% of the average RT)
```

Finally, we assume that the observed variable is the 'contagious population recognizable' CPR, distorted with an 'observation error':

```
   A OBS.K=CPR.K+SAMPLE(ONOIS.K,OINT,0.0)
   C OINT=1.0   Sampling Interval
   A ONOIS.K=NORMRN(0.0,OSTDV.K)
   A OSTDV.K= 0.05*CPR.K   (5% observation error)
```

This model which represents the synthetic system was not validated in any formal way because we did not think that this was necessary. What was important for this project was that the synthetic system had a certain realism so that the entire research would be more than a purely academic exercise. The realistic properties of the synthetic system include the non-linear relationships, coupling of several positive and negative loops, inclusion of system noise and observation errors and an order high enough (5 third order delays and 2 accumulations, adding to 17) to give the system a realistic

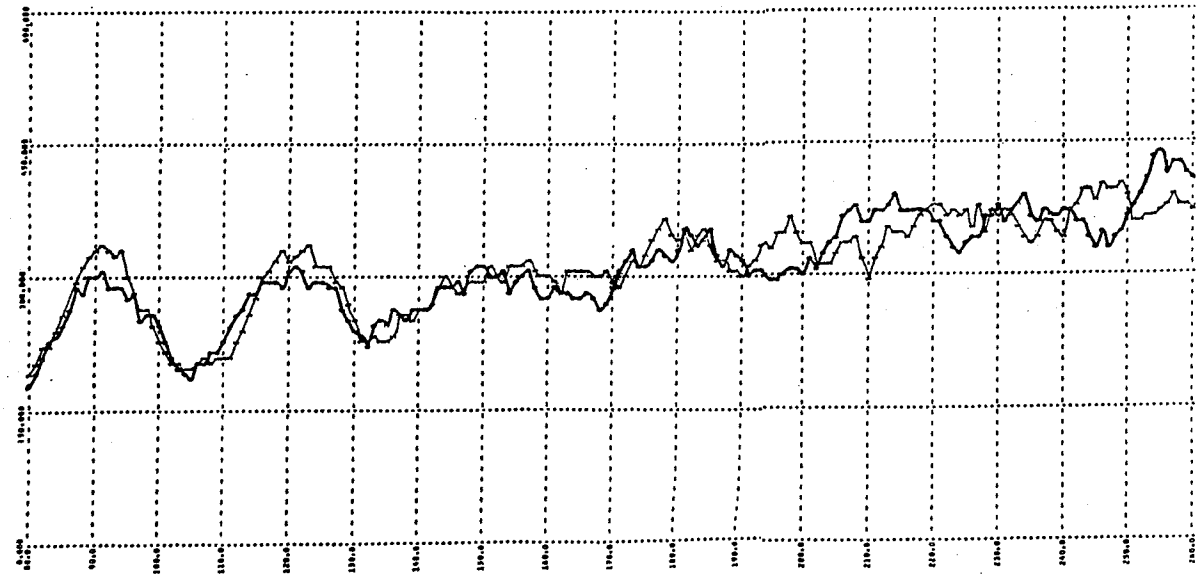
internal momentum. Another important property is the generic structure: The structure of the epidemic model can also be used to represent the dynamics of the spread of information among people. With some modifications, it can explain the predator-prey interactions and the dynamics of marketing. Finally, an important characteristic of the synthetic system is its ability to generate three different types of patterns. When both the immunity losing loop and the population growth loops are active, the system exhibits the oscillatory growth pattern seen in figure 2a. When the growth loops are turned off, we observe constant-mean oscillations (2b), and when the immunity is assumed to be permanent, then the behavior changes to the 'recurrent epidemics' type seen in figure 2c.

V- THE RESULTS

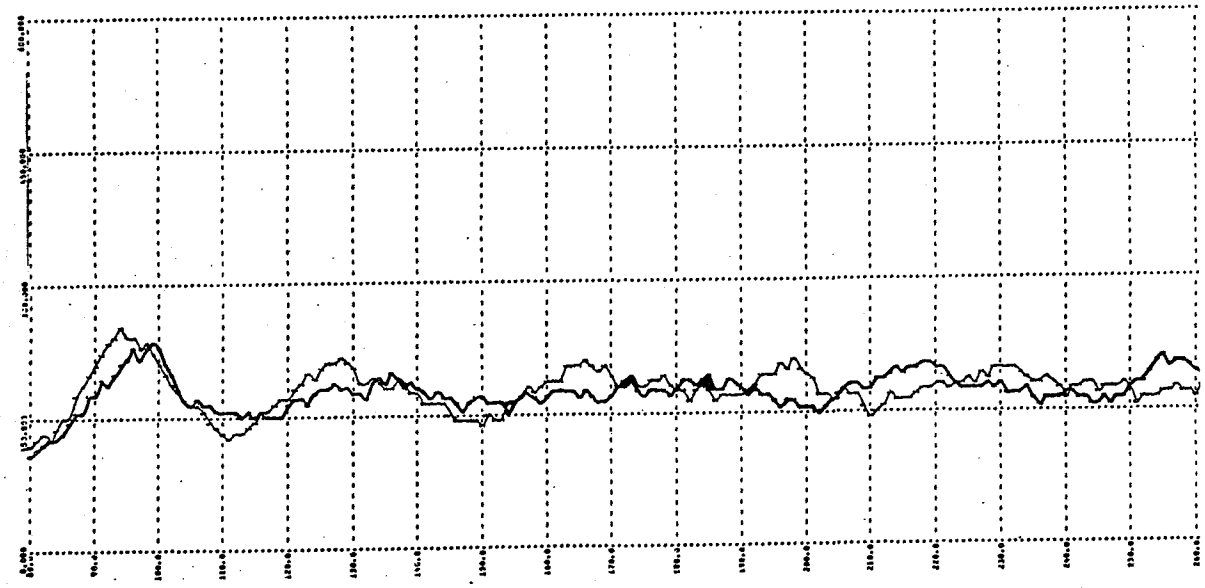
For the synthetic system, the noise seed is set to 7654321 and then, for each experimental design, the model is run with four different noise seeds (6654321, 5654321, 4654321, 3654321). Then, the 'real' output is compared to the four different model outputs by using the quantitative procedure outlined in section III. A FORTRAN program is written to do the computations and the plotting*. The computations are all based on the segments of the time histories between t=50 and 300. The program first fits a linear regression line to the two behavior patterns and detrends them if the analyst wishes to do so. Next, the autocorrelation and partial autocorrelation functions, their variances, the crosscorrelation function, the 'percent error in the means', the 'percent error in the variations' and the 'discrepancy cofficient' U are computed and printed out. The program also plots the autocorrelation functions, the differences between them (together with 2se bands) and the autocorrelations of the residual series $e_i = S_i - A_i$.

A- The Effect of the System Noise Only: Consider first the effect of 'moderate noise' (INT1=3, STDV1=0.005, INT2=0.5 and STDV2=0.2) with no observation error. The system behavior and the model behavior with seed 6654321 for this case are shown in figure 2. As an illustration, consider the runs of figure 2a ('growth case'). The corresponding autocorrelation functions (of the detrended patterns) are plotted in figure 3a. Note that the autocorrelation functions estimate the major periods of the referent time patterns. In figure 3b we see that the two functions are not significantly different at any lag. The slope estimates for the 'real' and model-generated behavior patterns are 0.818 and 0.785 which are not significantly different. The crosscorrelation function reaches a maximum of 0.783 at lag 0 (meaning no phase lag). Next, in figure 4, we plot the autocorrelation function of the residual series $S_i - A_i$. Note that, although the only difference between the model and the real system is the sequences of random numbers used in the two, the autocorrelation function of the residuals shows that the residuals are strongly autocorrelated and even cyclic. This is a clear demonstration of the absurdity of expecting accurate 'point forecasts' from SD models. Finally, the other statistics are printed in table 1. The 'percent eror in the means' E1=0.028 and the 'percent error in the variations' E2=0.0915. Notice how large the discrepancy coefficient (U=0.3645) can be even in this limiting case of 'perfect' model with moderate

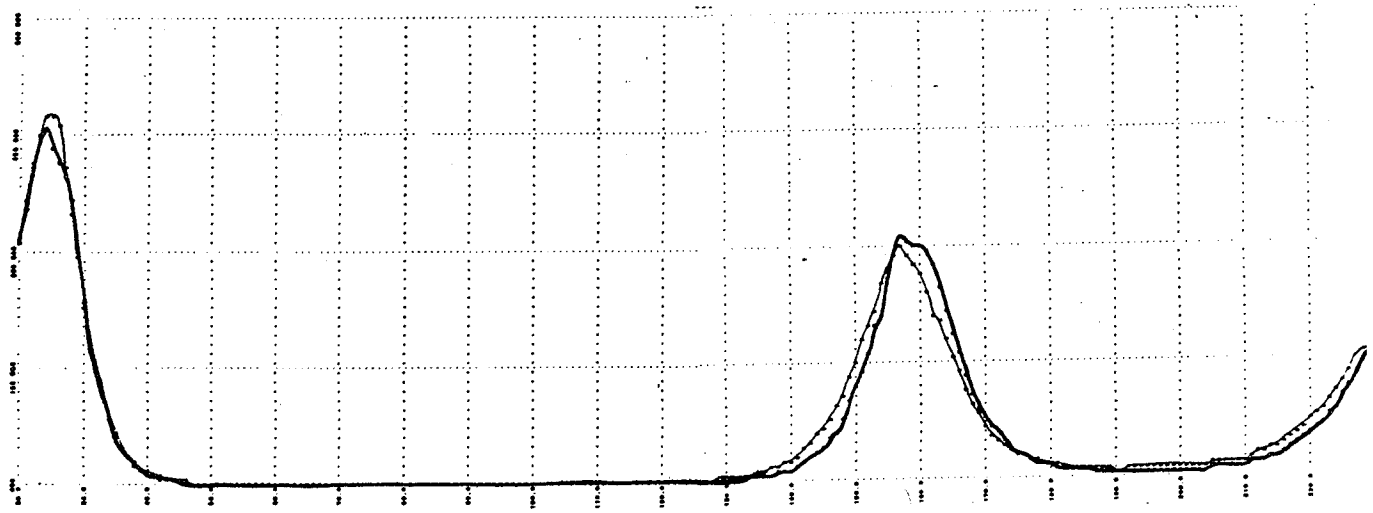
* A listing of the program is not provided in the appendix because the space limitation does not permit and the program is not well-documented yet. Interested persons may contact the author for a listing of the program.



(a)



(b)



(c)

Figure 2. Three Types of Patterns Exhibited by the System:
(a) Growth, (b) No Population Growth, (c) With Permanent Immunity.

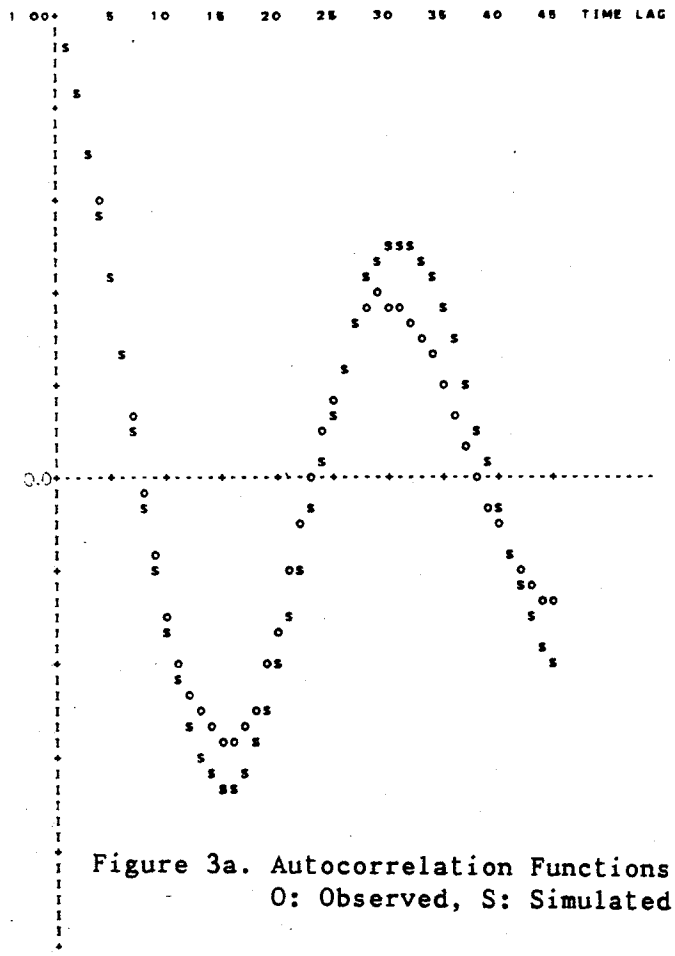


Figure 3a. Autocorrelation Functions: O: Observed, S: Simulated.

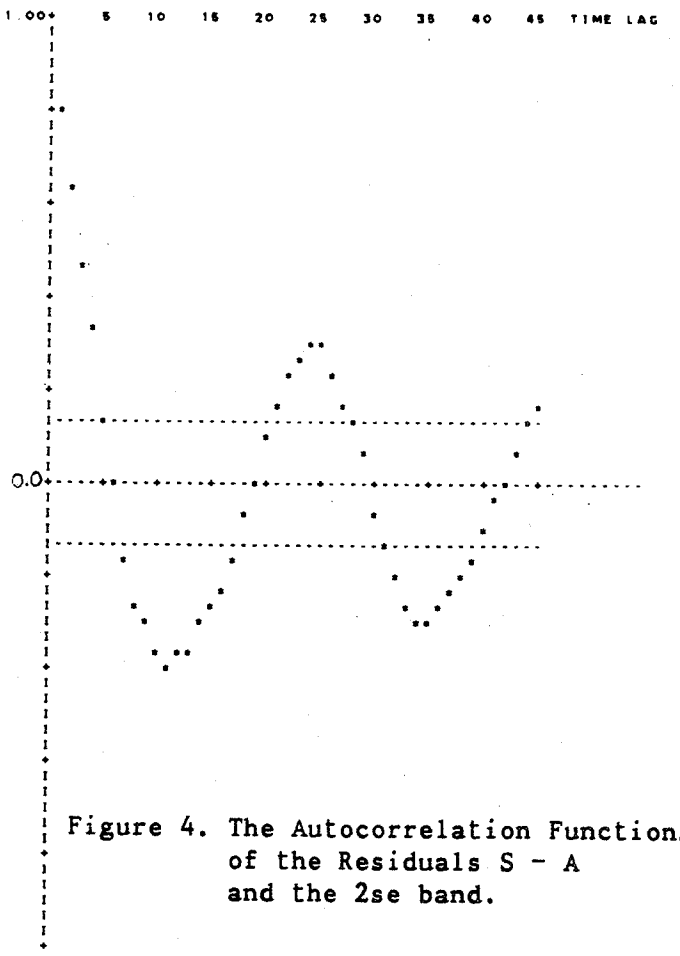


Figure 4. The Autocorrelation Functions of the Residuals S - A and the 2se band.

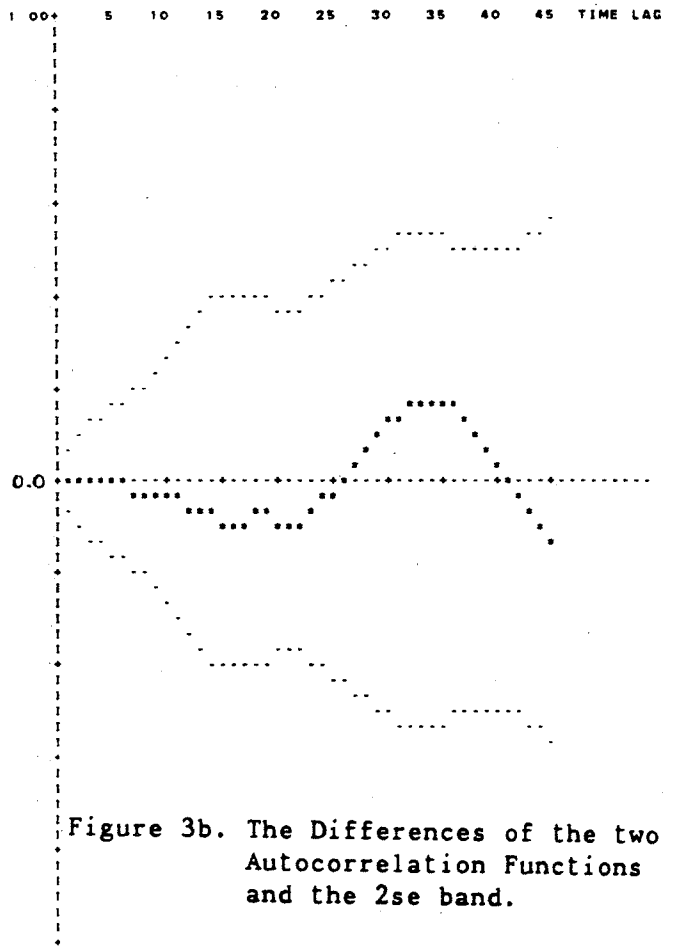


Figure 3b. The Differences of the two Autocorrelation Functions and the 2se band.

Table 1. Statistics Computed for the Runs of Figure 2a.

REGRESSION COEFFFICIENTS FOR OBSERVED TREND

B0: 212.6 VAR(B0): 22.70 B1: 818 VAR(B1): .0011

REGRESSION COEFFICIENTS FOR SIMULATED TREND

B0: 218.5 VAR(B0): 18.73 B1: 785 VAR(B1): .0009

SUMMARY STATISTICS

OBSERVED :	M1:	212.8
	M2:	37.4
	M3:	30.3
	M4:	83.1
SIMULATED :	M1:	218.8
	M2:	34.0
	M3:	22.0
	M4:	45.8

PERCENT MOMENT ERRORS

E1: .0280
E2: .0915

DISCREPANCY COEFFICIENT U : .3845

noise.

We repeat the same experiments with the other noise seeds (5654321, 4654321, 3654321) and then change the experimental condition to the 'increased system noise'. This is accomplished by doubling STDV1 (to 10%) and INT2 (to 1.0). The summary results are shown across the second experimental condition in table 2. When compared to the moderate noise setting, the quantitative measures naturally indicate poorer correspondence. The crosscorrelations drop, E1, E2 and U increase. But the trend estimates and the autocorrelation functions are still in good agreement.

And finally, the above experiments are repeated for the 'no-growth' case, with both the moderate and increased noise conditions. The corresponding results are summarized across the first two experimental conditions of the second half of table 2. Comparing these results to the ones obtained in the growth case, we observe that in the no-growth case the quantitative measures in general indicate better correspondence.

B- The Effect of the Observation Errors: In this set of experiments, we introduce 5% observation errors. First we assume inependent errors by letting OINT=1.0. Naturally, the correspondence measures deteriorate compared to the cases with no observation errors (compare for instance the third and the first experimental rows of table 2). It is important to note that the inclusion of independent observation errors gets reflected in the autocorrelation functions. In three of the four replicates, the autocorrelation functions exhibit 'critical' differences ('critical' meaning that the difference is smaller than 2se but larger than 1.5se). In all of those three runs, the difference occurs within the first 3 to 4 lag autocorrelations. The high frequency observation errors significantly lower the autocorrelations at very small lags. In such cases, since we know that the low lag differences are caused by the presence of hi-frequency observation errors, we can ignore these low lag differences in reaching a conclusion about the behavior validity. One can actually make use of the autocorrelation functions to find out if one time pattern has hi-frequncy components not present in the other.

Next, we make the observation errors correlated by setting OINT=3.0. The results are shown in the fourth row of table 2. Comparing these results to the results of the previous row, we see that in general none of the similarity measures change substantially. The only obvious change is observed in the autocorrelation functions: since the observation errors are now correlated, they do not cause the low lag autocorrelations to drop substantially. Hence, the autocorrelation tests in this case do not yield critical diferences.

When these experiments are repeated for the 'no-growth' case, the results are very similar to the previous ones. What we have observed for the growth case are also valid for the 'no-growth' case. Before concluding, we must add that, in both the 'growth' and the 'no-growth' cases, the effect of moderate observation errors is quite small compared to the effect of moderate system noise. The latter, by travelling through the feedback loops, can cause substantial behavior distortions such as irregular phase shifts and amplitude variations.

C- The Effect of the Parameter Errors: As the first set of parameter errors we increase IP from 15 to 18, RTAVG from 2 to 2.4 and CF from 0.004 to 0.0048 (all

Table 2. Summary of Results obtained under different experimental conditions. Each experimental cell has four replicate runs.

The experimental condition :		Significant difference in : Trends		Significant difference in : Autocorr. functions		Maximum Cross-correlation and its lag		Percent Error in the: Means E1		Percent Error in the: Variations E2		Discrepancy Coefficient U	
GROWTH	Moderate system noise	No	No	No	No	.738 at 0	.647 at 0	.028	.008	.091	.004	.364	.420
		No	No	No	No	.802 at 2	.794 at-1	.008	.005	.186	.206	.354	.353
	Increased system noise	No	No	No	No	.443 at 1	.268 at 1	.028	.022	.167	.058	.534	.608
		No	No	No	No	.610 at 3	.660 at-2	.007	.034	.286	.333	.511	.466
	Independent observation errors	No	No	Crit.	Crit.	.706 at 0	.611 at 1	.032	.015	.012	.076	.384	.446
		No	No	Crit.	No	.757 at 2	.762 at-2	.023	.015	.254	.337	.398	.405
	Correlated observation errors	No	No	No	No	.727 at 1	.634 at 1	.032	.028	.038	.044	.374	.431
		No	No	No	No	.758 at 2	.747 at-2	.023	.016	.246	.372	.389	.417
	'Hi-gain' parameter error	No	Crit.	Crit.	No	.803 at 1	.758 at 1	.190	.163	.631	.637	.390	.421
		Crit.	Crit.	Crit.	Crit.	.827 at 2	.820 at-1	.145	.164	.944	.906	.444	.425
	'Lo-gain' parameter error	Yes	Yes	No	Crit.	.462 at 13	.362 at 13	.248	.276	.353	.015	.821	.803
		Yes	Yes	Yes	Crit.	.512 at 11	.276 at 7	.286	.264	.029	.048	.769	.694
NOGROWTH	Moderate system noise	—	—	No	No	.742 at 0	.768 at 0	.012	.006	.045	.005	.360	.341
		—	—	No	No	.870 at 0	.814 at-2	.004	.010	.261	.108	.278	.347
	Increased system noise	—	—	No	No	.395 at 0	.449 at 1	.019	.020	.130	.022	.553	.530
		—	—	No	No	.727 at 1	.656 at-3	.005	.012	.340	.222	.405	.493
	Independent observation errors	—	—	Crit.	Yes	.733 at 0	.725 at 1	.008	.006	.008	.036	.365	.375
		—	—	No	Crit.	.839 at 1	.789 at-2	.0002	.009	.271	.182	.308	.383
	Correlated observation errors	—	—	No	No	.730 at 0	.737 at 1	.016	.004	.048	.020	.368	.363
		—	—	No	No	.828 at 0	.785 at-2	.002	.010	.293	.197	.318	.377
	'Hi-gain' parameter error	—	—	No	No	.775 at 1	.775 at 1	.163	.155	.992	1.05	.459	.477
		—	—	No	No	.867 at 1	.841 at 0	.156	.160	1.28	1.03	.468	.431
	'Lo-gain' parameter error	—	—	Crit.	Yes	.428 at-22	.456 at 16	.272	.278	.308	.037	.852	.849
		—	—	No	Crit.	.622 at 13	.389 at 14	.284	.272	.184	.146	.822	.779

changes are 20%). The main effect of this 'hi-gain' setting is to increase the amplitudes of the oscillations. This is clearly observed in the E2 values, across the fifth experimental condition of table 2. A secondary effect is to raise the mean values slightly, as reflected in the E1 values. For the 'no-growth' case, the effect of the 'hi-gain' parameter error is essentially the same (see the corresponding entries in the second half of table 2).

As the second set of parameter errors, IP is again increased to 18, but CF is decreased by 20% (to 0.0032). The main effect of this 'low-gain' setting is to lengthen the periods of the oscillations and to lower the mean values. The period errors are observed in the autocorrelation functions and the mean errors in the E1 values. Note also that the crosscorrelation functions have rather small maxima and they occur at large lags (such as +13 and -22). But these lags are not due to a phase lag between the model-generated and the observed behavior patterns. Rather, they are artifacts of the errors in the periods of the two behavior patterns. Therefore, if a significant period error is involved, the crosscorrelation function is not readily interpretable. Finally, note that in all four of the runs with 'low-gain' setting, there are significant errors in the trend components (the first column in table 2), which is a another effect of this setting. It is important to understand that whenever there is a significant error in the trend components, the 'percent error in the means' E1 loses its meaning . When the behavior patterns involve trends, the mean values are meaningfully comparable only if the trend components are not significantly different.

D- The Effect of the Input Error: The synthetic system is excited at t=8 with an impulse function, creating a sudden inflow of 100 infected people. To produce a phase lag, in this experimental condition we change the initiation time to t=18. When the resulting behavior pattern is compared to the 'real' behavior, we see that the crosscorrelation function reaches its maximum at about +10, in all four replicate run: For the growth runs, the estimates are 9, 9, 10 and 7 for the no-growth runs, they are 9, 10, 11 and 9. In these cases, we see that the crosscorrelation functions provide adequate estimates of the phase lag. The crosscorrelation function becomes problematic when the system noise is strong enough to cause the function to shift its maximum to a non-zero lag. Two criteria must be taken into account in deciding whether the phase lag indicated by the crosscorrelation function is systematic or random . First, when the phase lag is systematic, the lag at which the function reaches its maximum does not show a substantial variation from one noise seed to another. Second, in case of a systematic phase lag, the maximum and the minimum are quite large compared to the case of a random unsystematic phase lag. One may check the quantity {max-min} and suspect a systematic phase lag if it is 'large enough'. (In our experiments, {max-min} was always larger than 0.80 in presence of a systematic phase lag).

VI- CONCLUSIONS AND RECOMMENDATIONS

The experimental results suggest that the six-step quantitative procedure is appropriate for SD model behavior evaluation. Various tools used in the procedure are interdependent. Therefore, to be most informative, the tools must be used in a specific sequential order:

Step 1- Trend Comparison and Removal. The trend components (if any) must be compared as the first step. If a significant error in the trends is

discovered, then a model revision is called for. The trend components must be removed only after deciding that there is no significant difference between them.

Step 2- Comparing the Periods. The suggested autocorrelation test is able to detect significant errors in the periods. High frequency noise components affect only the very low lag autocorrelations. Hence, the test can also be used to find out if one behavior pattern has high frequency noise not present in the other. The test is quite insensitive to other types of pattern errors.

Step 3- Comparing the Means. The 'percent error in the means' E1 has rather small variability. In the no-growth case, E1 never goes above 0.02, unless there is a systematic parameter error. When growth is involved, the same limit is about 0.03. Note that if there is a significant difference in the trend components (i.e. Step 1 not passed), then E1 loses its meaning.

Step 4- Comparing the Variations. The 'percent error in the variations' E2 has large variability. For the growth case, E2<0.20 (approximately), unless there is a source of systematic error. For the growth case, the same limit is about 0.25.

Step 5- Testing for Phase Lag. The crosscorrelation function does provide an estimate of a potential phase lag. But the crosscorrelation function being maximum at a non-zero lag does not always indicate a systematic phase lag, because the system noise and/or autocorrelated observation errors may also cause minor irregular phase shifts. A systematic phase lag must be suspected if the lag at which the crosscorrelation function is maximum does not show substantial variation from one noise seed to another, and if the maximum and minimum crosscorrelations are susbstantial. (In our experiments, the quantity {max-min} was always larger than 0.80 in presence of a systematic phase lag). Finally, the interpretation of the crosscorrelation function becomes quite ambiguous if there is a significant error in the periods of the referent time patterns (i.e. Step 2 not passed), because period errors have substantial effect on the crosscorrelation function.

Step 6- As the last step, compute the discrepancy coefficient U, as a single summary measure of behavior accuracy. Since U is basically a point oriented measure, whereas SD models are pattern oriented, rather large U values must be tolerable in SD behavior evaluation. Experiments show that U can be as large as 0.60 even for a 'perfect' model with no structure or parameter errors. This result is in agreement with Rowland and Holmes (1978) who analyze Theil's coefficient in the context of dynamic mathematical models and suggest that values betwen 0.4 and 0.7 should imply average-to-good models.

The suggested quantitative tools are not appropriate for all types of behavior patterns. For instance, the 'recurrent epidemics' type of behavior of figure 2c, which is highly deterministic and transient, can not be evaluated by using such statistical tools. These types of patterns (highly deterministic, transient) must be evaluated by using graphical measures of specific behavior charactesistics. The statistical tools typically apply to more or less stationary, steady-state behavior patterns.

Experiments show that oscillatory growth pattern yields poorer similarity measures than the purely oscillatory pattern. Also, behavior accuracy exhibits

substantial variation from one noise seed to another. Therefore, the tests suggested above should always be carried out with several noise seeds, basing the decisions on the avarages of the several runs. (Experimental results of this paper are all based on 4 replicates). Much of the variation is caused by the system noise rather than the observation error. The effect of the observation error is quite weak compared to that of the system noise, which, by traveling through the loops, can cause significant behavior distortions such as irregular phase shifts and amplitude variations.

Finally, we must emphasize that passing the suggested tests does not imply model validity. Structural validity, which is not adressed by these tests, is the most important condition for model validity. Given that the model is structurally valid however, a positive outcome of the above tests does imply that the parameters and the input functions are accurately estimated and that the model exhibits an adequate behavior pattern.

The quantitative tools recommended in this paper need more extensive testing, on other types of systems and other types of behavior patterns. (We are currently in the process of testing them on Jay Forrester's Market Growth model). Such extensive experimentation is required before we can come up with 'reasonable' acceptance/rejection limits. As another extension, the appropriateness of other statistical tools such as spectral analysis or 'pattern recognition' techniques can be investigated by using the experimental methodology of this research.

APPENDIX - The DYNAMO Equations for the Synthetic System

```
*  EPIDEMICS
NOISE 7654321
L SP.K=SP.J+DT*(ILR.JK+IILR.JK-IR.JK-DR1.JK)   SUSCEPTIBLE POPULATION
N SP=SPN
C SPN=3000
R ILR.KL=(1-DF5)*DELAYP(RR.JK,IP,PHI)      IMMUNITY LOSING RATE
C IP=15  MONTHS   IMMUNE PERIOD
C DF5=0.02
R IILR.KL=(1-DF6)*DELAYP(BR.JK,IIP,II)     INFANT IMMUNITY LOSING RATE
C IIP=6  MONTHS   INFANT IMMUNITY PERIOD
C DF6=0.01
A NOIS1.K=NORMRN(0.0,STDV1)
A IPC.K=0.10+SAMPLE(NOIS1.K,INT1,0.0)      INFECTION PER CONTACT
C INT1=3.0            SAMPLING INTERVAL
C STDV1=0.005
R IR.KL=IPC.K*CR.K INFECTION RATE
A INP.K=PULSE(AMP,START,RPEAT)          INITIAL INPUT OF INFECTED
C AMP=400
C START=8.
C RPEAT=500.
R IRTOT.KL=IPC.K*CR.K+INP.K   TOTAL INFECTION RATE
N IRTOT=IR
R DR1.KL=DF1*SP.K   DEATH RATE 1
C DF1=0.002 PER MONTH   DEATH FRACTION 1
C LP=2 MONTH     LATENCY PERIOD
R LLR.KL=(1-DF2)*DELAYP(IRTOT.JK,LP,PI)    LATENCY LEAVING RATE
N LLR=LLRN
C LLRN=0
C DF2=0.02
C ASP=2  MONTH    ASYMPTOMATIC PERIOD
R SSR.KL=(1-DF3)*DELAYP(LLR.JK,ASP,CP)     SYMPTOMS SHOWING RATE
C DF3=0.03
A CR.K=CF.K*SP.K*CP.K         CONTACT RATE
A CF.K=0.004           PER MONTH CONTACT FRACTION
L CPR.K=CPR.J+DT*(SSR.JK-RR.JK-DR4.JK)  CONTAGIOUS POPULATION RECOGNIZABLE
N CPR=CPRN
C CPRN=0
```

```
R RR.KL=CPR.K/RT.K   RECOVERY RATE
A RT.K=RTAVG+SAMPLE(NOIS2.K,INT2,0.0)   RECOVERY TIME
C RTAVG=2.         AVERAGE TIME TO RECOVER
C INT2=0.5  SAMPLING INTERVAL
A NOIS2.K=NORMRN(0.0,STDV2.K)
A STDV2.K=0.10*RTAVG
R DR4.KL=DF4*CPR.K   DEATH RATE 4
C DF4=0.003
R BR.KL=(1-PBDF.K)*DELAYP(TCR.JK,PP,CB)      BIRTH RATE
C PP=9  MONTHS  PREGNANCY PERIOD
R TCR.KL=FF*(CP.K+PI.K+SP.K+PHI.K)  TOTAL CONCEPTION RATE
C FF=0.006 PER MONTH  FERTILITY FRACTION
A PBDF.K=DFLL+(DFUL-DFLL)*FS.K   PRE-BIRTH DEATH FRACTION
C DFLL=0.02
C DFUL=0.03    DEATH FRACTION UPPER LIMIT
A FS.K=(PI.K+CP.K+CPR.K)/(PI.K+CP.K+CPR.K+SP.K+PHI.K)  FRACTION SICK
NOTE
A OBS.K=CPR.K+SAMPLE(ONOIS.K,OINT,0.0)    OBSERVED CPR
C OINT=1.00
A ONOIS.K=NORMRN(0.0,OSTDV.K)
A OSTDV.K=OFRAC*CPR.K
C OFRAC=0.05
NOTE
SPEC DT=0.25/LENGTH=300/PRTPER=2/PLTPER=1
PLOT OBS=O(0,600)/PHI=M/SP=S/II=I
PRINT OBS,PHI,SP,PI,II
RUN REAL
```

REFERENCES

Anderson, O.D. and M.R. Perryman, eds., Applied Time Series Analysis, Netherlands: North-Holland Publishing Co., 1982.

Ansoff, H.I. and Slevin, D.P., "An Appreciation of Industrial Dynamics", Management Science, Volume 14, 1968, pp. 383-397.

Forrester, Jay, Industrial Dynamics, Cambridge: MIT Press, 1961.

Forrester, Jay, "A Response to Ansoff and Slevin", Management Science, Volume 14, 1968, pp. 601-618.

Forrester, Jay and Peter M. Senge, "Tests for Building Confidence in System Dynamics Models", in Forrester, Jay et al., eds., System Dynamics, NewYork: North-Holland, 1980.

Griner, G. et al, "Validation of Simulation Models Using Statistical Techniques", Proceedings Summer Computer Simulation Conference, 1978, pp. 54-59.

Peterson, D.W., "Statistical Tools for System Dynamics", in J. Randers, ed., Elements of the System Dynamics Method, Cambridge: MIT Press, 1980.

Rowland, J.R. and W.M. Holmes, "Simulation Validation with Sparse Random Data", Comput. Elect. Engng. , Vol. 5, 1978, pp. 37-49.

Senge, Peter M., "Statistical Estimation of Feedback Models", Simulation, Volume 28, 1977, pp. 177-184.

Sterman, John D., "Appropriate Summary Statistics For Evaluating the Historical Fit of System Dynamics Models", Dynamica, Volume 10, 1984, pp. 51-66.

Theil, H., Economic Forecasts and Policy, Amsterdam: North-Holland, 1958.

Wright, R.D., "Validating Dynamic Models: An Evaluation of Tests of Predictive Power", Proceedings Summer Computer Simulation Conference, 1972, pp. 1286-96.