FORECASTING S-SHAPED DIFFUSION PROCESSES VIA RESPONSE MODELING METHODOLOGY (RMM)

JORS. MS #6971.

Revision 1. Submitted: Dec., 2005

ABSTRACT

Diffusion processes abound in various areas of corporate activities, such as the time-dependent behavior of cumulative demand of a new product, or the adoption rate of a technological innovation. In most cases, the proportion of the population that has adopted the new product by time t behaves like an S-shaped curve, which resembles the sigmoid curve typical to many known statistical distribution functions. This analogy has motivated the common use of the latter for forecasting purposes. Recently, a new methodology for empirical modeling has been developed, denoted Response Modeling Methodology (RMM). The error distribution of the RMM model has been shown to model well variously shaped distribution functions, and may therefore be adequate to forecast sigmoid-curve processes. In particular, RMM may be applied to forecast S-shaped diffusion processes. In this paper, forty-seven data sets, assembled from published sources by Meade and Islam (1998), are used to compare the accuracy and the stability of RMM-generated forecasts, relative to current commonly applied models. Results show that in most comparisons RMM forecasts outperform those based on any individually selected distributional model.

Keywords: Diffusion processes; Distribution fitting; Forecasting; Inverse normalizing transformations; Sigmoid curves; Non-linear regression; Response Modeling Methodology;

INTRODUCTION

Diffusion processes are stochastic processes that describe the time-dependent evolution of the diffusion of a random phenomenon, such as the penetration into the market of a new product, or the adoption rate by the relevant population of a technological innovation. Most often, the response of interest is expressed as the time-dependent percentage of the relevant population, where the diffusion has already taken place, for example, the population proportion, P_T, that has already purchased the new product by time T. The forecasting problem is then that of forecasting, at time T, the value of the response at time T+L, given the adoption history up to and including T. More formally, denote by P_t the cumulative diffusion rate by time t ($0 \le P_t \le 1$). The required forecast is that of $E(P_{T+L} \mid P_1, P_2, \ldots, P_T)$. Occasionally, inverse forecasts are also needed, namely, forecast the time, T_P, by which the cumulative adoption rate would surpass a given desirable threshold value, P, given the adoption history. More formally, find estimate for $E(T_{P(k+l)} \mid T_{P1}, T_{P2}, \ldots, T_{Pk})$, where T_P is the time by which the cumulative penetration rate has reached the value P. For example, forecast $T_{0.5}$, the time by which half the target population has already adopted the technological innovation, given the k points of the time-series, $\{p_1, p_2, \ldots, p_k\}$.

Various approaches have been pursued to model S-shaped diffusion processes. A brief description of the main models associated with these approaches follows. It is based on a more comprehensive survey given in Meade and Islam (1998, henceforth MI)[1], and the interested reader may refer to this source for further details.

Models of diffusion processes used for forecasting may be divided into three major categories. The first includes the familiar trend curve, where

$$P_t = f_1(t) + \varepsilon_t, \qquad (1)$$

$f_1(t)$ is a sigmoid-curved function of time, and the error term is assumed to be homoscedastic (homogenous).

The second category comprises linearized trend curves, which assume that a non-linear transformation of Y_t may be expressed as a linear function of t:

$$f_2(P_t) = a + bt + \varepsilon_t \qquad (2)$$

These models resemble the link function, used in generalized linear

modeling (GLM), where a non-linear transformation of the response mean is assumed to be expressible in terms of the linear predictor. Note, however, that here the non-linear transformation is applied to the data, $\{P_t\}$, and not to the mean function. As noted by MI (based on Young and Ord[2]), for some data sets the error structure in (2) may be more appropriate than that of (1).

A third category of models use non-linear auto-regressive (AR) relationships of first order, namely,

$$P_t = f_3(P_{t-1}) + \varepsilon_t \qquad (3)$$

Occasionally, models in this group are obtained from those of the first, by setting

$$dP_t = P_t - P_{t-1} = \partial f_1(t)/\partial t + \varepsilon_t \qquad (4a)$$

or from those of the second category, by setting

$$[df_2(P_t)/dP_t][dP_t/dt] = b,$$

from which, with dt=1, one obtains, approximately,

$$dP_t = P_t - P_{t-1} = b / [df_2(P_{t-1})/dP_{t-1}] + \varepsilon_t \qquad (4b)$$

Based on a comprehensive literature review, MI have identified altogether twenty-nine models, which belong to either one of these three categories. A close examination of these models reveals that about half (15) are either trend curve models or linearized trend models. A common feature shared by all these models (and AR models derived thereof) is the attempt to formulate a relationship between P_t and t, either by expressing P_t explicitly in terms of t (trend curve models), or by finding a non-linear transformation of P_t that would result in a linear transformation of t (linearized trend models).

Surveying commonly used models, which belong to either of the above categories, shows that the majority of the models are based on an adaptation of some known statistical distribution (like the logistic, the log-normal or Gompertz). This is due to the fact that most trend models (or derivatives thereof) take advantage of the fact that the S-shaped curve, characteristic to a diffusion process, also typifies most known statistical distributions. Given a distribution function F(x), where X is the random variable, a plot of F(x) vs. x would most likely reveal a typical sigmoid-shaped curve. Substituting P_t for F(x) and t for x, relationships that have originally been derived to depict distribution functions may thus serve to describe the relationship $P_t=f(t)$. As a

result, some well known and commonly applied distributions, like the normal, the logistic, the extended logistic, Gompertz, the log-normal and Weibull, are routinely used to model P_t. MI classify the various models according to the position of the point of inflexion of the curve, since its timing coincides with the maximum rate of diffusion (or penetration). MI classification scheme is described in Appendix 1.

A major implication of this classification is that different data sets (or different assumptions regarding these sets) may require the adoption of different models that reflect the actual conduct of the time series (or allied assumptions) vis-à-vis the inflexion point. Thus, different statistical distributions may show different degrees of consistency (or goodness-of-fit) for a given data set. This is a bothering aspect of the current state-of-the-art of modeling S-shaped diffusion processes, which the current paper attempts to resolve.

Recently, a new approach for empirical modeling has been developed (Shore[3], and references therein), which is intended to serve as a general platform for empirical modeling of both systematic variation and random variation. The new approach is denoted Response Modeling Methodology (RMM), and its error distribution has been shown to include as special cases many known statistical distributions, transformations and approximations (Shore[3,4]). Furthermore, the quantile function of the error distribution is expressed in terms of the standard normal quantile, and has been shown to represent well distributions with a wide range of distributional shapes. Based on the RMM error quantile function, a uniform approach may be developed for modeling S-shaped diffusion processes, which would eliminate the need to fit different models to different data sets. This would spare the forecaster the need to specify in advance the assumed distributional model, as prescribed by current common practices, or the requirement to prepare a combined forecast based on a weighted scheme, as done in MI.

In this paper, RMM is applied to model S-shaped diffusion processes. Real data sets would serve to compare the effectiveness of the new approach to current distributional models. The comparison will be conducted by applying the various models to 47 real data sets assembled by MI, and which the latter were kind enough to provide us with. The focus in this study is on modeling

the relationship between P_t and t (first and second category), and no attempt is made at auto-regressive modeling (third category). This self-imposed limitation has allowed us to conduct the comparative evaluation of the various models on such a large scale (fitting, via non-linear regression, forty-seven data sets to each of the models in the comparison set). The new approach, however, may easily be extended to auto-regressive models, not unlike some current auto-regressive models, derived by extending trend- or linearized-trend models (refer, for example, to models (22) and (16), respectively, in Appendix 1 of MI). Furthermore, the new approach, though not including auto-regressive modeling, will be compared to some current auto-regressive models.

In the next section we give a brief introduction of the RMM model, and derive its quantile function. Since the latter is expressed in terms of the corresponding standard normal quantile, we denote this function an Inverse Normalizing Transformation (INT). Several INTs may be derived from the quantile function of the RMM model. Those that will be later used for forecasting are introduced. Normalizing transformations (NTs), explicit expressions for the standard normal quantile in terms of the corresponding quantile of the modeled response, which may also be derived from the quantile function of the RMM model, are introduced too. Both INTs and NTS will be used for modeling and ultimately forecasting S-shaped diffusion processes. In the pursuing section we develop effectiveness measures that will be utilized to evaluate the relative merits of the various models compared in this study. The next four sections assess the effectiveness of the new RMM-based models relative to existent models. The results in these sections are partitioned according to two dimensions:

I. *The response to be forecast* (cumulative adoption proportion at time T, P_T, or the time, T_P, when adoption proportion has exceeded a given threshold value, P).

II. *The purpose of the comparison* (how well a given model describes given data sets, or how well a given model forecasts future values of the response).

Accordingly, the first two sections examine goodness-of-fit statistics that indicate how well the RMM model may describe the given data sets (P_T in terms of T, and T_P in terms of P). Complete data sets are used in the fitting procedures of these sections. The accuracy and the stability of forecasts

generated via RMM are evaluated in the next two sections, and compared to current models. This is done by using, for each of the data sets, the first two thirds of the data for parameter estimation, while the last third serves to examine accuracy and stability of forecasts, generated by the estimated (fitted) model. The last section discusses the results and their implications.

Response Modeling Methodology - The basic model and derivatives

Let Y be a response (a random variable, r.v.), and let $\{X_1, X_2,..,X_k\}$ be K variables that transmit systematic variation to the response via the linear predictor:

$$\eta = B_0 + B_1X_1 + \ldots + B_kX_k \quad (5)$$

Let ε_1 be an additive random error associated with the linear predictor and let ε_2 be another random error associated directly with the response. The basic RMM model is (refer to Shore[3] and references therein)

$$Y = \exp\{(\alpha/\lambda)[(\eta + \varepsilon_1)^{\lambda} - 1] + \mu_2 + \varepsilon_2\}, \quad (6)$$

or

$$W = \log(Y) = (\alpha/\lambda)[(\eta + \varepsilon_1)^{\lambda} - 1] + \mu_2 + \varepsilon_2\} = (\alpha/\lambda)[(\eta + \sigma_1Z_1)^{\lambda} - 1] + \mu_2 + \sigma_2Z_2\} \quad (7)$$

where Z_1 and Z_2 are random variables from a bi-variate standard normal distribution with correlation ρ. Expressing Z_2 in terms of Z_1 *and* another independent standard normal variable, Z:

$$Z_2 = \rho Z_1 + (1-\rho^2)^{(1/2)} Z \quad (8)$$

and also assuming that no systematic variation is transmitted to the response (putting, arbitrarily, η=1), we obtain from (7):

$$W = \log(Y) = (\alpha/\lambda)[(\eta + \sigma_1Z_1)^{\lambda} - 1] + \mu_2 + \sigma_2[\rho Z_1 + (1-\rho^2)^{(1/2)} Z] \quad (9)$$

From this expression, with $\rho = \pm 1$, we finally obtain for the quantile of W, denote it by w, expressed in terms of a corresponding standard normal quantile, z:

$$w = (\alpha/\lambda)[(1 + \sigma_1 z)^{\lambda} - 1] + \mu_2 + \sigma_2\rho z] \quad (10)$$

Re-parameterized, we obtain from (10) for the quantile of Y (denote it by y):

$$y = (M) \exp\{[B/(C/A)][(1+Az)^{C/A} - 1] + Dz\} \quad (11)$$

where M is the median of Y (the quantile of Y corresponding to z=0), and {A,B,C,D} are parameters that need to be estimated. Eq. 11 is denoted the

"Origin" INT. From this INT, parameter-reduced "Off-spring" INTs may be derived (find details in Shore, Chapter 20):

$$y = (M)\exp\{(B)[\exp(Cz) - 1] + Dz\}, \quad (12)$$

$$y = (M)\exp\{(B)[\exp(Cz) - 1]\}, \quad (13)$$

$$y = (M)\exp\{(B/C)[(1+Az)^{C} - 1]\},\ z>-1/A, \quad (14)$$

$$y = (M)\exp[ABz/(1+Az)], \quad (15)$$

$$y = M\,[1 + z/(BC)]^{B}, \quad (16)$$

Note, that (16) is derived from (14) on setting in the latter C=0. The parameter A is then replaced by 1/(BC) so that the inverse of (16) results in the familiar Box-Cox transformation[5]. Also note that for (13)-(16), we may express z explicitly in terms of y. This implies that the inverse of these expressions may be used as normalizing transformations (NTs).

Both INTs and NTs will be later used for forecasting. The following NTs will be used (for easy further reference, these models will later be re-written in terms of the forecasting variables):

Model NT_1 [from (15)]:

$$z = (A)\log(y/M) / [B-\log(y/M)]. \quad (17)$$

Model NT_2 [from (16)]:

$$z = (A/B)\,[\,(y/M)^{B} -1\,]. \quad (18)$$

Using INTs and NTs for forecasting- The basic procedures

The correspondence between the original variables of the INTs and the NTs, and the time series variables associated with the diffusion process, can now be summarized by the following fundamental transformations (in (12)-(18), replace y with t):

Using NT:

$$t \rightarrow z_t \rightarrow \Phi(z_t) = P_t\ ; \quad (19)$$

Using INT:

$$P_t = \Phi(z_t) \rightarrow z_t \rightarrow t\ , \quad (20)$$

where arrows imply transformations (via the fitted NT or INT, or through available tables of the standard normal distribution).

We may now define formally the procedures to fit a model to a given time series and then use it for forecasting. To fit any of the INTs (or NTs) to a time

series representing cumulative relative demand, we transform the given data by two different procedures, each representing a different forecasting objective.

Procedure I

Objective: For a Given Time Series $\{P_t\}$- Model the time series (via an INT), and then use the fitted model to forecast T_P in terms of a specified P.

(Ia) Modeling:

To model the time series, pursue the following procedure:

- Replace y, in the expression for the INT, by t;
- For each observation, find $\Phi^{-1}(P_t) = z_t$ [$\Phi^{-1}(.)$ is the standard normal *inverse* CDF];
- Fit an INT to the new data set, $\{z_t, t\}$, using non-linear least-squares (NL-LS).

(Ib) Forecasting T_P, given P:

- Find $\Phi^{-1}(P) = z$; Introduce z into the fitted (estimated) INT to forecast T_P.

Fitting an INT to a given time series thus allows one to forecast the time, T_P, at which cumulative demand is expected to cross a given threshold value, P.

Procedure II

Objective: For a Given Time Series $\{P_t\}$- Model the time series (via an NT), and then use the fitted model to forecast P_T in Terms of a specified T

(IIa) Modeling:

To model the time series, pursue the following procedure:

- Replace y, in the expression for the NT, with t;
- For each observation, find $z_t = \Phi^{-1}(P_t)$;
- Fit an NT to the new data set, $\{t, z_t\}$, using NL-LS.

(IIb) Forecast P_T, given T:

- Replace y in the NT with T; Use the fitted NT to find $z_T = f(T)$; Find the forecast $P_t = \Phi(z_T)$.

Measures of effectiveness

To determine the relative effectiveness of the various models, as manifested across the 47 data sets used for this study or as revealed for a particular data set across all models, the standard deviation of the residuals (actual minus calculated, or forecast) was computed for each combination of fitted model and data set. From these statistics, averages and standard deviations were calculated either for each model (across all data sets) or for each data set (across all models):

I. Average for Model j (across all data sets):

$$\mu_j(STD)= \sum_{i=1}^{47} STD_{ij} / 47 \qquad (21)$$

where STD_{ij} is the standard deviation of the residuals from fitting model j to data set i. For this measure, small values are desired.

II. Standard deviation for Model j (across all data sets):

$$\sigma_j(STD)= \{ \sum_{i=1}^{47} [STD_{ij}- \mu_j (STD)]^{2} / (47-1) \}^{1/2} \qquad (22)$$

Unlike measure I, which provides average performance, this measure provides a yardstick for model consistency across various data sets. Small values are desired.

III. A weighted measure: $\mu_j(STD)x\sigma_j(STD)$ [namely, (I)x(II)]

This measure (the product of the two previous measures) may provide an overall verdict regarding the adequacy of each model. Small values are desired.

IV. Likelihood of Model j to be a "Winner"

For each data set, models were ranked according to the residual standard deviations. The winner (or the first two or three winners) were then selected (models with smallest standard deviations). For each model, the frequency (or relative frequency) of being a winner, or being included in the group of the first two or three winners, was identified across all data sets. The resulting measures indicated the likelihood that a model, selected for a given data set, would perform better than the other models examined in this study.

A description of the various data sets is given in Appendix 3 of MI[1].

Performance of INTs as models for S-shaped diffusion processes

In this section the new INTs are assessed, relative to current models, with respect to their capability to represent time series derived from diffusion processes. We note that although a certain model can succeed in depicting a given data set, it may perform badly for forecasting purposes. On the other hand, it is extremely unlikely that a model that generally fails in representing a given data series will succeed in forecasting future values. Therefore, in this section and the next we first examine how apt are the above INTs (this section) and NTs (next section) in describing the general behavior of the given 47 data sets, relative to current models.

To prepare for the implementation of fitting procedures, the given $\{P_t\}$ were transformed into $\{z_t\}$ for all 47 data sets. The transformed data were then used to fit INT models, using non-linear least-squares (NL-LS) procedure, as programmed in MATHEMATICA®. The resulting goodness-of-fit statistics, described in the previous section, were then calculated.

Six models were included in this part of the study: The INT given by (12) (henceforth denoted INT_1) and non-AR models that belonged to the restricted set of models, recommended by MI as the most adequate for forecasting (among 29 models). The latter models are displayed in Appendix 1. Re-writing INT_1 in terms of t and P_t, one obtains:

$$t = \exp\{\log(M) + (B)\{\exp[C\Phi^{-1}(P_t) - 1] + D\,\Phi^{-1}(P_t)\}\} \qquad (23)$$

The reason for using INT_1 only was that this is the closest to the "origin" INT, still having a reduced number of parameters (4) that is relatively easy to manage in NL-LS. Unsatisfactory fit achieved for this INT would halt further attempts to compare the other INTs to current models. The converse, however, is not necessarily true, namely, best performance of INT_1, relative to current models, does not guarantee best performance, in terms of goodness-of-fit, of the other "off spring" INTs. These other INT models have to be examined individually relative to their goodness-of-fit capabilities.

Regarding MI models, included in this part of the study, note that although the new approach, as applied here, results in models that are non auto-regressive, two of MI restricted set of models are. Since the latter do not have explicit inverse functions, they may only be compared with the new approach

when fitting P_t (given t) and not for fitting T_P (given P). Consequently, only five of MI models are included in this part of the analysis, and altogether we have six models that are compared across 47 data sets. This implies altogether 282 (6X47) NL-LS fitting runs, with manually delivered initial parameters for each run.

Table 1 displays results for Measures I-III, and Table 2 displays the results for measure IV.

Insert Table 1 about here

Insert Table 2 about here

By any measure, the new INT_1 performs uniformly better than the other models, with Model 5 (the Extended Logistic B) a close second. Restricting the comparison only to the first four "winners" (according to the left part of Table 2), and identifying thereof a single "winner" for each data set, the results in the right-hand side of Table 2 are obtained. Model INT_1 now is shown to perform appreciably better than any other model.

Performance of NTs as models for S-shaped diffusion processes

In this section the adequacy of NTs as models for P_t, in terms of t, is examined. The two normalizing transformations NT_1 (17) and NT_2 (18) are compared to the set of seven models, recommended by MI (Appendix 1).

The NTs are re-written in terms of P_t and t:

Model NT_1:

$$\Phi^{-1}(P_t)= (A)\mathrm{Log}(t/M) / [B-\mathrm{Log}(t/M)]; \quad (24)$$

Model NT_2:

$$\Phi^{-1}(P_t)= (A/B) [(t/M)^B -1], \quad (25)$$

where M is the median of t, namely, the value of t for which $P_t= 0.5$ (M is estimated from the fitting procedure).

Table 3 displays results for Measures I-III for the nine models, corresponding to the results in Table 1.

Insert Table 3 about here

The auto-regressive (AR) models (Models 6 and 7) yield the best results. This is to be expected since AR models forecast next period value based on the available most recent value. Therefore, comparing them to "trend" models

may be inappropriate. Furthermore, the AR models are incapable of one-step forecasting of future periods (apart from next period). Therefore, the AR models were deleted from further comparisons

Table 4 displays results for Measure IV for the seven models (with the AR models excluded). Since this time seven models had been compared, "First two Winners" were defined as models included in the set of the best two models (out of the best five), and "Winner" is the best model out of the best four models.

Insert Table 4 about here (formerly 5)

In terms of likelihood of being a "winner", for all comparisons Model 5 (the Extended Logistic B) performs best. However, in terms of Measures I-III (Table 3 without the AR models), the best performer is NT_2 (based on a re-scaled inverse of the Box-Cox transformation), with NT_1 second best. The Extended Logistic B (Model 5) appears as the worst both in terms of Measures II and III. This implies that while the Extended Logistic B may have, for a given data set, the highest probability of being the best, its performance may occasionally be quite deviant (bad). Thus, its long term (average) performance over an extended period of time, and using incoming data sets from various sources, may be unacceptable. Since NT_2 seems to perform best in terms of average performance, while it is only second to the Extended Logistic B in terms of the rate of being a "Winner" (Table 5), NT_2 can be considered the best model among all models considered.

It should again be noted that in both this section and the previous one, the adequacy of the various models, in terms of how well they represent given data sets, has been examined. The complete data set participated in each NL-LS run. In the next two sections, these analyses are repeated, however forecasts are made for observations not participating in the fitting procedures, and the standard errors of these forecasts are compared for the various models.

Forecasting T_P, given P

The 47 data sets have been subjected to the same analysis as before, however only the first two thirds of each time series was employed in the fitting, and forecast errors for the remaining observations were calculated and used to compute the effectiveness measures. In addition to the measures depicted

earlier, an additional measure was introduced. This is the correlation between the standard deviation of the residuals for observations employed in the fitting procedure (the first two thirds of each time series), and the standard deviation of forecast errors, calculated for the rest of the observations. Correlations were calculated for each model across all 47 data sets. This measure is intended to reflect the capability of the standard deviation, obtained from fitting a model to available observations, to predict the accuracy of future forecasts. Thus, a lower correlation would suggest that if satisfactory goodness-of-fit is achieved from already available observations, this has little bearing regarding the expected size of forecast errors, obtained from the fitted model.

The analysis in this part of the study was conducted in two stages (though not so intended in the first place). In the first stage, the same models that had been examined in previous sections were re-fitted and the forecast accuracy assessed, according to the {2/3,1/3} data partition. However, for INT_1, fitting to seven data sets resulted in an expression that was not monotone increasing, as assumed by the underlying S-shaped model of the diffusion process. These data sets (#18, 27, 32, 37, 38, 42, 43) were therefore discarded, and the rest of the analysis conducted based on the remaining 40 data sets.

The results for the effectiveness measures are detailed in Tables 5 and 6.

Insert Table 5 about here

Insert Table 6 about here

The long term average performance (Measures I-III) for Model 1 (Simple Logistic) and Model 2 (Gompertz) are the best. Furthermore, the correlation between the standard deviations of the residuals in the fitting stage and in the forecasting stage is the highest for these models. In terms of frequency of appearance as "Best performing model", INT_1 appears to be the "Winner", namely, the probability that INT_1 would deliver best performance is the highest. However, one may still wonder why INT_1, which displayed the best goodness-of-fit when complete sets were used in model fitting, has less satisfactory average performance at this stage of the analysis.

The solution for this seems to be that since INT_1 has four parameters, while the Simple Logistic and Gompertz have each only two, fitting INT_1 may at times result in "over-fitting" that would damage the accuracy of forecasts. This is particularly evident by data sets 22 and 41, for which the standard

deviations of the forecast errors are particularly large (31.3 and 30.6, respectively). Deleting these data sets from the analysis resulted in dramatic improvement in the average performance of INT_1, relative to the Simple Logistic (Model 1) and Gompertz (Model 2), as may be realized from Table 7.

Insert Table 7 about here

Note also the relative improvement in the position of Model 5 (Extended Logistic B), which has one additional parameter relative to the Simple Logistic and Gompertz, and therefore may also show unsatisfactory average long-term performance due to occasional over-fitting. These results imply that INT_1 may be considered as the preferred model provided that its performance in the first several forecasting periods is monitored. If satisfactory forecast accuracy is achieved in these periods, the model will probably perform well in the rest of the forecast periods, and very likely better than any other model, as Measure IV (Table 7) clearly indicates.

To test our hypothesis that it is probably due to a too large number of parameters (4) that INT_1 may occasionally provide bad forecasts, three additional "off-spring" INTs, with a reduced number of parameters, were examined. These are INT_3-INT_5 (eqs. 13, 15, 16), each with three parameters (including the median, M). The models are re-written here in terms of t and X_t:

Model INT_3 (from eq. 13):

$$t = (M)\exp\{(B)\{\exp[C\Phi^{-1}(P_t)] - 1\}\} \quad (25)$$

Model INT_4 (from eq. 15):

$$t = (M)\exp\{AB\,\Phi^{-1}(P_t) / [1 + A\,\Phi^{-1}(P_t)]\} \quad (26)$$

Model INT_5 (from eq. 16):

$$t = (M)\,[1 + \Phi^{-1}(P_t)/(BC)]^B \quad (27)$$

Note that INT_3-INT_5 include the log-normal model as a limiting case. Thus, in (25) as C approaches zero and B becomes large so that (B)(C)= const., the behavior of INT_3 tend to the log-normal (Class II_{3B} according to MI classification, relate to Appendix 1). Likewise, in (26), as $B \rightarrow 0$ and C becomes large so that (B)(C)= const., the behavior of INT_4 also approaches that of a log-normal variable. Finally, INT_5 (27) will asymptotically approximate the log-normal as $B \rightarrow \pm\infty$.

Analyzing all 47 data sets (with only the first two thirds of each set used for

fitting), we realize that the problems encountered with INT_1 no longer exist. Therefore, no data set has to be discarded from the analysis. The results of this analysis are displayed for the INTs only in Table 8, and for all models in Table 9.

Insert Table 8 about here

Insert Table 9 about here

Several conclusions may be drawn from these tables (and other results not shown here). First, given the asymptotic log-normality of all INTs examined (which will manifest itself by the extreme values of the parameters obtained in the NL-LS fitting procedure), we realize that out of the 47 data sets fitted at this stage of the analysis, for five sets the parameters' values indicated that the log-normal may indeed be the best model for forecasting.

Secondly, while INT_4 appears as the best model ("Winner") both among all models (LHS of Table 10), and in comparison with non-INT models (RHS of Table 10), the long-term performance (as indicated by averages calculated over all data sets) positions INT_4 (and also INT_5) as the non-preferred choice. This may be realized from Table 10, where results for INT_3-INT_5 are given for the same 38 data sets compared in Table 9.

Insert Table 10 about here

From Tables 8 and 11 the best long-term performing models are Model 5 (the Extended Logistic B) and INT_3, both of which have only three parameters. INT_3, however, has the additional property that it includes the log-normal as a limiting case.

To sum up this analysis, it was realized that while INT_4 delivers best performance for the majority of the data sets, INT_3 may be the preferred choice (next to the Extended Logistic B) in terms of average performance in the long-term. If forecasts for the first several forecasting periods prove that INT_4 delivers satisfactory performance (no extremely outlying forecast errors), this would probably be the best model for the rest of the periods. Alternatively, we may opt for a dynamic updating of parameters based on the most recent observations. This will circumvent the possibility of "over-fitting", the most probable cause for the occasional bad performance of INT_4.

Forecasting P_T, given T

In an earlier section, the adequacy of the new two NTs (eqs. 17 and 18) in modeling S-shaped diffusion processes was examined in terms of goodness-of-fit statistics. In this section, we repeat the analysis in order to evaluate the efficacy of the NTs in forecasting. This will be done, as in the previous section, by using only two thirds of the time-series data for fitting, while the rest of the observations are used to evaluate forecasts accuracy. The reader is reminded that to forecast P_t in terms of t, t is first transformed to z_t (via the fitted NT), and then $\Phi(z_t)$ provides the required forecast of P_t.

Two NTs (eqs. 17 and 18), and five models from MI set of seven (AR models are omitted for reasons expounded earlier) participate in this analysis.

The results are displayed in Tables 11 and 12.

Insert Table 11 about here

Insert Table 12 about here

Examination of the results in these tables, and also results not displayed here (which relate to particular combinations of a data set and a model), show that no difficulties were encountered in the implementation of the NL-LS procedure for any combination (unlike the difficulties encountered with INT_1, as expounded earlier). Referring to the average-performance measures (Measures I-III), we realize from Table 12 that NT_2 (the Box-Cox transformation applied to the re-scaled t), Gompertz and Weibull provide comparable results, and they comprise the set of best average-performing models. However, in terms of frequency of appearance as "Winners" among the 47 data sets (according to the standard deviation of the forecast errors), the new INTs consistently deliver superior results (40% and 36% of the cases appearing among the top two "Winners"), while Gompertz and Weibull are at the bottom of the list (with 21% and 13%, respectively).

We conclude that both in terms of long-term average performance and in terms of the probability of delivering best performance for a particular data set, the three-parameter NT_2-based model delivers forecast accuracy appreciably better than all other models.

Conclusions

Due to the wide use of the diffusion process as the underlying model for the

spread of technological innovation or the penetration of a new product into the market, forecasting S-shaped diffusion processes have been the subject of much research effort. In a comprehensive study by MI, twenty-nine different current models have been identified, and classified according to the behavior of the inflexion point of the S-shaped curve. Since S-shaped curve is characteristic also to most statistical distributions, the majority of these models are based on the selection of some known distribution. However, arbitrary selection of any of these models to produce forecasts (based on a given data set) has consistently proved to result in bad forecasts. On the other hand, identification procedures that would indicate the most appropriate model have proved to be evasive. This has motivated the use of either weighted or simple average of a set of forecasts derived from different models. MI suggested a scheme to generate weights for the combined forecast, based on membership probabilities obtained for a particular data set from discriminant analysis.

In this study we pursued a different approach, based on a new general methodology for empirical modeling. This approach produced INTs and NTs, which have consistently been shown to provide good representation to distributions belonging to any of the four classes defined in MI study. Thus, the need to identify which class a particular data set most probably belongs to is eliminated by the new approach. Furthermore, overall the new INTs (or NTs) provide better forecasts for S-shaped diffusion processes than any single existing model. Thus, our claim that they may be universally used, irrespective of which underlying statistical model is the "correct" model, has been corroborated.

References

1. Meade, N., T. Islam (1998). Technological forecasting-model selection model stability and combining models. *Management Science.* 44(8) 1115-1130.

2. Young, P., K. Ord (1989). Model selection and estimation for technological growth curves. *International J. Forecasting.* 5 501-513.

3. Shore, H. (2005). *Response Modeling Methodology- Empirical Modeling for Engineering and Science.* World Scientific Publishing Co. Ltd., Singapore.

4. Shore, H. (2004). Response Modeling Methodology (RMM)- Current

distributions, transformations and approximations as special cases of the RMM error distribution. *Communications in Statistics (Theory& Methods)*, 33(7), 1491-1510.

5. Box, G. E. P., D. R. Cox (1964). Analysis of transformations. *J. Royal Statistical Society* B, 26 211-243.

Appendix

MI distinguish between symmetric models (inflexion point at P_t=0.5, like the normal and the logistic models), non-symmetric models (P_t<0.5 but constant, like Gompertz and linearized Gompertz), and flexible models (where the inflexion point can occur anywhere within the same model, including P_t=0.5; Typical examples are the log-normal, Weibull and the extended logistic). MI denote these classes by classII_1 (symmetric), class II_2 (asymmetric) and Class II_3 (flexible). Based on a carefully selected subset of seven models that belong to the different classes (out of the original twenty-nine), MI introduce the following classification:

Class II_1: Simple Logistic, Mansfield (AR)

ClassII_2: Gompertz, Floyd (AR)

Class II_{3A}: Weibull, Extended Logistic (B)

Class II_{3B}: Cumulative Lognormal,

where the third class (II_3) is subdivided, based on results from stepwise discriminant-analysis, which showed that the lognormal "behaves differently from the other flexible curves" (MI[1], p. 1122).

These models are displayed below (all express P_t in terms of t):

1. Simple Logistic: $\frac{1}{1+c\cdot\exp(-bt)}+\varepsilon_t, b>0, c>0.$

2. Gompertz: $\exp(-c\cdot\exp(-bt))+\varepsilon_t, b>0, c>0.$

3. Cumulative Log-normal: $\int_0^t \frac{1}{y\sqrt{2\Pi}\sigma}\exp(-\frac{(\ln(y)-\mu)^2}{2\sigma^2})dy+\varepsilon_t, \sigma>0.$

4. Weibull: $1-\exp(-(\frac{t}{\alpha})^{\beta})+\varepsilon_t, \alpha>0, \beta>0.$

5. Extended Logistic (B): $\frac{1-c\frac{p}{q}\exp(-(p+q)t)}{1+c\exp(-(p+q)t)}+\varepsilon_t, p>0, q>0.$

6. Mansfield (AR): $P_t - P_{t-1} = bP_{t-1}(1-P_{t-1}) + \varepsilon_t, b > 0.$

7. Floyd (AR): $P_t - X_{t-1} = bX_{t-1}(1-X_{t-1})^2 + \varepsilon_t, b > 0$

Table 1. Measures I-III for goodness-of-fit comparisons (T_P is the response)

Model	I. μ_j(STD)	II. σ_j(STD)	III. μ_j(STD) σ_j(STD)
IN	0.8568	0.81	0.70237
1	2.0746	1.17	2.44359
2	1.6682	1.04	1.74412
3	1.5164	1.14	1.74165
4	1.3259	0.97	1.29179
5	1.0113	0.87	0.88462

Table 2. Measure IV for goodness-of-fit comparisons (T_P is the response)

Model	Winner (of 2 best models)	%	Winner (of 4 best models)	%
IN	45 (9	9	41	8
1	2	4	-	-
2	3	6	-	-
3	11	2	1	2
4	4	8	0	0
5	29	6	5	1

Table 3. Measures I-III for goodness-of-fit comparisons (P_t is the response)

Model	I.. μ_j(STD)	II. σ_j(STD)	III. μ_j(STD) σ_j(STD)
N	0.0225	0.0198	0.00045
N	0.0190	0.0132	0.00025
1	0.0302	0.0211	0.00064
2	0.0269	0.0180	0.00049
3	0.0272	0.0241	0.00066
4	0.0260	0.0210	0.00055
5	0.0233	0.0334	0.00078
6	0.0179	0.0131	0.00024
7	0.0163	0.0105	0.00017

Table 4. Measure IV for goodness-of-fit comparisons (P_t is the response). AR models excluded.

Model	First two winners (of	In percentage	The winner (of preceding	In percentage
N	12	0.255	-	-
N	20	0.425	9	0.191
1	-	-	-	-
2	-	-	-	-
3	17	0.361	9	0.191
4	15	0.319	4	0.085
5	30	0.638	25	0.531

Table 5. Measures I-III for goodness-of-forecast comparison (T_P is the response). "Correlation" refers to correlations between STDs of residuals from fitting and from forecasting, calculated for each model over 40 data sets.

Mo	Measure I	Measure II	Measure III	Correlation
I	2.7692	6.6308	18.362	0.1508
1	1.4733	0.8751	1.2893	0.6906
2	1.4976	1.1096	1.6618	0.4472
3	5.9149	12.659	74.877	0.3182
4	3.7847	8.2109	31.076	0.3925
5	1.7354	2.7189	4.7187	0.2397

Table 6. Measure IV for goodness-of-forecast comparison (T_P is the response). 40 data sets.

Model	First two	In percentage	The winner	In percentage
I	21	0.525	13	0.325
1	11	0.275	-	-
2	11	0.275	-	-
3	17	0.425	11	0.275
4	8	0.200	8	0.200
5	12	0.300	8	0.300

Table 7. Measures I-III for goodness-of-forecast comparison (T_P is the response). 38 data sets (excluding deviant samples 22, 41). "Correlation" refers to correlations between STDs of residuals from fitting and from forecasting, calculated over 38 data sets.

Mo	Measure I:	Measure II:	Measure III:	Correlation
I	1.2863	1.0816	1.3914	0.5296
1	1.4659	0.8945	1.3113	0.7739
2	1.4350	0.9755	1.3999	0.6919
3	4.3693	10.892	47.593	0.4213
4	2.9425	7.5149	22.113	0.4668
5	1.1579	0.9690	1.1221	0.5898

Table 8. Measures I-III for goodness-of-forecast comparison (T_X is the response). 47 data sets.

Mo	Measure I:	Measure II:	Measure III:	Correlation
I	2.2901	4.9251	11.279	0.1194
I	9.1510	40.608	371.61	0.0475
I	2.8922	5.7322	16.578	0.1877

Table 9. Measure IV for goodness-of-forecast comparison (T_P is the response). 40 data sets.

Mod	Fir	In	The	In	Fir	In	The	In
I	4	0	-	-	-	-		
I	1	0	1	0	1	0	1	0
I	6	0	-	-	-	-	-	-
I	1	0	7	0	-	-	-	-
1	9	0	9	0	2	0	9	0
2	7	0	-	-	1	0	7	0
3	1	0	4	0	2	0	-	-
4	7	0	-	-	8	0	-	-
5	8	0	8	0	1	0	1	0

Table 10. Measures I-III for goodness-of-forecast comparison (T_X is the response). 38 data sets (excluding also deviant sets 22, 41).

Mo	Measure I:	Measure II:	Measure III:	Correlation
I	1.3156	0.9092	1.1963	0.4104
I	2.7578	4.5741	12.615	0.5703
I	2.6566	4.4369	11.787	0.2788

Table 11. Measures I-III for goodness-of-forecast comparison (P_T is the response). 47 data sets.

Mo	Measure I:	Measure II:	Measure III:	Correlation
N	0.0531	0.0905	0.0048	-
N	0.0246	0.0185	0.0004	0.5636
1	0.0306	0.0184	0.0005	-
2	0.0259	0.0180	0.0004	-
3	0.0254	0.0227	0.0005	0.1345
4	0.0249	0.0176	0.0004	0.1562
5	0.0283	0.0321	0.0009	-

Table 12. Measure IV for goodness-of-forecast comparison (X_T is the response). 47 data sets.

M	First	In	The	In
N	19	0.4042	15	0.3191
N	17	0.3617	13	0.2766
1	14	0.2978	9	0.1914
2	10	0.2127	-	-
3	13	0.2766	-	-
4	6	0.1276	-	-
5	15	0.3191	10	0.2127