DELAYS AND AGGREGATION IN SYSTEM DYNAMICS MODEL

by

R. Joel Rahn
Faculté des Sciences de l'Administration
Université Laval, Québec

ABSTRACT

This paper focuses on the aggregation that is implicit in the use of distributed delays in dynamic models. The aggregation process relates the continuous time-dependent response of a delay structure to the underlying distribution of delay times of the disaggregated events which constitute the delay. The discussion covers in particular the special case of exponential delays used in system dynamics models.

TABLE OF CONTENTS

INTRODUCTION

One of the characteristic features of a system dynamics study is the emphasis placed on delays in the system. From one point of view, delays are non-events; something happens in a system and no change is observed immediately elsewhere, but after a delay or a time-lag a reaction occurs somewhere else in the system. For example, a salesman writes an order for one hundred units and after some period of time, a delay, the order is delivered to the customer. In this sense a delay means a time-interval between a causal action and the effect of that action.

Another aspect of delay structures that is frequently used in system dynamics modeling is the creation of average values of system variables that change with time. Typically, a special delay structure called SMOOTH is used to define an average over past values of a variable with the weight or relative contribution of the values declining continuously (exponentially) the further the value is in the past. Thus a sudden change in the variable being averaged is distributed over some characteristic time. The full impact is not felt immediately because the average does not change immediately to the new value of the input variable. The process of smoothing and the effects of this smoothing are given excellent treatment in Appendix E and chapter 9 of Industrial Dynamics (Forrester, 1961).

But in system dynamics studies more emphasis is placed on another aspect of so-called delay structures. (A delay structure is any combination of rates and levels in series). This aspect is the distortion or transformation that such a delay structure imposes on any input to the system. In the example of our salesman, it may well happen that all one hundred units ordered are not delivered on the same date-part of the order could be filled from the inventory, part from current production and perhaps part must be bought from

a competitor to meet a delivery deadline. Thus the response to the impulse given by the order for one hundred units may be the delivery of one hundred units distributed over several days depending on the source of the unit, method of transport etc. It is this kind of transformation of the original input rate of an order for one hundred units into a delivery rate stretching over several days that characterizes the delay structure of the system.

From the simple example of the salesman's order it is clear that the transformation aspect is an aggregation of discrete events. Each unit may have a serial number and a precise delay time may be associated with the delivery of each unit. But if all units are functionally equivalent, for example 100 television sets for a new hotel, it is useful to consider the aggregate result expressed in terms of the delivery rate following the order i.e. in terms of the number of TV sets delivered per day. The aggregation process converts a set of discrete delay events into a continuous flow rate of output in response to a flow rate of inputs.

The purpose of this paper is to describe the relationship between the first aspect of a delay as a discrete lag and the third aspect of delays as structures that transform the behaviour of input variables. A great deal of time is spent on the transformation aspect of delays when teaching system dynamics. It is hoped that a clearer perception of the nature of the aggregation process implicit in many uses of delays will aid in the conceptualization and comprehension of the relationship between the model and the system being modeled.

AGGREGATION OF DISCRETE EVENTS

Consider a typical discrete process as mailing a single letter, or ordering a single widget. We can imagine the process to consist of two events. First, a causal event such as placing the letter in a mailbox or asking a salesman for a widget. Second, a result event such as the arrival of the letter at its destination or the receipt of a widget from the retailer. These two events are separated by an interval of time, a delay or a time-lag of a specified length. Some types of discrete-event simulation deal with just such discrete, specific time-lags in a system and accumulate statistics based on repeated evaluation of the events occurring in a system in which the time-lags between events are random variables i.e. have values determined by a given probability distribution.

In system dynamics models one considers instead many such delay-events aggregated according to two principles. Namely, many similar events are assumed to be occuring simultaneously and independently. For example, many letters are posted each day, many orders are written each week. From these examples simultaneity is interpreted not in the strict physical sense of exact coincidence in time (not all letters are mailed at 11.32 a.m.), but in the sense that several events occur over a time interval, DT, that is short compared to the time-horizon of the model. In this case we speak of a flow of events (the number of letters mailed per day in a model with a horizon of months or years) as representing simultaneity. These events are also assumed to occur independently, that is, the occurrence of one event does not influence the occurrence or non-occurrence of another. There is no feedback between simultaneous events.

In the aggregation process, the lumping together of many causal events to form a flow of causal events or input leads to a lumping together of the result events to form a flow of results or output, modified by the distribution of values for the lags of the individual events. For example, the mailing rate (letters per day) results in a delivery rate influenced by the distri-

bution of the delay times of the individual letters aggregated into the mailing rate.

To make the nature of the influence of the distribution of delays more explicit, consider the following definition:

$C(t)$ = the inflow rate of letters at time t.

Then $C(t-\tau)$ = the inflow rate of letters at time $t-\tau$.

Let $p(\tau)d\tau$ = the probability that a letter is delivered between τ and $\tau+d$ time-units after it is mailed.

Consider the case when there are no letters in the mail system initially, then the delivery rate is just the sum of all the letters mailed at a previous time (that is, at time $t-\tau$) multiplied by the probability that these letters are delivered after a delay of τ time-units. That is the delivery rate at time t (the flow of results) is

$$R(t) = \int_0^\infty C(t-\tau)\, p(\tau)\, d\tau \qquad (1)$$

$$= \int_{-\infty}^{t} C(t')\, p(t-t')\, dt' \qquad (2)$$

Eq. (1) means that the aggregate result from a delay process with a distribution of delay times is a weighted sum over all past aggregate causal events where the weighting function is the probability distribution[1] of the delay times. The form of this result is well known to engineers.[2] However, the relationship between the distribution of discrete events (the delay times of each letter) and the aggregate delays are rarely emphasized. I have found this interpretation extremely useful when considering the justification for

1. In this paper (probability) distribution functions denote what many statisticians prefer to call frequency functions (McGee, 1971).

2. The above derivation was inspired by lecture notes prepared by T. Manetsch (Manetsch and Park, 1973).

using a particular delay form and when considering the relationship between difference equation and differential equation simulation models. On this latter point, it is interesting to note that Dhrymes (Dhrymes, 1971) in his magnum opus on distributed lags in econometric models, describes lag structures for which all co-efficients are non-negative and sum to one and refers to an interpretation of these structures in terms of discrete probability distributions. By the above argument this interpretation is elevated to a theorem: with $p(\tau)$ a discrete distribution of delay times, eq. (1) says that the output of such a system is a sum over past inputs at discrete times in the past weighted by the probability that such inputs are delayed by the appropriate discrete amounts. The importance of having non-negative lag co-efficients cannot be over-emphasized. Distributed lags with negative co-efficients do not represent distributed delays in the sense of this paper. Instead such structures are attempts to replace more complicated causal structures by "black-boxes" or transformations of the inputs.

Let us now examine the relationship between eq. (1) and first-order delays as they are used in system dynamics. The differential equation for a first-order delay is

$$\frac{dL}{dt} = C(t) - \frac{L}{AT} \tag{3}$$

The solution is

$$L(t) = L_o e^{-t/AT} + \int_o^t C(t')e^{-(t-t')/AT}dt' \tag{4}$$

where L_o is the initial content of the level. If $L_o = 0$, as we assumed in deriving eq. (1), the output rate is

$$\frac{L}{AT} = \int_o^t C(t')\frac{e^{-(t-t')/AT}dt'}{AT} \tag{5}$$

Referring to eq. (2) and noting that for a system that starts at time $t = 0$ we have $C(t') = 0$ for $t' < 0$, we get for eq. (2).

$$R(t) = \int_o^t C(t')\, p(t-t')dt' \tag{6}$$

Thus the probability distribution of delay times for a first-order delay is

$$P(t) = \frac{e^{-t/AT}}{AT} \tag{7}$$

In fact the first-order exponential delay is a member of a family of delays for which the distribution of delay times are the Erlang distributions, namely

$$P(t,\ AT;n) = \frac{t^{n-1}e^{-t/(AT/n)}}{(\frac{AT}{n})^n\ (n-1)!} \tag{8}$$

When $n = 3$ we have the distribution of delay times corresponding to a third-order delay. From eq. (6), if the input rate is a unit pulse at time zero we see that the output rate is

$$R(t) = p(t) \tag{9}$$

so that the shape of the distribution is evident as the response to a pulse.

Two generalizations of the delay structures available in DYNAMO suggest themselves immediately. First, we can consider the sum of several distributions of delay times, appropriately weighted so that the cumulative distribution function of the sum tends to one in the limit as the delay time goes to infinity. Thus one could have

$$p(t) = a_1 p(t,AT_1;3) + a_2 p(t,AT_2;3) \tag{10}$$

to represent a two-humped distribution if AT_1 and AT_2 are sufficiently different. Further in this line of thought, if empirical evidence of a particular delay distribution were available, one could attempt to fit the data with a combination of exponential delays of different orders and different average delay times. Non-linear estimation techniques would be required.[3]

Another generalization derives from eq. (3) where the differential equation of the first-order delay is given. There we see that a delay can be determined from a non-homogeneous differential equation for the distribution function with the input rate as the non-homogeneity. Thus if we have a differential equation for a distribution function, the corresponding aggregate delay is easily written. Unfortunately, distribution functions suitable for use with DYNAMO are not easy to find and in fact seem to be restricted to the Erlang family of distributions otherwise known as exponential delays.

C O N C L U S I O N

This paper has attempted to make explicit the relationship between a mass of separate events occuring at discrete times and the underlying probability distributions governing those events in order to justify the use of aggregate variables in simulation modeling. Although the formal structure of system dynamics models is deterministic, the probabilistic nature of many real systems is not thereby ignored. Rather certain statistical aspects are retained in delay structures and the question becomes one of verifying that the representation of the delay time distributions is accurate or at least satisfactory.[4]

---

3. Some idea of the problems involved even in the simple, case of one distribution are discussed by M. Hamilton in these proceedings.

4. This latter question is treated in detail in Hamilton's paper in these proceedings.

As a final remark, it will be noted that the relationship between delay structures and the underlying distribution of delay times has been developed for material delays only. For this type of delay, as the Markov process interpretation makes clear, the definition of an event as the change from one state to another is particularly easy to make. In the case of information delays, it may be inappropriate to speak of a delay event and the distribution of delay times since the information delay may actually represent a formal data-smoothing process used by a decision-maker as part of his policy or rate equation definition. Such a formal transformation of data has no direct, physical, relationship with an implicit distribution of delay events.

B I B L I O G R A P H Y

Dhrymes 1971 Distributed Lags: Problems of Estimation and Formulation, P.J. Dhrymes, Holden-Day, San Francisco, 1971.

Forrester 1961 Industrial Dynamics, J.W. Forrester, M.I.T. Press, Cambridge, Mass, 1961.

Manetsch and Park 1973 System Analysis and Simulation With Application to Economic and Social Systems, T.J. Manetsch, G.L. Park, Premilinary Edition, Dept. of Electrical Engineering and System Science, Michigan State University, East Lansing Mich., 1973.

McGee 1971 Principles of Statistics: Traditional and Bayesian, V.E. McGee, Appleton Century Crofts, New York, 1971.