CART: A Decision Support Model for Estimating Conditional Response Times

James M. Tien, Rensselaer Polytechnic Institute, Troy, New York 12181
Chai-Yi Chou, Bell Communications Research, Red Bank, New Jersey 07701 (1985)

ABSTRACT

There are many organizations which provide on-site services that must be scheduled on an adhoc or real-time basis. Increasingly, the dispatch or communication centers in these organizations are installing on-line, computer-aided dispatch (CAD) systems. Given timely information regarding the state of the system (e.g., how many response units are busy, how many calls are in queue awaiting service, etc.), together with appropriate assumptions concerning the call arrival and service patterns, a model-based, response time algorithm could be developed and coded in a CAD system. Such a computer-assisted response time (CART) algorithm is developed herein. The algorithm can be straightforwardly implemented in an existent CAD system, and would support the call-taker in making an appropriate decision regarding the response time to each call for service.

I. INTRODUCTION

There are many organizations which provide on-site services that must be scheduled on an adhoc or real-time basis; they include services in both the public (e.g., police, ambulance, fire, etc.) and private (e.g., delivery/pick-up, repair/maintenance, etc.) sectors. Increasingly, the dispatch or communication centers in these organizations are installing on-line, computer-aided dispatch (CAD) systems. Yet, in replying to a call for service, the call-taker in such a modernized center is still being non-specific in advising the caller as to when a response unit will be on-site to handle the call. Certainly, given timely information regarding the state of the system (e.g., how many response units are busy, how many calls are in queue awaiting service, etc.), together with appropriate assumptions concerning the call arrival and service patterns, a model-based, response time algorithm could be developed and coded in a CAD system. Thus, when a call arrives and is entered into the CAD system, a corresponding response time can be estimated; the call-taker can then transmit this estimate to the caller. Moreover, because calls can be generally classified as being critical (i.e., requiring an immediate response) or non-critical (i.e., not requiring an immediate response), formal dispatch procedures can also be identified for handling these two types of calls in an optimum manner, given the available resources. A family of such procedures is modeled in the computer-assisted response time (CART) algorithm that is developed herein. The algorithm can be straightforwardly implemented in an existent CAD system, and would support the call-taker in making an appropriate decision regarding the response time to each call for service.

In order to provide a context within which to view CART, we describe the problem that motivated this research -- a problem in the police dispatch area. Typically, as Tien and Valiante note [1979], citizens who call for police service are always being advised that a "patrol car will be right out", even though considerable delays may occur either because no patrol cars are available for dispatch, or because the few cars that are available are being reserved for dispatch to more crtical calls for service, or because the car that is assigned to the geographic sector in which the call originates is busy. Whatever the reason, citizens are being needlessly frustrated. Certainly, the frustration can be mitigated, if not eliminated, by formally advising citizens of potential delays. Indeed, because citizen satisfaction is a function of expectation [Kansas City Police Department, 1977; Tien et al., 1978; Tien and Valiante, 1979] and because some 86.1 percent of all calls for police service are non-critical in nature [Tien et al., 1978; Sumrall et al., 1980], a considerable portion of police demand can be "managed" and, more specifically, a formal delay procedure is one approach for managing such demand.

In 1976, the Wilmington Department of Police (WDP), Wilmington, Delaware, implemented a formal delay procedure; that is, when all patrol cars were busy, callers requesting service for a non-critical matter were told by the call-taker to expect a 30-minute delay. As an element of both the Wilmington split-force patrol experiment [Tien et al., 1978] and the Wilmington management of demand program [Cahn and Tien, 1981], this formal delay procedure was judged to be very effective; the citizens' attitude toward a delay -- of which they were formally advised -- is best summarized by one of the telephone survey respondents who said, "I am a taxpayer. If it helps to keep my taxes down, then I'm all for the police to take their time in showing up to non-emergency situations — but I would like to be told of such a delay so that I'm not waiting around for them" [Tien and Valiante, 1979, p. 23].

It should, however, be noted that Wilmington's formal delay procedure is fixed or static; that is, callers receiving a formal delay are each advised of the same constant delay -- a 30-minute delay. Certainly, this need not and should not be the case. As alluded to earlier, the expected delay for a non-critical call is variable and is dependent or conditioned on the state of the system, together with appropriate assumptions concerning the call arrival and service patterns. Thus, what is needed is a dynamic (i.e., state or queue dependent) procedure for delaying responses to non-critical calls. Moreover, the procedure must simultaneously satisfy two conflicting objectives: first, the probability of a critical or high priority call being delayed must be small, and, second, the delay of a non-critical or low priority call must not be excessive.

A family of such dynamic delay procedures are modeled and discussed in Section II, followed by a review of the corresponding algorithmic developments in Section III and simulation results in Section IV. Some concluding remarks are contained in Section V.

II. Model

There are, of course, several approaches to modeling a dynamic delay procedure that must simultaneously satisfy the two above stated objectives. However, after reviewing the literature and considering practical requirements of implementation, we decided on the following dynamic delay procedure. Simply stated, if, in terms of the police response environment, the call-taker receives a high priority (i.e., critical) call and at least one of the N total patrol cars is not busy, then the call-taker would inform the caller that a patrol car will respond with an expected (with, say, 95 percent confidence) response time equal to the expected travel time. If, on the other hand, there is no free patrol car, then the high priority (HP) call is queued in the HP queue and the caller is advised of an expected response time equal to the expected delay time in the HP queue and the expected travel time. If the call-taker receives a low priority (i.e., non-critical) call and n, the number of busy patrol cars, is less than some parameter R, then the call-taker would inform the caller that a patrol car will respond with an expected response time equal to the expected travel time.

if, on the other hand, $n \geq R$, then the low priority (LP) call is queued in the LP queue and the caller is advised of an expected response time equal to the expected delay time in the LP queue and the expected travel time. Whenever a patrol car becomes free, it attempts, first, to service the next HP call if there is any; second, if $n < R$ or if the LP queue length is greater than some parameter M, it attempts to serve the next LP call, if there is any; otherwise, it remains free until the next call arrives.

Notationally, the delay component of the above described procedure can be defined as D(N;R,M;Q,Q), where

- N = total number of patrol cars;
- R = cut-off for the number of busy patrol cars; if the number of busy patrol cars is equal to or greater than R, then only HP calls are served;
- M = cut-off for the number of calls in the LP queue; if the number of calls in the LP queue is greater than M, then the R cut-off does not apply and LP calls are served as long as there is no HP calls waiting to be served and as long as there is at least one free patrol car; and
- Q = designation that the calls are queued (and served on a first-come, first-served basis) — the first "Q" indicates that HP calls are queued, while the second "Q" indicates that LP calls are queued. (In some situations, an "L" designation is employed for the HP calls to reflect the fact that these calls are "lost" if they cannot be served immediately.)

It can be seen that the procedure is quite mindful of the need to have enough patrol cars available to respond to high priority calls (i.e., the R cut-off), while at the same time not allow the low priority calls to be queued for too long (i.e., the M cut-off). Additionally, we note that i) when $M = \infty$, the R cut-off is always in force; and ii) when $M = 0$, the R cut-off is always turned off and the procedure is equivalent to a procedure with $R = N$ and $M = \infty$. Moreover, the procedure actually describes a family of similarly structured procedures, depending on the particular values of the R and M parameters.

In an initial attempt to develop a tractable model of the procedure, as is the case in Section III, the following three assumptions can be made. First, the arrival of HP and LP calls are independent, homogeneous Poisson processes with average rates λ_1 and λ_2, respectively. Second, all calls within each priority are served on a first-come, first-served basis, with the HP calls being served before the LP calls. Third, each patrol car takes an exponentially distributed time -- with average $1/\mu$ — to serve a call. Fourth, each call requires only a single patrol car response. While the first two assumptions are quite appropriate [Larson, 1972; Taylor, 1976; Green, 1978], the latter two assumptions do not hold in general; that is, the service time is less random than an exponentially distributed random variable and calls for service do sometimes require a multiple car response. These two assumptions are appropriately relaxed in Section IV.

Not surprisingly, the related models in the literature are, for the most part, also based on the four stated assumptions. For the sake of brevity, Exhibit 1 contains a summary of our literature review; it focuses on four distributions across three sets (i.e., no cut-off, one cut-off, and two cut-off) of queueing models, with each set containing both the situations of HP loss and HP queue. Two important points should be made concerning Exhibit 1. First, although four distributions are considered in the exhibit, the real purpose of our effort is to obtain the first distribution (i.e., distribution of conditional delay times); nevertheless, for purpose of validation, we also obtain the second distribution (i.e., distribution of steady state probabilities) in order to derive the third distribution (i.e., distribution of unconditional delay times), which can likewise be directly obtained — through a transform relationship of Little's [1961] formula -- from the fourth distribution (i.e., distribution of busy servers). The fact is, however, that, except in the simple situation of no cut-off, the literature does not address conditional delay times; instead, the literature concentrates on unconditional delay times, which is the reason we also obtain these times -- for purpose of validation. Second, Exhibit 1 also indicates that the literature is devoid of any reference to the D(N;R,M;Q,Q) model. In sum, the focus of our research has not been dealt with in the literature; consequently, our results are not only valuable from a decision support perspective, but also constitute a contribution to the queueing literature.

Model	Distribution of Conditional Delay Times	Distribution of Steady State Probabilities	Distribution of Unconditional Delay Times	Distribution of Busy Servers
D(N;N,∞;L,Q): No cut-off with HP loss	*	*	*	*
D(N;N,∞;Q,Q): No cut-off with HP queue	Drossin & Raich [1957].		Cobham [1954]: Only the expected value. Drossin & Raich [1957]: Only for N-1. Davis [1966].	
D(N;R,∞;L,Q): One cut-off with HP loss		Sonick & Jackson [1973]: Only on an algorithmic basis.	Taylor & Templeton [1980]:	Jaiswal [1968]. Taylor & Templeton [1980].
D(N;R,∞;Q,Q): One cut-off with HP queue			Taylor & Templeton [1980]: Only the transform.	Jaiswal [1968]. Taylor & Templeton [1980].
D(N;R,M;L,Q): Two cut-off with HP loss			Taylor & Templeton [1976]: Only the expected value for R=N-1.	Taylor & Templeton [1976]: Only for R=N-1.
D(N;R,M;Q,Q): Two cut-off with HP queue				

Exhibit 1. Summary of Literature Review

III. ALGORITHMS

The state of the D(N;R,M;Q,Q) model can be denoted by three variables; that is, (n,h,ℓ), where n is the number of busy servers in the system, h is the HP queue length, and ℓ is the LP queue length. The conditional delay times for the HP calls or customers in the D(N;R,M;Q,Q) models are exactly the same as those in the $D(N;R,\infty;Q,Q)$ model. More specifically, because of the prioritized first-come, first-served queue discipline, the conditional delay time distribution of the kth customer in the HP queue at state (N,h,ℓ) is Erlang of order k distributed with scale parameter equal to $N\mu$. (Obviously, if $n < N$, then there would be no HP queue and an arriving HP call would experience no delay.)

The LP conditional delay times are considerable more difficult to obtain. Indeed, these times cannot be analytically derived; instead, we have been able to develop several numerical algorithms to approximate the first and second moments of these times, as well as the system-wide LP unconditional delay time. These algorithms are quite involved and lengthy; they are not contained herein but may be in found in Chou [1984]. In brief, the first algorithm includes 22 steps and is focused on finding $LD(n,h,\ell;k;s)$, the Laplace or s transform of the conditional delay time of the kth customer in the LP queue at state (n,h,ℓ). Several

insights contributed to the development of this key algorithm. First, our scrutiny of the underlying state transition diagram of the D(N;R,M;Q,Q) model helped us both to group similar states and to sequence the steps of the algorithm. Second, we recognized that one and only one of the following three events could occur at any instant in time: i) a server completes a service and becomes available for assignment; ii) an HP customer arrives; or iii) an LP customer arrives. Third, we noted that, unlike the one cut-off situation, an arriving LP customer may have an effect on the conditional delay times of those customers who are already in the LP queue. More specifically, an arriving LP customer would have i) an effect on the conditional delay times of all customers in the LP queue if $\ell \leq M$; ii) no effect on the first $(\ell-M)$ customers in the LP queue if $\ell > M$; and iii) and effect on the remaining M customers (i.e., the $(\ell-M+1)$st to the th customer) in the LP queue if $\ell > M$. Consequently, if $D(n,h,\ell;k)$ denotes the conditional delay time distribution of the kth customer in the LP queue at state (n,h,ℓ), then

$$D(N,h,\ell;k) = D(N,h,M+k;k) \text{ for } h = 0,1,2,\ldots; \quad \ell = M+k, M+k+1,\ldots; \quad k = 1,2,\ldots \tag{1}$$

Fourth, we showed that the following property holds for $1 \leq k \leq \ell$:

$$LD(N,h,\ell;k;s) - W(s)\, LD(N,h-1,\ell;k;s) \to 0 \text{ uniformly for all } s\varepsilon[0,1] \text{ as } h \to \infty, \tag{2}$$

where

$$W(s) = [(s + N\mu + \lambda_1) - (s + N\mu + \lambda_1)^2 - 4N\mu\lambda_1]/2\,\lambda_1 \tag{3}$$

This property provided a means to approximate the infinite state structure of the D(N;R,M;Q,Q) model.

The Laplace transform algorithm served as a basis for the subsequent development of algorithms for computing $ED(n,h,\ell;k)$, the expected value of the conditional delay time of the kth customer in the LP queue at state (n,h,ℓ), and $SD(n,h,\ell;k)$, the standard deviation of the conditional delay time of the kth customer in the LP queue, given the system is at state (n,h,ℓ). In order to illustrate these algorithms, the ED and SD values are noted in Exhibit 2 for the D(25;22,3;Q,Q) model. A number of interesting observations emerge from Exhibit 2. First, as expected, the busier the system, the longer the LP delay. For example, $ED(23,0,1;1) = 2.62 > ED(22,0,1;1) = 1.42$. Second, surprisingly, the ED and SD values are dependent not only on the number of busy servers and the queue position but also on the LP queue length. In particular, the longer the LP queue, the shorter the LP delay; this is because of the likelihood of the M cut-off being triggered when the LP queue length is longer. For example, $ED(22,0,3;1) = 0.91 < ED(22,0,2;1) = 1.27 < ED(22,0,1;1) = 1.42$. Third, in contrast to the previous observations and as indicated in (1), whenever there are at least M -- in this case, 3 -- customers behind the kth customer in the LP queue, a longer LP queue length would not affect the customer's conditional delay time. Fourth, again surprisingly, the coefficient of variation of the conditional delay time of the kth customer in the LP queue decreases as the system gets busier but increases as the LP queue length increases.

Finally, a 29-step numerical algorithm has been developed to approximate $p(n,h,\ell)$, the steady state probabilities. In turn, the system-wide, expected unconditional delay time can be obtained from the following expression:

Expected Unconditional Delay Time =

$$\sum_{\ell=0}^{M-1} \sum_{n=R}^{N-1} p(n,h,\ell)\, ED(n,h,\ell+1;\ell+1) + \sum_{n=R}^{N-1} p(n,h,M)\, ED(n+1,h,M;M) + \sum_{\ell=0}^{K} \sum_{h=0}^{H} p(N,h,\ell)\, ED(N,h,\ell+1;\ell+1) \tag{4}$$

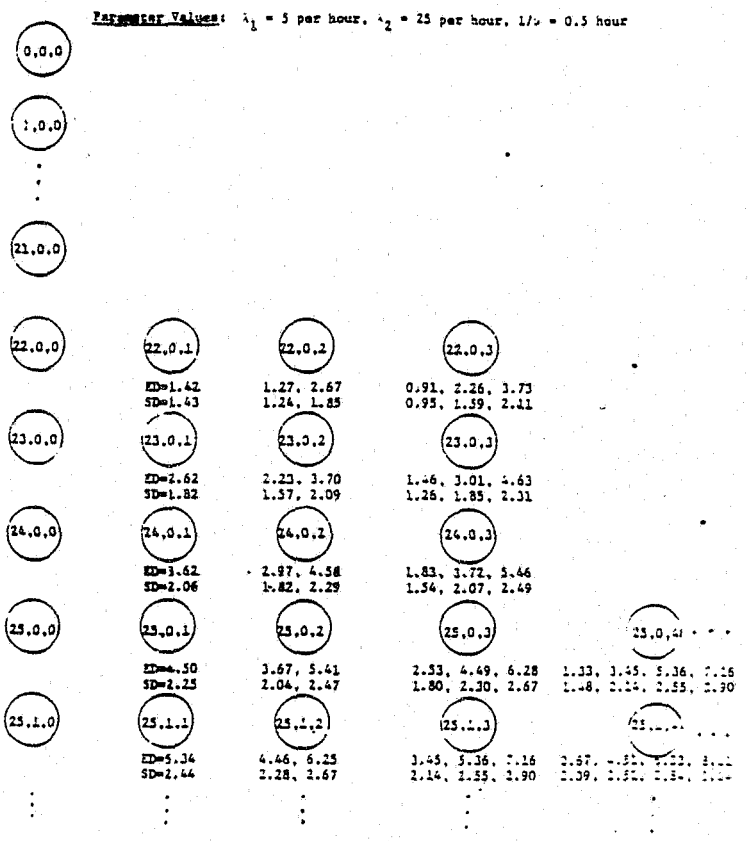

Exhibit 2. D(25;22,3;Q,Q): Expected Value and Standard Deviation of Low Priority Conditional Delay Times (In Minutes)

Our approximate numerical algorithms have at least been partially validated by comparing the results using (4) with the exact solutions obtained for the special case of M=1 — that is, for the D(N;R,1;Q,Q) model. Our approach to finding the exact solution of the LP unconditional delay time for D(N;R,1;Q,Q) has been, first, to define the probability generating function over the LP queue length; second, to solve these probability generating functions for the distribution of busy servers; third, to use the distribution of busy servers to compute the expected LP queue length; and, fourth, to use Little's [1961] formula to compute the expected value of the LP unconditional delay time. Actually, this is a typical approach to obtaining the expected unconditional delay, as documented in the queuing literature. Such an approach, however, cannot be successfully applied to the general D(N;R,M;Q,Q) model.

IV. SIMULATION

Our approximate numerical algorithms have been further validated by the results of our simulation analysis, which has employed the General Purpose Simulation System (GPSS) language. We have simulated the transitions in the D(N;R,M;Q,Q) model using a second as the basic time unit of simulation. Six runs, each run for 10 days and using different sets of initialization seeds for the HP and LP customer arrival and service completion processes, have been made. Unfortunately, since the GPSS output does not summarize

the conditional delay time statistics in terms of (n,h,ℓ), the state variables, and k, the queue position of the LP customer, we have had to develop a major FORTRAN program to collect from the GPSS all the individual simulated results and then to summarize the results appropriately.

Our simulation analysis has also been used to relax the service time assumption. Several researchers have shown that the service time expended on a call is much less random than that suggested by an exponentially distributed random variable [Taylor, 1976; Tien et al., 1978; Cahn and Tien, 1981]. Indeed, in the police environment, the service time is almost constant at about 25 minutes; in actuality, most calls take less time to serve, but the patrol officer would tend to take a "deserved break" in the remaining time before calling the radio dispatcher to report the completion of service. In comparing the simulation results and as expected, the ED and SD values for the LP conditional delay time are significantly larger in the constant service time situation than the corresponding values in the exponential service time situation.

Similarly, our simulation analysis has been extended to allow for the fact that a call for service may require the assistance of more than one response unit [Tien et al., 1977; Green, 1980; Green, 1981; Green and Kolesar, 1984]. Furthermore, the first or "primary" response unit has a longer service time than the backup or "assist" unit(s), as the latter unit(s) could leave the scene as soon as the incident is under control. Interestingly and as illustrated in Exhibit 3, when we used as input to our multiple-response simulation the empirically obtained data by Tien et al. [1977] — in which i) the distribution of response unit requirements (for the police environment) is such that 74.2% of all customers or calls require 1 unit, 18.2% require 2 units, and 7.6% require 3 units, and ii) the service times of both primary and assist units are approximately constant at 24 and 15 minutes, respectively — we found that the resultant conditional delay time statistics are, for the most part, somewhat comparable to the corresponding results obtained by our numerical algorithms (which consider the single-response, exponential service time situation). This somewhat surprising result can be explained by the fact that while the multiple-response requirement tends to inflate the LP conditional delay times, the constant service times of the multiple responders tend to deflate those same statistics, with the net impact being somewhat of a cancellation of the two effects.

Finally, our simulation analysis has been used to obtain the underlying distribution of the conditional response times in a multiple-response environment, assuming that the travel time is Erlang of order 2 distributed [Tien et al., 1977] and independent of the delay time. In particular, we have sought to obtain a conservative estimate of the conditional response time so that the probability that the response is within the estimate is 0.95. If the conditional response times were normally distributed, then the 0.95 quantile would be at a distance of 1.65 times the standard deviation to the right of the expected value. However, because the response time distributions are skewed to the right, we have found that a factor of 1.95 is more appropropriate; that is,

$$R_{0.95} = (ED + ET) + 1.95\sqrt{VD + VT} \qquad (5)$$

where $R_{0.95}$ = 0.95 quantile of conditional response time
ED = Expected value of conditional delay time
ET = Expected value of travel time
VD = Variance of conditional delay time
VT = Variance of travel time

Parameter Values: $\lambda_1 = 5.81$ per hour, $\lambda_2 = 36$ per hour, $1/\mu_B = 0.4$ hour, $1/\mu_A = 0.25$ hour
Server Requirements: 74.2% required 1 server, 18.2% require 2 servers, and 7.6% require 3 servers

	Low Priority Conditional Delay Time Statistics (In Seconds)						
	A. Algorithmic (Single-Response, One Exponential Service Time: $1/\mu_B$)		B. Simulated (Multiple-Response, Two Constant Service Times: $1/\mu_B, 1/\mu_A$)			Relative Difference (A-B)/A x 100%	
State (n,h,ℓ;k)	Expected Value	Standard Deviation	Expected Value	Standard Deviation	Number of Observations	Expected Value	Standard Deviation
(23,0,1;1)	68	72	73	74	8131	-7.4%	-2.8%
(23,0,2;1)	67	69	69	70	6922	-3.0%	-1.4%
(23,0,3;1)	64	63	64	65	3641	0.0%	-3.2%
(23,0,4;1)	57	53	58	57	4627	-1.8%	-3.6%
(23,0,5;1)	40	42	47	56	3533	-17.5%	-33.3%
(23,0,2;2)	134	97	148	101	6922	-10.4%	-4.1%
(23,0,3;2)	129	91	139	95	3641	-7.8%	-4.4%
(23,0,4;2)	119	82	128	85	4627	-7.6%	-3.7%
(23,0,5,2)	101	71	118	86	3533	-16.8%	-21.1%
(23,0,3;3)	196	113	217	116	3641	-10.7%	-2.7%
(23,0,4;3)	186	105	202	105	4627	-8.6%	0.0%
(23,0,5;3)	169	95	192	107	3533	-13.6%	-12.6%
(23,0,4;4)	254	124	278	122	4627	-9.4%	1.6%
(23,0,5;4)	239	115	267	123	3533	-11.7%	-7.0%
(23,0,5;5)	309	133	345	134	3533	-11.7%	-0.8%
(24,0,1;1)	131	93	127	88	1949	3.1%	5.4%
(24,0,2;1)	126	89	118	81	1450	6.3%	9.0%
(24,0,3;1)	117	81	112	75	1402	4.3%	7.4%
(24,0,4;1)	100	70	94	70	1216	6.0%	0.0%
(24,0,5;1)	65	57	77	70	3039	-18.5%	-22.8%
(24,0,2;2)	193	122	195	108	1450	-1.0%	11.6%
(24,0,3;2)	184	104	187	99	1402	-1.6%	4.8%
(24,0,4;2)	168	94	167	91	1216	0.6%	3.2%
(24,0,5;2)	142	85	157	95	3839	-10.6%	-11.8%
(24,0,3;3)	252	126	264	114	1402	-4.8%	9.5%
(24,0,4;3)	238	115	243	110	1216	-2.1%	4.3%
(24,0,5;3)	215	106	237	114	3839	-10.2%	-7.5%
(24,0,4;4)	307	133	316	126	1216	-2.9%	5.3%
(24,0,5;4)	287	125	313	128	3839	-9.0%	-2.4%
(24,0,5;5)	359	142	394	142	3839	-9.7%	0.0%

Exhibit 3. D(25;23,5;Q,Q): Comparison of Algorithmic Single-Response and Simulated Multiple-Response Results

V. CONCLUSION

The above sections describe the essential elements of our computer-assisted response time (CART) algorithm, while Exhibit 4 summarizes the algorithm. As depicted in the exhibit, at the beginning of each tour, a dispatcher enters into CART the delay-related and travel-related parameter values. Then, CART computes and stores the expected values and variances of the HP and LP conditional delay times. When a call arrives at, say, state (n',h',ℓ'), the call-taker determines the priority of the call and enters it into CART, which then retrieves the corresponding HP or LP conditional delay time, as well as travel time, statistics from its memory bank and computes the $R_{0.95}$ value using (5). This value is then provided to the call-taker who, in turn, advises the caller.

As indicated in Exhibit 4, the module for computing the expected value and variance of the HP conditional delay times is straightforward. More specifically, the HP conditional delay time of the jth customer in the HP queue is Erlang of order j distributed with scale parameter equal to Nμ. Hence, the expected value and variance are $j/N\mu$ and $j/(N\mu)^2$, respectively. The modules for computing the expected value and variance of the LP conditional delay times are those outlined in Section II. However, an efficient mapping technique can be employed so that only the first M LP conditional delay times need to be calculated. In particular, if we denote D(N,h,k;k) as the conditional delay time distribution of the kth customer in the LP queue state (N,h,k) and if k is greater than M by an amount i, we can think of the first i LP

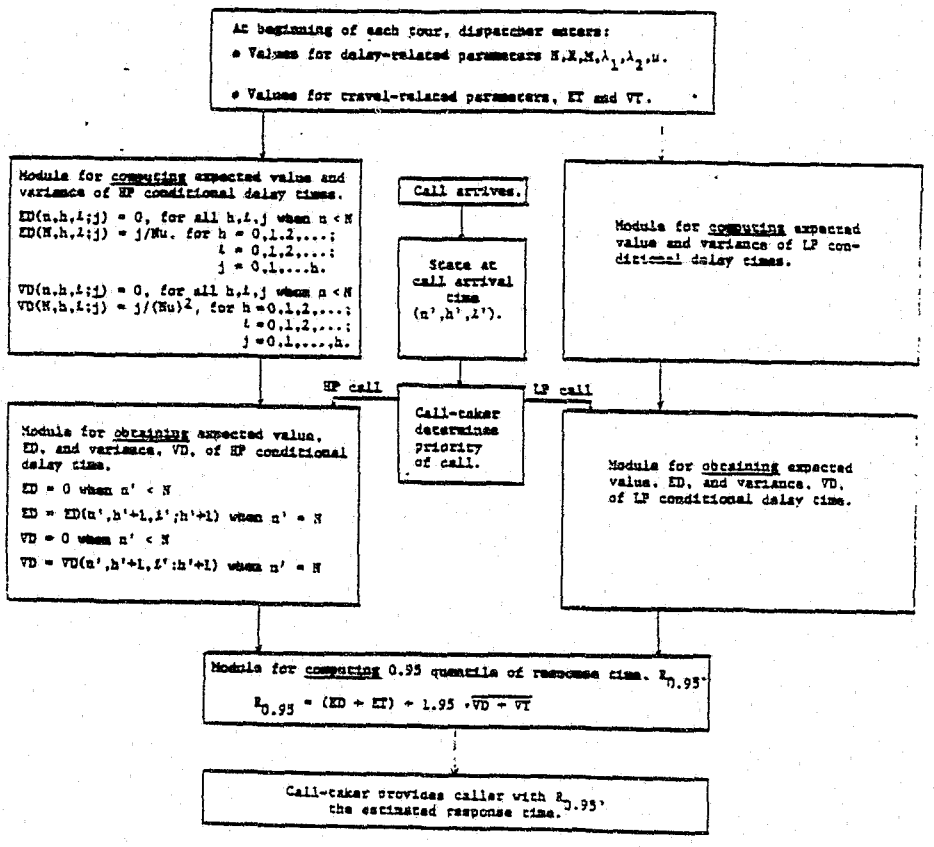

Exhibit 4. A Computer-Assisted Response Time (CART) Algorithm

customers as HP customers. The reason is because when any of these i LP customers takes the first queue position in the LP queue, there are definnitely more than M customers remaining in the LP queue, and this implies that the R cut-off is still turned off by the M cut-off. Hence,

$$D(N,h,k;k) = D(N,h+i,k-i;k-i) = D(N,h+i,M;M), \quad \text{for } k = M+i \qquad (6)$$

This efficient mapping technique can have a significant impact on both computation time and memory space. For example, in the case of the D(25;23,10;Q,Q) model, less than 6,000 — instead of some 25,000 — real values need to be stored for figuring out the delay time statistics for the first 50 LP customers. In sum, CART can be implemented on a microcomputer with no more than 48K bytes of memory.

Two additional remarks should be made regarding CART. First, although not indicated in Exhibit 4, CART can provide an updated response time to any caller who calls back to ask for an updated response time. Based on the name of the caller, CART can identify the caller's position in the queue, so that an updated response time can be appropriately obtained. Second, again although not indicated in Exhibit 4, CART can keep track of the response time provided to each LP caller, and, whenever it seems like a response time might be exceeded, CART can change the call's priority to that of an HP so that it can be more immediately dispatched. Inasmuch as the initial response time provided to the caller is a 0.95 quantile estimate, we can expect that such priority changes would occur with no more than 5% of all LP calls.

In terms of future research, one possible extension to the D(N;R,M;Q,Q) model is for M to be a function of n, the number of busy servers. That is, we can define an increasing function M(n) (for n = 0,1,...,N) such that if ℓ, the number of customers in the LP queue, is greater than M(n) and less than or equal to M(n+1), then (N-n) — instead of a fixed (N-R) — servers are set aside for HP customers. In this extension, the LP queue length builds up gradually instead of suddenly at n=R, which is the case considered herein.

Another effort that can be undertaken concerns further calibrating our approximate numerical algorithms so that the results would more closely correspond to situations with service times that are less random than exponential and calls for service which require multiple responses. One could either appropriately modify the values of the input parameter values — as considered by Green and Kolesar [1984] — or change the basic mechanics of the algorithms.

Finally, it should again be stated that although this effort is motivated in the police environment, there are similar situations in other response-oriented systems that would benefit from the results reported in this paper.

REFERENCES

[1] M.F. Cahn and J.M. Tien, An Alternative Approach in Police Response: The Wilmington Management of Demand Program, Cambridge, MA: Public Systems Evaluation, Inc., March 1981.

[2] C.Y. Chou, Dynamic Queue-Dependent Dispatching Procedures, Ph.D. Dissertation, Rensselaer Polytechnic Institute, 1984.

[3] A. Cobham, "Priority Assignment in Waiting Line Problems", Operations Research, 2, pp. 70-76, 1954.

[4] R.H. Davis, "Waiting Time Distribution of a Multi-Server Priority Queueing System", Operations Research, 14, pp. 133-136, 1966.

[5] S.A. Dressin and E. Reich, "Priority Assignment on a Waiting Line", Quart. Appl. Math., 15, pp. 208-211, 1957.

[6] L. Green, "Queues Which Allow a Random Number of Servers Per Customer", Ph.D. Dissertation, Yale University, 1978.

[7] L. Green, "A Queuing System in Which Customers Requires a Random Number of Servers", Operations Research, Vol. 28, pp. 1335-1346, 1980.

[8] L. Green, "Comparing Operations Characteristics of Queues in Which Customers Require a Random Number of Servers", Management Science, 27, pp. 65-74, 1981.

[9] L. Green and P. Kolesar, "A Comparison of the Multiple Dispatch and M/M/C Priority Queuing Models of Police Patrol", Management Science, 30, pp. 665-670, 1984.

[10] N.K. Jaiswal, Priority Queues, New York, NY: Academic Press, 1968.

[11] Kansas City Police Department, Response Time Analysis, Washington, D.C.: U.S. Government Printing Office, 1977.

[12] R.C. Larson, Urban Police Patrol Analysis, MIT Press, 1972.

[13] J.D.C. Little, "A Proof for the Queuing Formula L $= \lambda W$", Operations Research, 9, pp. 383-387, 1961.

[14] W. Shonick and J.R. Jackson, "An Improved Stochastic Model for Occupancy-Related Random Variables in General-Acute Hospitals", Operations Research, 21, pp. 952-965, 1973.

[15] R. Sumrall et al., Alternative Strategies for Responding to Police Calls for Service, Birmingham, Alabama: Birmingham Police Department, Unpublished Report, 1980.

[16] I.D.S. Taylor, A Priority Queuing Model to Measure the Performance in the Ontario Ambulance System, Ph.D. Dissertation, University of Toronto, 1976.

[17] I.D.S. Taylor and J.G.C. Templeton, "Multi-Server Priority Queues with Modified Cut-Off Queue Discipline", University of Toronto, Department of Industrial Engineering, Working Paper #76-018, October 1976.

[18] I.D.S. and Templeton, J.G.C., "Waiting Time in a Multi-Server Cut-Off Priority Queue and Its Application to an Urban Ambulance Service", Operations Research, 28, pp. 1168-1188, 1980.

[19] J.M. Tien, J.W. Simon and R.C. Larson, An Alternative Approach in Police Patrol: The Wilmington Split-Force Experiment, Washington, D.C.: U.S. Government Printing Office, No. 027-000-0068-0, April 1978.

[20] J.M. Tien and Valiante, N.M., "A Case for Formally Delaying Non-Critical Calls for Service", The Police Chief, March 1979.