A PREDICTIVE MODEL FOR THE POLICE RESPONSE FUNCTION

Deepak Bammi
Systems Engineering Department, University of Illinois at Chicago Circle, Chicago, Illinois 60680

and

Nick T. Thomopoulos
Industrial and Systems Engineering Department, Illinois Institute of Technology, Chicago, Illinois 60616

ABSTRACT

The expected response time to a call for service (CFS) for a given configuration of police beats is developed. The effect of downtime calls on the response time to a CFS is determined. Consideration is given to both travel time and waiting time. Travel time and service time distributions are isolated. The model is valid for Poisson arrivals and arbitrary service time distributions. A probabilistic assignment policy is determined for each beat. The fraction of incoming calls arriving in beat k answered by unit ℓ is obtained. Pre-emptive priorities are allowed. Application to the Aurora, Illinois, Police Department is shown.

INTRODUCTION

For the purpose of law enforcement, the city is divided into a number of police districts. A district in turn is divided into a number of beats. A beat is an area within a district to which a patrol unit is assigned. Calls for police service are telephoned into the communication center at police headquarters. If the patrol unit of the beat of occurrence of call is available, it is dispatched to answer the call. If it is unavailable, a unit from an adjoining beat answers the call. After the completion of an out-of-beat assignment the patrol unit returns to its beat. When not answering calls for service, the unit patrols the beat. A patrol unit may be unavailable for dispatching if it is presently servicing a call, or if it is off duty for administrative or personal reasons.

CRITERIA FOR DESIGNING BEATS

The International City Manager's Association[1] classified objectives of the patrol division under six headings: (1) prevention of crime, (2) suppression of criminal activity, (3) apprehension of criminals, (4) preservation of the peace, (5) regulation of conduct (non-criminal), and (6) protection of life and property. The criteria to be chosen for designing beats should have a high measure of effectiveness with respect to these six objectives.

Probability of arrest seems to be inversely related to response time in the relevant range. In a study conducted by the Los Angeles Police Department[2] it was found that when response time was 1 minute, 62 percent of the cases resulted in arrest; whereas, when all cases with response time under 14 minutes were groups together, only 44 percent led to arrest. Arrest probability as a function of response time is plotted in Figure 1.

It is proposed that patrol beats of a police department be designed to minimize response time of the patrol units. Minimization of response time should result in higher probability of arrest as shown in Figure 1. Assuming that the conditional probability of being convicted given that a citizen is arrested is unchanged, the probability of a criminal being convicted increases with the minimization of response time. Actually, the conditional probability of being convicted given that a citizen is arrested is likely to increase with reduced response time because of being able to gather more evidence with quick arrival. An increased probability of being convicted reduces the utility of committing a crime to a potential criminal. Thus, the minimization of response time results in an increase in the prevention and suppression of criminal activities. Peace is preserved by preventing crimes, by quick arrival of police at the location of crime, and by arresting criminals. Regulation of non-criminal conduct should also be improved by more rapid response to calls. Life and property have an increased degree of protection when a reduction in response time takes place. Minimization of response time, thus, satisfies the six objectives listed

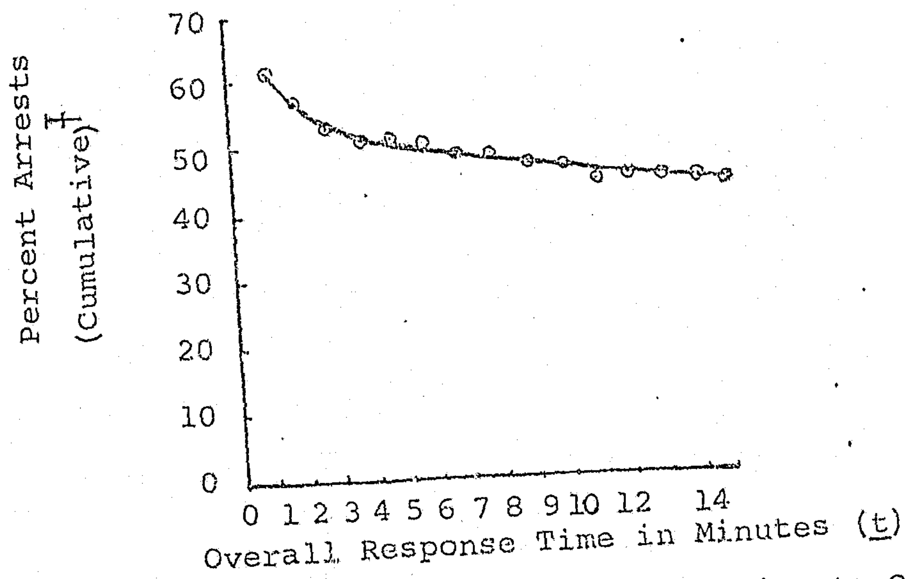

Figure 1. Percent of Arrest in Relation to Overall Response Time+

+ from Science and Technology[2].

ǂ Percent Arrests (Cumulative)

$$= 100 \times \frac{\text{Number of Arrests}}{\text{Number of Cases with Response time less than } t}$$

by the International City Managers' Association and also reduces crime disutility to the citizen. As suggested by Smith[3] response time has the additional advantage of being policy sensitive. That is, it is directly affected by decisions on the size and distribution of the patrol force.

Response time is the time elapsed from when need for police service arises until a patrol unit arrives at the location of the call. It is composed of (1) delay in reporting the incident to the communication center, (2) delay in the communication center in filling a report and in waiting for a patrol unit if all units in the district are unavailable, and (3) the travel time of the patrol unit from its present location to the scene of the incident. Delay in reporting incidents of crime to the police could be improved by strategic location of telephones, the ability to call the police without having to deposit a coin, and by greater cooperation by the citizenry.

In this paper it is assumed that we have no control over the delay in reporting incidents to the communication center. We also assume no control over the time spent in filling reports at the communication center.

If a call for service occurs when all patrol units in the district are unavailable, then there is a waiting time at the communication center. This waiting time is a function of how soon units become available again after an assignment. It is assumed that the service time at the scene of incident does not depend on the configuration of beats.

The fraction of the total response time that is due travel time is a function of the average service time, the number of units deployed and the geography of the city. This model was based on information available from the City of Aurora, Illinois. Aurora has a population of about 80,000 and is fifty miles from Chicago. The average service time for calls for service was 17.4 minutes. For sixteen patrol units deployed during the busiest shift an average travel time of 2.0042 minutes was noted when the average response time was 3.4915 minutes. Thus the travel time was 57 percent of the total response time and certainly warranted inclusion in the objective function. Further, travel time for the sixteen patrol units varied from a low of 0.9352 minutes to a high of 4.9548 minutes with ten of sixteen units having travel time within 30 percent of the average.

In a large city like Chicago, Nilsson[4] reported a service time of about 40 minutes. The higher service time may tend to make patrol units busier than those in Aurora (unless beats and their arrival rates are proportionately reduced in size). For busier units the average response time would be higher and the percentage of total response time that is due to travel time may be less than the 57 percent observed in Aurora. This would reduce the importance of travel time in the objective function but we feel that it would still be meaningful to include it in the optimization.

As reported by Smith[3], using minimum average response time as the objective function raises certain problems.

Calls for service do not tend to be distributed evenly over a city, but rather are usually heavily concentrated in certain areas. Minimizing average response time would lead to a heavy concentration of patrol units in heavy-crime areas and sparse deployment in low-crime areas. This could result in unacceptably long response times for calls from low-crime areas. Furthermore, this could lead to a rise in crime in these previously low-crime areas as criminals would probably shift their activities to areas which they find offer less risk of arrest.

In order to protect against these kinds of results, a constraint was added to the objective. Ideally, we would like the constraint to be that maximum response time will not exceed a specified upper limit anywhere in the city, but this proved to add very substantially to the cost of computation. Therefore, we substituted the constraint that maximum travel time will not exceed a specified upper limit anywhere in the city. This accomplishes almost the same result, as response time in low-crime areas tends to be largely travel time. This constraint is treated by making sure that for every beat the travel time between the centers of any two nodes does not exceed the pre-defined maximum.

So the objective for the predictive and optimization models for the police response function is to minimize average response time throughout the city and in all time periods, subject to a constraint that maximum travel time will nowhere exceed a specified upper limit.

PREVIOUS PREDICTIVE MODELS

Larson[5] developed a number of quantitative models for use in the allocation of police patrol forces. He wrote a simulator in the MAD (Michigan Algorithm Decoder) language. Larson's first model determined the probability law for travel distances to an incident in a beat and the corresponding optimal beat design on the assumptions that calls for service (CFS) and car location are independent and are uniformly distributed over the beat. He also assumed that the unit is always available.

In his second model, Larson considered an infinitely large command comprised of square beats, each of unit area. He assumed a "strict center of mass" dispatching strategy in which the unit is assumed to be at the center of its beat and the call is assumed to be at the center of the beat of occurrence. The dispatching strategy is then to choose that available unit with the minimum total travel distance.

After defining deterministic and probabilistic assignment policies and determining some state probabilities, Larson concludes that a model involving queueing considerations for N servers is difficult to solve.

Next, he finds an approximate solution for a finite command with the following additional assumptions: (1) demands for service are generated within the command by a simple Poisson process with parameter λ_c demands per hour, (2) average total time to service a call = $(1/\mu_c)$, (3) the "busy" probability of each patrol unit is approximated to be independent of the state (busy or patrolling) of every other patrol unit. The busy probability of each of the N_c patrol units is $\rho_c = \lambda_c/(N_c\mu_c)$, and (4) the probability that a queue of waiting calls will form is very small, and that either the beat car associated with the incident or at least one car in the four contiguous beats is always available for dispatch.

A dynamic programming model is developed to assign patrol units to geographically distinct commands by minimizing achievable delay cost per hour. The assumptions of the priority queueing model used are (1) Poisson arrivals, (2) negative exponential services (same service rate for all priority classes), (3) first-come, first-served queue discipline within each priority class, and (4) no pre-emption. Application to the New York City Police Department is shown.

Overlapping beats are explored in a system where car positions are known exactly. It is shown that the expected travel time in such a system is about the same as in a dispatching system with mutually exclusive beats and no car position information. In a previous model Larson showed that perfect car position information reduces travel time by 10 to 20 percent. It could be inferred, then, that for the same dispatching system overlapping beats involve larger travel times.

Larson also discusses repositioning (reassignment or patrol units to areas other than they are currently assigned) and preventive patrol.

A more detailed discussion of dispatching across beat boundaries (intersector dispatching) and other concepts appear in Larson[6]. He finds the optimal beat design for two beats to minimize the average travel distance under intersector cooperation and repositioning. Larson[7] analyzed spatially distributed queueing systems with up to 12 response units for Poisson arrivals and negative exponential service times.

PREDICTIVE MODEL OF RESPONSE

Before an attempt can be made to minimize response time there needs to be developed a procedure that will determine the expected response time for a particular configuration of beats.

A district can be divided into a number of mutually exclusive and collectively exhaustive contiguous geographical units. If each geographical unit is represented by a node, the district can be viewed as a network of nodes. A beat will be formed by combining a number of these nodes. A feasible configuration of beats should cover all the nodes in the district with the available patrol units. Division of a district into nodes is discussed in Appendix A of Bammi[8].

It will be assumed that arrivals of calls for service are Poisson. The theoretical reasoning for this is that there is a large population capable of producing calls for service, and any one of them has a small probability of producing a call for service in a short interval of time t. Larson[9] showed that the Poisson distribution was a good approximation for Boston. The Poisson assumption for arrivals of calls for service was validated for Aurora, Illinois by Thomopoulos[10].

The service-time distributions will be left arbitrary. Larson[9] and Nilsson[4] both showed that the service-time distributions are not negative exponential. The St. Louis Project[11] used a Poisson input, negative exponential service time, multiserver model in which the mean service rate is the same for all patrol units.

Queueing Model for Independent Beats

We will make the following assumptions in this section:

1. Each beat has one patrol unit;
2. Arrivals of calls at a node follow the Poisson distribution
3. Each patrol unit will service its own beat calls only, i.e., there will be no dispatching across beat boundaries;
4. Calls of all types are serviced with the same priority;
5. Time to service a call is a function only of the type of call and not a function of the node of occurrence or of the patrol unit assigned the call.

In subsequent models we will drop assumptions 3 and 4.

The notations used in this paper appear in the section titled summary of definitions.

Expectation and Variance of Service Time in Beat k.

Information on arrival rate of calls, and expectation and variance of service time can be obtained for each node by analysis of historical data on calls for service.

For Poisson arrivals, the arrival rate of calls in beat k can be obtained by summing the arrival rate of calls at each of the I_k nodes within beat k. See for example, Conway et al[12]. Thus,

$$\lambda_{(k)} = \sum_{i=1}^{I_k} \lambda_i$$

The expected service time for a call in a beat is a function of the expected service time for calls at each of its constituents nodes, weighted by the fraction of total calls in the beat at each node. For Poisson arrivals we have

$$E(t_{s(k)}) = \sum_{i=1}^{I_k} (\lambda_i E(t_{si}))/\lambda_{(k)}$$

Similarly, variance of service time for a call in a beat is given by

$$V(t_{s(k)}) = E(t_{s(k)}^{2}) - E(t_{s(k)})^2$$

Collection of data based on nodes is essential to allow for different beat designs. Further, since a node is a small enough geographical unit, statistical analysis on calls for service by nodes helps the police administrator perceive changes in crime trends over time.

Expectation and Variance of Travel Time in Beat k.

To determine the expected travel distance per call for a patrol unit answering calls in its own beat we need to determine the probability q_{imk} of patrol unit k traveling from node i to node m, $i=1,2,3,\ldots, I_k$, $m=1,2,3,\ldots,I_k$, given that the unit travels from node i to m.

Following Parzen[13] we have the expected travel distance of patrol unit k to answer a call in its own beat

$$E_{kk}(d) = \sum_{i=1}^{I_k} \sum_{m=1}^{I_k} q_{imk} E(d_{im}) \qquad (1)$$

The probability of unit k traveling from node i to m is equal to the probability of unit k being at node i multiplied by the probability of unit k traveling from node i to node m, given that it is at node i; i.e.,

$$q_{imk} = q_{ik} \times q_k \, (i \rightarrow m \mid i)$$

Neglecting the strategic aspects of crime location on the part of the criminals, the arrival of calls in different nodes of a beat should be independent. For independent Poisson arrivals the probability of unit k traveling from node i to node m given that it is at node i, will be equal to the fraction of calls of beat k that occur at node m.

Thus,

$$q_k \, (i \rightarrow m \mid i) = \lambda_m / \sum_{m=1}^{I_k} \lambda_m \qquad (2)$$

For Poisson arrivals, the fraction of time unit k is at node i while it services a call in its own beat equals

$$\lambda_i \, E\,[t_{si}] / \sum_{i=1}^{I_k} (\lambda_i \, E[t_{si}])$$

When a patrol unit is not answering a call for service it might be on downtime or on preventive patrol. These can be carried out under one of the two following policies:

(1) preventive patrol or downtime is concentrated in various nodes in proportion to the fraction of time unit k spends servicing a call in that node,

(2) preventive patrol or downtime is distributed uniformly over all nodes in the beat.

It seems policy 1 for preventive patrol would be more effective in combatting crime than policy 2. A third policy for downtime could be one which shows a higher proportion of downtime for some specific nodes such as nodes containing city courts or some popular restaurants.

Under policy 1, the fraction of time unit k is at node i while on preventive patrol or downtime is equal to the fraction of time unit k is at node i while servicing calls in its own beat.

Under policy 1 of preventive patrol and policy 1 of downtime we have

$$q_{ik} = \lambda_i E[t_{si}] / \sum_{i=1}^{I_k} (\lambda_i E[t_{si}]) \qquad (3)$$

Under policy 1 of preventive patrol and policy 2 of downtime, we have

$$q_{ik} = (1-\rho_d) \lambda_i E[t_{si}] / \sum_{i=1}^{I_k} (\lambda_i E[t_{si}]) + \rho_d / I_k \qquad (4)$$

where ρ_d is the fraction of time the patrol unit is down and not available.

Under policy 2 of preventive patrol and policy 1 of downtime, we have

$$q_{ik} = (\rho_{(k)} + \rho_d) \lambda_i E[t_{si}] / \sum_{i=1}^{I_k} (\lambda_i E[t_{si}]) + (1-\rho_{(k)} - \rho_d)/I_k \qquad (5)$$

Under policy 2 of preventive patrol and policy 2 of downtime, we have

$$q_{ik} = \rho_{(k)} \lambda_i E[t_{si}] / \sum_{i=1}^{I_k} (\lambda_i E[t_{si}]) + (1 - \rho_{(k)})/I_k \qquad (6)$$

The expected travel distance between two nodes i and m is derived in Bammi[8]. Dividing it by the average velocity we obtain the travel time t_{im} between nodes. Then, modifying equation (1) and using equation (2) and one of the equations (3), (4), (5) or (6), we obtain the expected travel time of patrol unit k to answer a call in its own beat. For Aurora, Illinois, the police administrators chose equation (4) so that preventive patrol was concentrated in various nodes in proportion to workload at the nodes, and downtime was uniformly distributed over the beat.

$$E(t_{rkk}) = \sum_{i=1}^{I_k} \sum_{m=1}^{I_k} q_{imk} \, t_{im}$$

Similarly,

$$E(t_{rkk}^2) = \sum_{i=1}^{I_k} \sum_{m=1}^{I_k} q_{imk} \, t_{im}^2$$

and

$$V(t_{rkk}) = E(t_{rkk}^2) - E(t_{rkk})^2$$

<u>Expected Response Time</u>. Since we assumed that unit k answers all calls in its beat, utilization rate of unit k assigned to beat k while servicing its own calls

$$\rho_{(k)} = \lambda_{(k)} \times (E(t_{rkk}) + E(t_{s(k)}))$$

If travel time and service time distributions are independent, the variance of calls answered by unit k

$$\sigma^2_{(k)} = V[t_{rkk}] + V[t_{s(k)}]$$

Downtime. We distinguish two types of downtime. Fixed downtime represents the type of duties that have to be answered by the patrol force during a given shift and is not dependent on the number of patrol units in operation. Variable downtime is that part of downtime which increases linearly with the number of units in operation. The arrival rate of downtime calls is given by

$$\lambda_d = \lambda_{fd}\, C_o/K + \lambda_{vd} \tag{7}$$

where λ_{fd} is the arrival rate of fixed downtime calls per unit when the number of average units in operation was C_o. K is the number of units for which beats are being designed. λ_{vd} is the average arrival rate of variable downtime calls per unit. $E[t_{fd}]$ and $E[t_{vd}]$ are the expectations of fixed downtime and variable downtime calls.

The utilization factor for downtime calls is given by

$$\rho_d = \lambda_{fd}\, C_o\, E[t_{fd}]/K + \lambda_{vd}\, E[t_{vd}] \tag{8}$$

The expectation of a downtime call is given by

$$E[t_d] = \rho_d/\lambda_d \tag{9}$$

The variance of a downtime call is given by

$$V[t_d] = (\lambda_{fd}\, C_o\, V[t_{fd}]/K + \lambda_{vd}\, V[t_{vd}])/\lambda_d \tag{10}$$

Average Number of Waiting Calls. In determining the average number of waiting calls, we must distinguish two types of calls: calls for service (source 1) and downtime calls (source 2). Response time is to be calculated only for calls for service (source 1). The average number of waiting calls for source 1 is affected by the arrival of downtime calls (source 2). If no precedence is assumed, the Pollaczek-Khintchine formula may be used to give the average number of waiting calls for both sources in beat k (see, for example, Saaty[14]) as

$$L'_{q(k)} = \frac{(\lambda_{(k)} + \lambda_d)^2}{2(1-\rho_{(k)} - \rho_d)} \int_0^\infty t^2\, b(t)\, dt \tag{11}$$

where b(t) is the service-time density, i.e.,

$$b(t) = \frac{\lambda_{(k)}}{\lambda_{(k)} + \lambda_d}\, h_k(t) + \frac{\lambda_d}{\lambda_{(k)} + \lambda_d}\, h_d(t)$$

where $h_k(t)$ is the service-time density of calls for service answered by unit k, and $h_d(t)$ is the density of downtime calls.

If $h_k(t)$ is the m-th member of the Erlang family of service time distributions and $h_d(t)$ is the n-th member of the Erlang family

$$b(t) = \frac{\lambda_{(k)}}{\lambda_{(k)} + \lambda_d} \frac{(\mu_{kk} m)^m}{(m-1)!} t^{m-1} e^{-m\mu_{kk} t} + \frac{\lambda_d}{\lambda_{(k)} + \lambda_d} \frac{((E[t_d])^{-1} n)^n}{(n-1)!} t^{n-1} e^{-n(E[t_d])^{-1} t} \tag{12}$$

where

$$\mu_{kk} = (E[t_{rkk}] + E[t_{s(k)}])^{-1}$$

Substituting equation (12) in (11) and integrating

$$L'_{q(k)} = \frac{(\rho_{(k)} + a\rho_d)(\rho_{(k)}\frac{m+1}{2m} + \frac{\rho_d}{a}\frac{n+1}{2n})}{(1-\rho_{(k)} - \rho_d)}$$

where

$$a = \frac{E[t_{rkk}] + E[t_{s(k)}]}{E[t_d]}$$

Since the variance of the m-th member of the Erlang family is $(m\mu^2)^{-1}$

$$\sigma^2_{(k)} = (m\,\mu^2_{kk})^{-1}$$

and $$m = (\sigma^2_{(k)}\,\mu^2_{kk})^{-1}$$

and $$V[t_d] = n^{-1}\,E[t_d]^2$$

$$n = E[t_d]^2/V[t_d]$$

Substituting values of m and n we obtain

$$L'_{q(k)} = \frac{(\rho_{(k)} + a\rho_d)(\rho_{(k)}\frac{1+\sigma^2_{(k)}\mu^2_{kk}}{2} + \frac{\rho_d}{a}\frac{1+V[t_d]E[t_d]^{-2}}{2})}{(1-\rho_{(k)} - \rho_d)} \tag{13}$$

The average number of waiting calls from source 1 (calls for service)

$$L_{q(k)} = \frac{\lambda_{(k)}}{\lambda_{(k)}+\lambda_d}\;L'_{q(k)}$$

$$L_{q(k)} = \frac{\rho_{(k)}(\rho_{(k)}\frac{1+\sigma^2_{(k)}\mu^2_{kk}}{2} + \frac{\rho_d}{a}\frac{1+V[t_d]\,E[t_d]^{-2}}{2})}{(1-\rho_{(k)} - \rho_d)} \tag{14}$$

If the service-time density of calls for service and the density of downtime calls are both distributed negative exponentially

$$L_{q(k)} = \frac{\rho_{(k)}\,(\rho_{(k)} + \rho_d/a)}{(1-\rho_{(k)} - \rho_d)} \tag{15}$$

If we assume that the calls for service have precedence over downtime calls and we have Poisson arrivals and negative exponential services, it has been shown that the average number of waiting calls from source 1 (calls for service) is (see for example, Saaty)[14],

$$L_{q(k)} = \frac{\rho_{(k)}\,(\rho_{(k)} + \rho_d/a)}{(1-\rho_{(k)})} \tag{16}$$

Comparing equations (15) and (16), we note that the $L_{q(k)}$ differ only by a factor in the denominator. If this same factor holds for arbitrary service time distributions, the average number of waiting calls from source 1 (calls for service) when calls for service have precedence over downtime calls

$$L_{q(k)} = \frac{\rho_{(k)}(\rho_{(k)}\frac{1+\sigma^2_{(k)}\mu^2_{kk}}{2} + \frac{\rho_d}{a}\frac{1+V[t_d]\,E[t_d]^{-2}}{2})}{(1-\rho_{(k)})} \tag{17}$$

Equations (13), (14), (15), (16), (17) are valid for $\rho_{(k)}$, $\rho_d < 1$, $k = 1,2,\ldots,K$, where K is the number of beats in the district.

A simulation model was developed to compare the value of $L_{q(k)}$ given by equation (17) and that obtained from simulation. The model simulated one beat having Poisson arrivals of calls for service and downtime calls. Calls for service had precedence over downtime calls. Several distributions were used to generate service time on calls for service and downtime calls. When calls for service followed the negative exponential distribution there was 2.13 percent difference between the value of the average number of clalls for service in waiting line obtained from simulation and that obtained from equation (17). When the Erlang 2 distribution was used the percent error was 3.22 The Erlang 5 distribution yielded a percent error of 1.89 The uniform distribution calls for service showed an error of 1.57 percent. From these results we concluded that the computer simulation validated the assumption made in deriving equation (17).

Expected number of calls in system (beat) for beat k

$$L_{(k)} = L_{q(k)} + \rho_{(k)}$$

where either equation (17) or (14) is used to obtain $L_{q(k)}$, depending on whether or not there is precedence of calls for service over downtime calls. For Aurora, Illinois precedence was assumed.

In order that a unit may respond to a call that just arrived in the beat, it must first service all the calls in the system (the system being defined as the beat), it must travel to all calls in the waiting line and finally travel to the new call. Thus, the expected response time to a call in beat k

$$\begin{aligned} E(t_{w(k)}) &= L_{(k)}E(t_{s(k)}) + L_{q(k)}E(t_{rkk}) + E(t_{rkk}) \\ &= (L_{q(k)} + \rho_{(k)})\, E(t_{s(k)}) + (L_{q(k)} + 1)E(t_{rkk}) \\ k &= 1,2,3,\ldots,K \end{aligned}$$

The expected response time to a call in the district then becomes

$$\begin{aligned} E(t_w) &= \sum_{k=1}^{K} (\lambda_{(k)}E(t_{w(k)}))/\sum_{k=1}^{K} \lambda_{(k)} \\ &= \sum_{k=1}^{K} (\lambda_{(k)}E(t_{w(k)}))/\lambda \end{aligned}$$

where the response times have been weighted by the arrival rates in the various beats.

Queueing Model with Dispatching Across Beat Boundaries

Having developed a model for independent beats (herein referred to as the "non-flying" problem) we relax the assumption that patrol units cannot be dispatched across beat boundaries+ (herein referred to as the "flying" problem). If a call occurs in beat k and patrol unit k is not busy, it services the call. If patrol unit k is busy, then an adjoining beat unit is questioned next regarding its availability. If this adjoining beat unit is not busy, it services the call. If it is busy, then another adjoining unit is questioned. This is continued until a patrol unit is assigned the call or it is found that all units are busy. In the latter case, the call joins a queue of waiting calls and is assigned the first available unit when its turn comes. As soon as a patrol unit finishes servicing a call, it returns to its own beat and starts preventive patrol

We make the following assumptions in this section:

1. each beat has one patrol unit,
2. arrival of calls at a node follows the Poisson distribution,
3. calls of all types are serviced with the same priority,
4. time to service a call is a function only of the type of call and not a function of the node of occurrence or of the patrol unit assigned to the call.

Service distributions are kept arbitrary.

Footnote for Page 22

+Larson[6] refers to dispatching across beat boundaries as intersector dispatching. He solves for the amount of intersector dispatching under certain special conditions and places bounds on it for a generalized dispatching algorithm.

<u>Determination of "Flying" Probabilities</u>. In order to solve this problem we need to determine the "flying" probabilities, $Q_{k\ell}$, fraction of calls arriving in beat k answered by unit ℓ.

<u>Two Beat Problem</u>. The fraction of incoming calls in beat 1 answered by patrol unit 1 is equal to the probability of unit 1 being available for dispatch plus the probability that patrol unit 1 is busy multiplied by the probability that unit 2 is busy multiplied by the probability that a queued call is answered by unit 1 (a call is termed "queued" if both patrol units 1 and 2 are unavailable for dispatch). A unit is unavailable if it is busy servicing a call or <u>if</u> it is on a type of administrative downtime which obviates its dispatch. This dispatch policy is shown in Fig. 2.

$$Q_{11} = (1-\rho_1') + \rho_1'\rho_2'v_1 \qquad (18)$$

where v_ℓ the probability that a queued call is answered by unit ℓ is given by equation (22) below, $\ell=1,2$, and $\rho_\ell'=\rho_\ell+\rho_d$, $\rho_\ell'<1$, $\ell=1,2$.

In equation (18) we multiplied the probability of unit 1 being busy by the probability of unit 2 being busy to obtain the probability of both units being busy. This is an approximation insofar as we can multiply probabilities of two events directly, only if they are independent. Recognization of these two events being dependent is taken when we evaluate ρ_1 and ρ_2 in equations (30). For example, in order to determine ρ_1 we consider the calls that unit 1

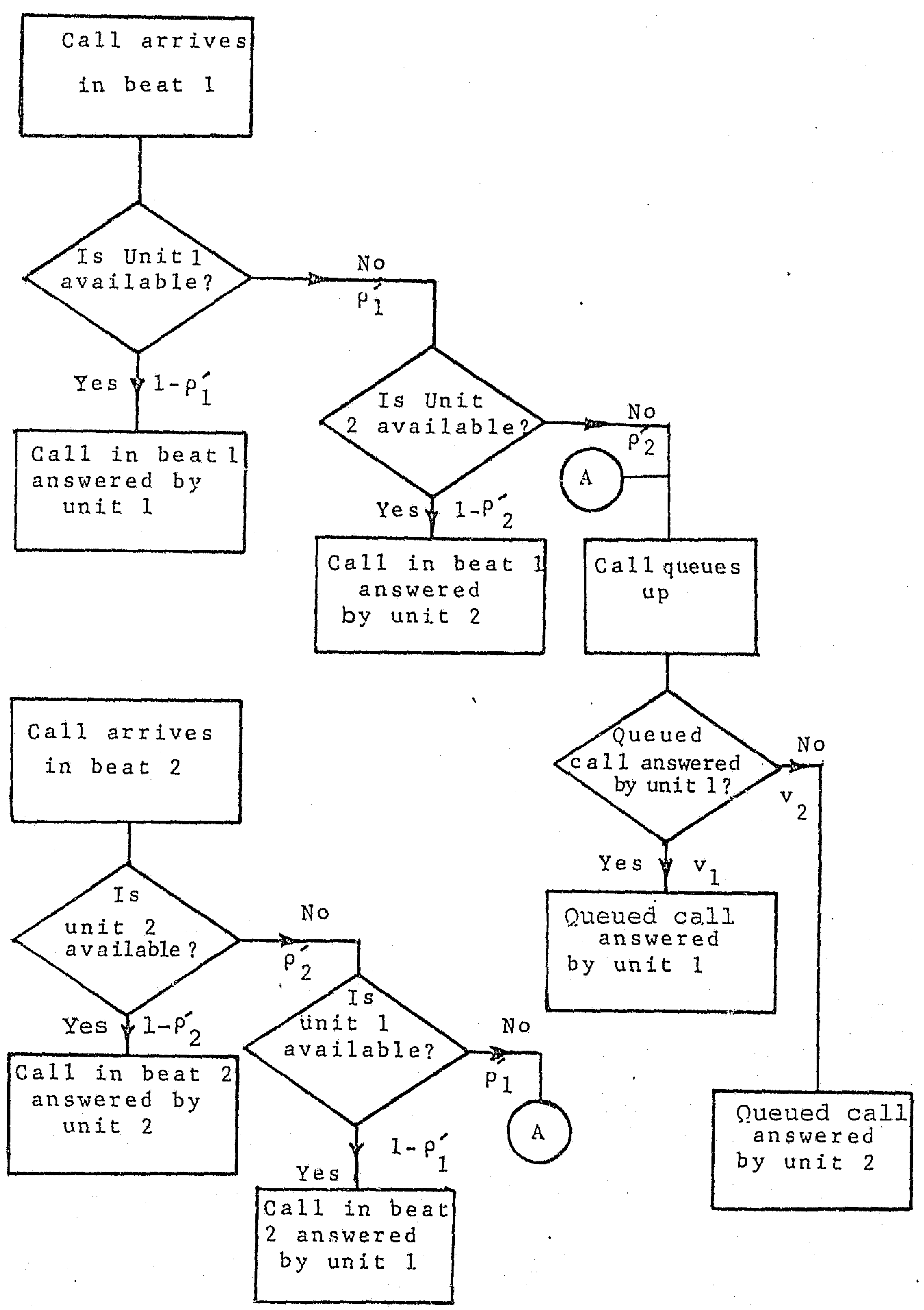

Figure 2. A Dispatch Policy for a Two-Beat District

answers in its own beat as well as those it answers in beat 2.

The fraction of incoming calls in beat 1 answered by unit 2 is equal to the probability that unit 1 is busy and unit 2 is available for dispatch plus the probability that both units 1 and 2 are busy multiplied by the probability that a queued call is answered by unit 2

$$Q_{12} = \rho_1'(1-\rho_2') + \rho_1'\rho_2' v_2 \qquad (19)$$

also,

$$\begin{aligned} Q_{11} + Q_{12} &= 1-\rho_1' + \rho_1'\rho_2' v_1 + \rho_1' - \rho_1'\rho_2' + \rho_1'\rho_2' v_2 \\ &= 1-\rho_1'\rho_2' + \rho_1'\rho_2' (v_1 + v_2) \\ &= 1-\rho_1'\rho_2' + \rho_1'\rho_2' \\ &= 1 \end{aligned}$$

Similarly for calls arriving in beat 2 we have

$$Q_{21} = \rho_2'(1-\rho_1') + \rho_1'\rho_2' v_1 \qquad (20)$$

$$Q_{22} = 1-\rho_2' + \rho_1'\rho_2' v_2 \qquad (21)$$

also $Q_{21} + Q_{22} = 1$

If all units are busy and unit ℓ is the first to finish servicing a call, it will be assigned to one of the queued calls. Since, the event that unit ℓ is assigned a queued call occurs if and only if unit ℓ is the first to finish servicing a call, we can say that the probability

that a queued call is assigned to unit ℓ should be the same as the probability that unit ℓ is the first to finish servicing a call.

In some situations there may be a built-in bias such that if all units are busy, it may be more likely that a particular unit is the first to become free. One way this may occur is if the dispatcher tends to assign the central (downtown) units more often to a call in an adjoining beat than an outlying unit because of a closer center-of-mass for the central units. Also, if some beats having low arrivals of calls are constrained not to be large geographically they may have a lower utilization factor than other beats and may be the last ones to become busy and therefore among the last to become free again. This built-in bias could be corrected (if present) by estimating the workload of each unit (from non-flying model) and applying a correction factor to equation (22) or (23). However, as later estimates show, the need to use these equations arises less than .39 percent of the time so this additional modification is probably not necessary from a practical point of view

By the above argument then,

v_ℓ = probability that unit ℓ finishes servicing a call before unit ℓ_1, $\ell_1 = 1,2,3,\ldots, L$, $\ell_1 \neq \ell$

= probability $(t_\ell < t_1, t_2, \ldots, t_{\ell_1}, \ldots, t_L, \ell_1 \neq \ell)$

$$= \int_{-\infty}^{\infty}\int_{t_\ell}^{\infty}\int_{t_\ell}^{\infty} \cdots \int_{t_\ell}^{\infty} \text{joint probability density function } (t_1, t_2, \ldots, t_L)\, dt_L dt_{L-1} \ldots dt_1 dt_\ell$$

If the service distribution of calls answered by unit ℓ and ℓ_1 are independent, $\ell \neq \ell_1$, $\ell, \ell_1 = 1,2,\ldots, L$

$$v_\ell = \int_{-\infty}^{\infty} h_\ell(t_\ell) \int_{t_\ell}^{\infty} h_1(t_1) \int_{t_\ell}^{\infty} h_2(t_2) \ldots \int_{t_\ell}^{\infty} h_{\ell_1}(t_{\ell_1}) \ldots \int_{t_\ell}^{\infty} h_L(t_L)\, dt_L \ldots dt_{\ell_1} \ldots dt_2\, dt_1\, dt_\ell$$

where $t_1, t_2, \ldots, t_\ell, \ldots, t_L$ are the service time remaining till completion.

Since t_ℓ cannot be negative

$$v_\ell = \int_{0}^{\infty} h_\ell(t_\ell) \int_{t_\ell}^{\infty} h_1(t_1) \int_{t_\ell}^{\infty} h_2(t_2) \ldots \int_{t_\ell}^{\infty} h_{\ell_1}(t_{\ell_1}) \ldots \int_{t_\ell}^{\infty} h_L(t_L)\, dt_L \ldots dt_{\ell_1} \ldots dt_2\, dt_1 dt_\ell \quad (22)$$

The service time distributions of calls answered by unit ℓ, $\ell = 1,2,\ldots, L$, can be obtained by a linear combination of the service time distributions of calls arriving in beat k, k= 1,2,..., K. These in turn can be obtained from

sampled data. If we assume negative exponential services, the distribution for the service time remaining till completion is the same as the total service time distribution. For negative exponential services,

$$v_\ell = \mu_\ell \Big/ \sum_{\ell=1}^{L} \mu_\ell \;, \; \ell = 1,2,\ldots,L \quad (23)$$

Equation (23) is only approximately true for service time distributions other than negative exponential. However, if the service rates are the same the approximation is exact for all Erlang distributions. If service rates are about the same the approximation is fairly close. For example, with three units each following the Erlang 2 distribution with $\mu_2=0.8\mu_1$ and $\mu_3 = 1.2\mu_1$, the probability that unit 1 is assigned a queued call is 0.3304 by equation (22) and 0.3333 by the approximate equation (23). Similarly, for three units each following the Erlang 3 distribution with $\mu_2=0.8\mu_1$ and $\mu_3=1.2\mu_1$, the probability that unit 1 is assigned a queued call is 0.3033 by equation (22) and 0.3333 by the approximate equation (23). Our experience has shown that the service rates for different units do not deviate more than 20 percent from the average service rate so testing for $\mu_2=.8\mu_1$ and $\mu_3=1.2\mu_1$ seems adequate.

In any case, equations (22) or (23) are used only if all units are busy. For a city (or district) deploying sixteen units which are busy about 15 percent of the time (a typical figure for Aurora, Illinois) the probability that all of them are busy (assuming independence) is only 656×10^{-16}. Even if a police department has its units busy on the average 50 percent of the time, the probability of all being busy is still only 0.0039 for eight units and .000015 for sixteen units.

Thus, since the approximation of equation (23) is needed only very seldom (less than one-half percent of the times) and the approximation itself is not bad for operating conditions, we can say that the model is for the most part valid for arbitrary service time distributions.

Three Beat Problem. A call arriving in beat 1 is answered by patrol unit 1 if it is available. If unit 1 is busy then the dispatcher must decide whether unit 2 or unit 3 should be questioned next regarding its availability. If the dispatcher knew the exact location of both units 2 and 3 at the time the call occurred in beat 1, then the nearest unit could be dispatched. However, in most police stations the dispatcher does not know the exact location of all units. Individual police departments have developed, either formally or informally, an assignment policy. We will allow here the possibility of a probabilistic assignment policy.

For example, if unit ℓ_1 is not available then unit ℓ_2 should be questioned next regarding its availability a fraction $w_{\ell_1 \ell_2}$ of the time, $\ell_1, \ell_2 = 1,2,\ldots,L$. For instance, if the expected travel distance from beat 2 to beat 1 is the same as that from beat 3 to beat 1, then if unit 1 is not available the dispatcher may question unit 2 next with probability 0.5 and unit 3 next with probability 0.5 Also, if the expected travel distance from beat 2 to beat 1 is much less than the expected travel distance from beat 3 to beat 1, then if unit 1 is not available the dispatcher may question unit 2 next always.

A more accurate representation of actual dispatching policies is obtained by determining an assignment policy for each node. If a call occurs at node i in beat k and unit k is not available, then the dispatcher questions that unit next which has the closest "center of mass" to this node. Center of mass of a beat is defined here as the center of gravity of the beat weighted by the "workload" of its component nodes.+ Workload of a node is obtained by multiplying the arrival rate of calls at that node by the expected service time at the node. By summing the assignment policies for its component nodes, a probabilistic assignment policy for calls arriving in a beat is developed. The programs in Bammi[8] demonstrate how this can be done on a digital computer.

The fraction of incoming calls in beat 1 answered by patrol unit 1 is equal to the probability of unit 1 being available for dispatch plus the probability that all three units are busy multiplied by the probability that a queued call is answered by unit 1. The dispatching policy for incoming calls to beat 1 is shown in Fig. 3.

Thus,

$$
\begin{aligned}
Q_{11} &= 1-\rho_1' + \rho_1' w_{12}\rho_2'\rho_3' v_1 + \rho_1' w_{13}\rho_3'\rho_2' v_1 \\
&= 1-\rho_1' + \rho_1'\rho_2'\rho_3' v_1 \\
Q_{12} &= w_{12} \rho_1'(1-\rho_2') + w_{13}\rho_1'\rho_3'(1-\rho_2') \\
&+ w_{12} \rho_1'\rho_2'\rho_3' v_2 + w_{13}\rho_1'\rho_3'\rho_2' v_2 \\
&= w_{12} \rho_1'(1-\rho_2') + w_{13} \rho_1'\rho_3'(1-\rho_2') + \rho_1'\rho_2'\rho_3' v_2 \\
Q_{13} &= w_{12} \rho_1'\rho_2'(1-\rho_3') + w_{13} \rho_1'(1-\rho_3') + \rho_1'\rho_2'\rho_3' v_3
\end{aligned}
$$

Footnote for Page 31

+We are assuming here that the average of the above function is a function of the average. A more accurate representation of the dispatching policy would be to find the expected distance between node i in beat k and unit ℓ in beat ℓ by summing the distance between node i in beat k and each of the nodes in beat ℓ weighted by the probability of unit ℓ being at each of the nodes in its beat.

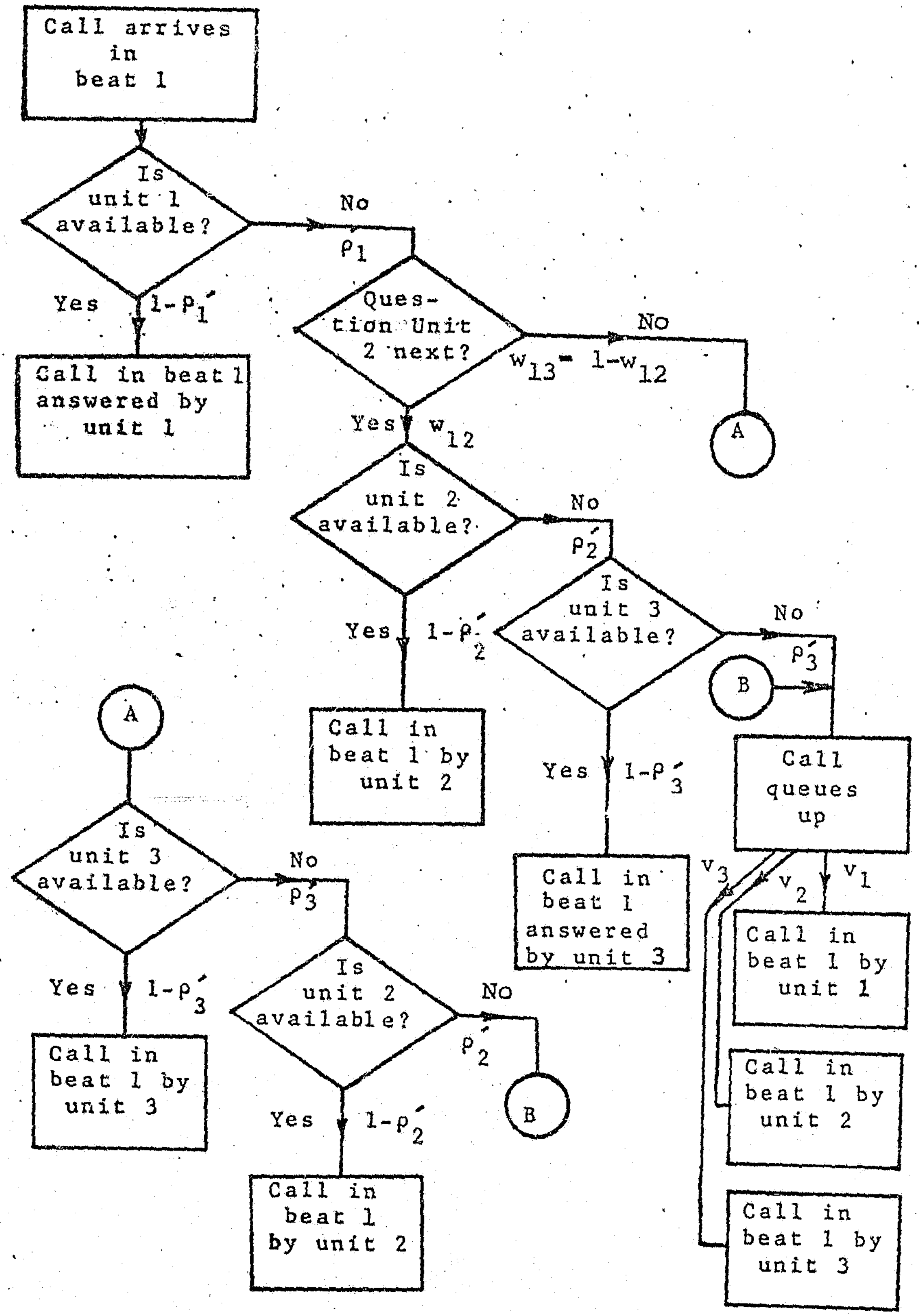

Figure 3. A Dispatch Policy for a Three-Beat District. Arivals in Beat 1

$$Q_{21} = w_{21}\ \rho_2'(1-\rho_1') + w_{23}\ \rho_2'\rho_3'(1-\rho_1') + \rho_1'\rho_2'\rho_3'\ v_1$$

$$Q_{22} = 1-\rho_2' + \rho_1'\rho_2'\rho_3'\ v_2$$

$$Q_{23} = w_{21}\ \rho_2'\rho_1'(1-\rho_3') + w_{23}\ \rho_2'(1-\rho_3') + \rho_1'\rho_2'\rho_3'\ v_3$$

$$Q_{31} = w_{31}\ \rho_3'(1-\rho_1') + w_{32}\ \rho_3'\rho_2'(1-\rho_1') + \rho_1'\rho_2'\rho_3'\ v_1$$

$$Q_{32} = w_{31}\ \rho_3'\rho_1'(1-\rho_2') + w_{32}\ \rho_3'(1-\rho_2') + \rho_1'\rho_2'\rho_3'\ v_2$$

$$Q_{33} = 1-\rho_3' + \rho_1'\rho_2'\rho_3'\ v_3 \tag{24}$$

where, $w_{12} + w_{13} = 1$

$w_{21} + w_{23} = 1$

$w_{31} + w_{32} = 1$

where, v_1, v_2, v_3 are obtained from equation (23).

where, $\rho_\ell' = \rho_\ell + \rho_d \quad , \quad \ell = 1,2,3,\ldots,L.$

In a like fashion we can determine $Q_{k\ell}$ for an arbitrary number of beats.

A simulator which validated values of $Q_{k\ell}$ obtained from equaitons (18),(19),(20),(21), and (24) is described in the Appendix.

In order to determine $Q_{k\ell}$ we need the combined utilization factor of unit ℓ considering calls for service and downtime calls. But the utilization factor of unit ℓ depends on $Q_{k\ell}$ (see equation (30)). As a first approximation, we obtained the utilization factor of unit ℓ from the non-flying problem (independent beats). To this was added the percent of time spent on downtime to obtain ρ_ℓ^1 the combined utilization factor for unit ℓ to be used in equation (24) for determining $Q_{k\ell}$. Next ρ_ℓ is found from equation (30). A series of computer runs were made to test the convergence of $Q_{k\ell}$ by repeatedly using equation (24) for $Q_{k\ell}$ and equation (30) for ρ_ℓ after the first iteration. It was found that $Q_{k\ell}$ converged very fast, and to save computer time only the first iteration was retained in the programs.

Expectation and Variance of Service Time of Calls Answered by Unit ℓ. When units are allowed to answer calls in beats other than their own, the input stream of calls generated for each unit is Poisson if the input stream of calls in each beat is Poisson. This can be seen by repeatedly applying two theorems proved by Conway et al.[12], viz., (i) the probabilistic selection of jobs from a single Poisson stream into several output paths yield independent Poisson streams, and (ii) the aggregation of several Poisson input streams results in Poisson stream.

In particular, if the arrivals of calls for service in beat k follow the Poisson distribution with parameter $\lambda_{(k)}$, $k = 1,2,3,\ldots, K$, then the calls answered by unit ℓ follow the Poisson distribution with parameter $\sum_{k=1}^{K} Q_{k\ell}\lambda_{(k)}$, $\ell = 1,2,3,\ldots,L$.

Then, expected service time of calls answered by unit ℓ

$$= E[t_\ell] = \sum_{k=1}^{K} (Q_{k\ell}\ \lambda_{(k)}\ E[t_{s(k)}]) / \sum_{k=1}^{K} Q_{k\ell}\,\lambda_{(k)}$$

Variance of calls answered by unit $\ell = \sigma_\ell^2$

$$= \sum_{k=1}^{K} (Q_{k\ell}\,\lambda_{(k)}\ (V[t_{rk\ell}] + V[t_{s(k)}])) / \sum_{k=1}^{K} (Q_{k\ell}\,\lambda_{(k)})$$

The covariance is zero if the arrival of calls in beat k is independent of the arrival of calls in beat ℓ.

Expectation and Variance of Travel Time for Calls Arriving in Beat k and Answered by Unit ℓ. The expected travel distance of unit ℓ to answer a call in beat k is given by

$$E_{k\ell}[d] = \sum_{i=1}^{I_\ell} \sum_{m=1}^{I_k} q_{i\ell mk}\ E(d_{im}) \quad (27)$$

where the summation is over all nodes in beat ℓ and all nodes in beat k. $k,\ell = 1,2,3,\ldots,K$.

As before, $q_{i\ell mk} = q_{i\ell} \times q_{k\ell}\ (i \to m|i)$ (28)

Patrol unit ℓ answers $Q_{\ell\ell}\lambda_{(\ell)}$ calls in its own beat. For Poisson arrivals these $Q_{\ell\ell}\ \lambda_{(\ell)}$ calls are divided among the I_ℓ nodes in beat ℓ in the same proportion as the total $\lambda_{(\ell)}$ calls in beat ℓ, i.e., unit ℓ answers $\lambda_i\ Q_{\ell\ell}$ calls at node i. Then, the fraction of time unit ℓ is at node i while it services a call in its own beat equals $\lambda_i\ E[t_{si}]/\sum_{i=1}^{I_\ell}(\lambda_i\ E[t_{si}])$

As before, equations (3), (4), (5), and (6) are used to determine $q_{i\ell}$, the probability of being at node i in beat ℓ.

For Poisson arrivals, the probability of unit ℓ traveling from node i in beat ℓ to node m in beat k given that it is at node i

$$q_{k\ell}(i\rightarrow m|i) = \lambda_m/\sum_{m=1}^{I_k}\lambda_m \qquad (29)$$

where we have cancelled $Q_{k\ell}$ from the numerator and denominator.

Modifying equation (27) as before and using equations (28) and (29) and one of the equations (3), (4), (5), or (6), we obtain the expected travel time of patrol unit ℓ to answer a call in beat k.

$$E[t_{rk\ell}] = \sum_{i=1}^{I_\ell}\sum_{m=1}^{I_k} q_{i\ell\,mk}\ t_{im}$$

Similarly, $E[t_{rk\ell}^2] = \sum_{i=1}^{I_\ell}\sum_{m=1}^{I_k} q_{i\ell\,mk}\ t_{im}^2$

and $V[t_{rk\ell}] = E[t_{rk\ell}^2] - E[t_{rk\ell}]^2$

also, expected travel time of calls answered by unit ℓ,

$$E[t_{r\ell}] = \sum_{k=1}^{K}(Q_{k\ell}\lambda_{(k)}\ E[t_{rk\ell}])/\sum_{k=1}^{K} Q_{k\ell}\lambda_{(k)}$$

<u>Expected Response Time</u>. The utilization rate of unit ℓ is given by

$$\rho_\ell = \sum_{k=1}^{K}(Q_{k\ell}\lambda_{(k)}\ (E[t_{rk\ell}] + E[t_{s(k)}])) \qquad (30)$$

As before, the arrival rate, utilization factor, expectation and variance of downtime are obtained from equations (7), (8), (9), and (10).

The service-time density in the flying case is given by

$$b(t) = \frac{\sum_{k=1}^{K}(Q_{k\ell}\lambda_{(k)})}{\sum_{k=1}^{K}(Q_{k\ell}\lambda_{(k)})+\lambda_d}\ \frac{(\mu_\ell m)^m}{(m-1)!}\ t^{m-1}\ e^{-m\mu_\ell t}$$

$$+ \frac{\lambda_d}{\sum_{k=1}^{K}(Q_{k\ell}\lambda_{(k)})+\lambda_d}\ \frac{((E[t_d])^{-1}n)^n}{(n-1)!}\ t^{n-1}e^{-n(E[t_d])^{-1}t}$$

where $\mu_\ell = (E[t_{r\ell}] + E[t_\ell])^{-1}$

Thus, the average number of waiting calls for service and downtime calls for unit ℓ

$$L'_{q\ell} = \frac{(\rho_\ell + a\rho_d)(\rho_\ell \frac{m+1}{2m} + \frac{\rho_d}{a} \frac{n+1}{2n})}{(1 - \rho_\ell - \rho_d)}$$

where

$$a = \frac{E[t_{r\ell}] + E[t_\ell]}{E[t_d]}$$

As before,

$$L'_{q\ell} = \frac{(\rho_\ell + a\rho_d)(\rho_\ell \frac{1+\sigma_\ell^2 \mu^2_\ell}{2} + \frac{\rho_d}{a} \frac{1+ V[d]\ E[d]^{-2}}{2})}{(1 - \rho_\ell - \rho_d)}$$

The average number of waiting calls from source 1 (calls for service) without precedence

$$L_{q\ell} = \frac{\sum_{k=1}^{K} Q_{k\ell}\, \lambda_{(k)}}{\sum_{k=1}^{K} Q_{k\ell}\, \lambda_{(k)} + \lambda_d} L'_{q\ell}$$

$$= \frac{\rho_\ell (\rho_\ell \frac{1+\sigma_\ell^2 \mu_\ell^2}{2} + \frac{\rho_d}{a} \frac{1+ V[t_d]\ E[t_d]^{-2}}{2})}{(1-\rho_\ell - \rho_d)}$$

The average number of waiting calls from source 1 (CFS) when the CFS have precedence over downtime calls

$$L_{q\ell} \simeq \frac{\rho_\ell (\rho_\ell \frac{1+\sigma_\ell^2 \mu_\ell}{2} + \frac{\rho_d}{a} \frac{1+V[t_d]\ E[t_d]^{-2}}{2})}{(1 - \rho_\ell)}$$

where $\rho_\ell, \rho_d < 1$, $\ell = 1,2,3,\ldots, L$, where L is the number of units in the district.

Expected number of calls in system for unit ℓ

$$L_\ell = L_{q\ell} + \rho_\ell$$

The expected response time for calls answered by unit ℓ

$$E[t_{w_\ell}] = L_\ell\, E[t_\ell] + L_{q\ell}\, E[t_{r\ell}] + E[t_{r\ell}]$$

$$= (L_{q\ell} + \rho_\ell)\, E[t_\ell] + (L_{q\ell} + 1)\, E[t_{r\ell}]$$

$$\ell = 1,2,3,\ldots,L$$

The expected response time to a call in the district then becomes

$$E[t_w] = \sum_{\ell=1}^{L} (\sum_{k=1}^{K} Q_{k\ell}\lambda_{(k)}) E[t_{w\ell}] \Big/ \sum_{\ell=1}^{L} \sum_{k=1}^{K} Q_{k\ell}\lambda_{(k)}$$

$$= \sum_{\ell=1}^{L} (\sum_{k=1}^{K} Q_{k\ell}\lambda_{(k)}) E[t_{w\ell}]/\lambda$$

where the response times have been weighted by the calls answered by the various units.

Priority and Non-Priority Calls

If calls for service can be classified as either priority or non-priority then we can determine the expected response time to the two types of calls. We do this by running the models in two steps. In the first step we feed as input only the priority calls and determine the expected response time to priority calls. This procedure of obtaining the response time to priority calls is valid if priority calls preempt non-priority calls and there is a first-come-first-served queue discipline. Next we feed as input the total calls for service and obtain the expected response time to all calls for service. Non-priority calls which were interrupted during service due to the arrival of a priority call resume service at a later time and appear as a new call for service. These repeater calls are included in the total calls for service. By subtracting the response time for priority calls from the response time to all calls we obtain the increase in response time due to non-priority calls.

The measure of effectiveness can then incorporate the weights to be given to priority and non-priority calls. If α_p represents the weight to priority calls and α_n the weight to non-priority calls, the measure of effectiveness is α_p x (expected response time to priority calls) plus α_n x (increase in expected response time due to non-priority calls), $0 \leq \alpha_p \leq 1$, $0 \leq \alpha_n \leq 1$. If α_p and α_n are both equal to one it implies that all calls are weighted equally. If only priority calls are to be used in designing beats we would set α_p equal to one and α_n equal to zero.

OPTIMIZATION MODEL

The predictive model developed in this paper determines the objective function used in the optimization model by Bammi[8]. In this model police patrol beats are designed to minimize the response time to calls for service in the city. The measure of response time may be the average for all calls for service, or for a weighted function of priority and non-priority calls for service. Optimization is subject to constraints on the maximum travel time within beats and on the numbers of men and cars available. An efficient computer program has been written and applied to the design of beats for the Aurora, Illinois Police Department.

The number of iterations and the total reduction in response time from the initial to optimal solution was found to be a function of how well the initial beat configuration was designed. A good initial solution was obtained by equalizing the workload (arrival rate of calls for service multiplied by expected service time) for the beats. We also found that since response time is a function of travel time as well as workload, beats with large areas should have a workload slightly less than the average workload for a beat to account for their larger travel times. A good initial solution sometimes afforded half of the total reduction in response time.

Based on such an initial solution we found a reduction of 6.46 percent in response time from initial to optimal solution when using sixteen beats. Similarly, a reduction of 5.1 percent was observed when deploying eight beats. On comparing the optimal solution for eight beats with the beats which Aurora was using before this study was made, we found a reduction of 12.2 percent in response time. This is approximately equal to a saving of two patrol units per shift which implies a saving about $162,000 a year.

SUMMARY OF DEFINITIONS

i,m — Subscript for node number

j — Subscript for type of call

k — Subscript for beat number

ℓ — Subscript for unit number

p — Priority calls

n — Non-Priority calls

I_k — Number of nodes in beat k

J — Number of types of calls

K — Number of beats in the district

L — Number of units in the district

P — Number of types of calls that are priority

N — Number of types of calls that are non-priority

$E_{k_1,k_2,\ldots,k_\ell,\ldots,k_L,m_1,m_2,\ldots,m_k,\ldots,m_K}$. State of the system, where k_ℓ = location of unit ℓ, m_k = number of calls for service in beat k

λ_{ij} — Arrival rate of calls of type j at node i

λ_i — Arrival rate of calls of all types at node i

$\lambda_{(k)}$ — Arrival rate of calls of all types at all nodes in beat k

λ — Arrival rate of all calls in the district

λ_{fd} — Arrival rate of fixed downtime calls per unit when the average number of units in operation was C_o

λ_{vd} — Arrival rate of variable downtime calls per unit

λ_d — Arrival rate of downtime calls

t_{sij} — Time to service (not including travel) a call of type j at node i

t_{si} — Time to service a call at node i

$t_{s(k)}$ — Time to service a call in beat k

t_ℓ — Time to service a call by unit ℓ

t_s — Time to service a call in the district

t_d — Downtime

t_{im} — Travel time between nodes i and m

$t_{rk\ell}$ — Travel time for a call in beat k answered by unit ℓ.

$t_{r\ell}$ — Travel time for a call answered by unit ℓ

$\rho_{(k)}$ — Utilization rate of unit k (for independent beats), arrival rate of calls in beat k multiplied by the expected service time and travel time to answer those calls

ρ_ℓ — Utilization rate of unit ℓ, arrival rate of calls answered by unit ℓ multiplied by the expected service time and travel time for those calls

ρ_d — Utilization factor for downtime calls

ρ'_ℓ — Combined utilization factor of unit ℓ considering CFS and downtime calls

$t_{w(k)}$ — Waiting time (response time) to a call in beat k

$c_{w\ell}$ Waiting time (response time) to a call for unit ℓ

t_w Waiting time (response time) to a call in the district

$L_{(k)}$ Expected number of calls in system for beat k

L_ℓ Expected number of calls in system for unit ℓ

$L'_{q(k)}$ Expected number of calls in waiting line for beat k considering CFS and downtime calls

$L_{q(k)}$ Expected number of calls in waiting line (not including the one in service) for beat k

$L'_{q\ell}$ Expected number of calls in waiting line for unit ℓ considering CFS and downtime calls

$L_{q\ell}$ Expected number of calls in waiting line for unit ℓ

$E(\ )$ Expected value

$V(\ )$ Variance

$\sigma^2_{(k)}$ Variance of beat k calls

σ^2_ℓ Variance of unit ℓ calls

$Q_{k\ell}$ Fraction of calls arriving in beat k answered by unit ℓ

$w_{k\ell_1\ell_2\ell}$ Probability of questioning unit ℓ regarding its availability for dispatch if call arrives in beat k and units ℓ_1 and ℓ_2 are busy

v_ℓ Probability that a queued call is answered by unit ℓ

$h_\ell(t_\ell)$ Service time distribution of calls answered by unit ℓ

$h_{d}(t)$ Density of downtime calls

$b(t)$ Weighted service time density of calls for service and downtime calls

$\mu_{k\ell}$ Total service rate of calls arriving in beat k answered by unit ℓ

μ_ℓ Total service rate of all calls answered by unit ℓ

q_{ik} Probability of unit k being at node i given that unit k is in beat k

q_{imk} Probability of unit k traveling from node i to node m given that unit k answers a call in its own beat

$q_k(i \to m|i)$ Probability of unit k traveling from node i to node m given that it is at node i

$q_{i\ell mk}$ Probability of unit ℓ traveling from node i in beat ℓ to node m in beat k given that unit ℓ answers a call in beat k

$q_{k\ell}(i \to m|i)$ Probability of unit ℓ traveling from node i in beat ℓ to node m in beat k given that unit ℓ is at node i

$E(d_{im})$ Expected travel distance from node i to node m

$E_{k\ell}(d)$ Expected travel distance of unit ℓ to answer a call in beat k

u_{im} average velocity of travel between nodes i and m

a_p	Weighting factor for priority calls
a_n	Weighting factor for non-priority calls
(x_1,y_1)	Coordinates of patrol unit when dispatch order is received
(x_2,y_2)	Coordinates of the call for service
x_r	Travel distance in x-direction
y_r	Travel distance in y-direction
d_r	Travel distance
f( )	Density function

ACKNOWLEDGMENTS

The first author wishes to thank his research adviser Spencer B. Smith for guiding his dissertation[8] and providing invaluable suggestions over the past three years. Chief of Police, VICTOR PUSCAS and Commander of Field Services Bureau, ALEX M. MIHALKA of the Aurora, Illinois Police Department devoted many hours in formulating the project and developing the data base. To JOSEPH H. ENGEL thanks for his encouragement, suggestions and for providing several hours of computer time.

This work was partially supported by the Illinois Law Enforcement Commission. Computations and data analysis were performed at the IBM 370/model 155 of the University of Illinois at Chicago Circle, Computer Center and on the UNIVAC 1108 at the Computation Center of the Illinois Institute of Technology.

REFERENCES

[1]INTERNATIONAL CITY MANAGERS' ASSOCIATION. Municipal Police Administration. 1313 East 60th Street, Chicago, Illinois, 1950.

[2]SCIENCE AND TECHNOLOGY. Task Force Report by the President's Commission on Law Enforcement and Administration of Justice, U.S.G.P.O., Washington, D.C., 1967.

[3]S. B. SMITH, Superbeat: A System For the Effective Distribution of Police Patrol Units, pp. 1-14, Illinois Institute of Technology, Chicago, 1973.

[4]E. NILSSON, "Police Systems Analysis," Doctoral Dissertation, Northwestern University, Evanston, Illinois, 1969.

[5]R. C. LARSON, "Models for the Allocation of Urban Police Patrol Forces," Technical Report No. 44, Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts, November, 1969.

[6]R. C. LARSON, Urban Police Patrol Analysis, the M.I.T. Press, Cambridge, Massachusetts, 1972.

[7]R.C. LARSON , "A Hypercube Queueing Model for Facility Location and Redistricting in Urban Emergency Services", International Journal of Computers and Operations Research Volume 1, March 1974.

[8]D. BAMMI, "Design of Police Patrol Beats to Minimize Response Time to Calls for Service," Doctoral Dissertation, Illinois Institute of Technology, Chicago, Illinois, December, 1972.

[9] R. C. LARSON, "Operational Study of the Police Response System," Technical Report No. 26, Operations Research Center, Massachusetts Institute of Technology, Cambridge, Mass., December, 1967.

[10]N. T. THOMOPOULOS, "Statistical Analysis of Calls for Service," in S. B. SMITH (ed.), Superbeat: A System for the Effective Distribution of Police Patrol Units, pp. 23-51, Illinois Institute of Technology, Chicago, 1973.

[11] Saint Louis Police Department: Allocation of Patrol Manpower Resources, Vols. I and II, St. Louis, Missouri, July, 1966.

[12]R. W. CONWAY, W. L. MAXWELL, and L. M. MILLER, Theory of Scheduling, Addison-Wesley Publishing Company, Reading, Mass., 1967.

[13]E. PARZEN, Modern Probability Theory and Its Applications, John Wiley and Sons, Inc., New York, 1960.

[14] T. L. SAATY, Elements of Queueing Theory, McGraw Hill Book Company, Inc. New York, 1961.

APPENDIX

A simulator was written in the FORTRAN language to evaluate values of $Q_{k\ell}$ obtained from equations (18), (19),(20),(21), and (24). It takes as input the arrival rates and distributions of calls for service in various beats and the service rates and distributions of calls answered by each unit in every beat.

The program is written to run for any number of eight hour shifts. An initialization period at the beginning of each shift ensures an operating state when collecting statistics. The program simulates the operations in the same shift on successive days. The program has been coded for Poisson arrivals and for either negative exponential or general service time distributions. Two beat, three-beat, and four-beat districts were analyzed.

Travel time is treated by feeding as input the increase in total service time when a call is answered by a unit outside the beat rather than by the unit assigned the beat. The average utilization factor for the district is obtained by the formula

$$\rho_{av} = \sum_{k=1}^{K} \lambda_{(k)} \;(E[t_{s(k)}] + E[t_{rkk}]\;)/K$$

The fraction of calls answered by a patrol unit in its own beat, Q_{kk}, decreases as the arrival rate of calls increases. When the average utilization for the district approaches or exceeds one, we find that calls in a beat are shared equally by all units.

Table 1 shows a set of runs for a three-beat district where the service rates are about the same for calls in different beats but the arrival rates are not. In fact, the arrival rate in beat 1 in one and a half times the arrival rate in beat 2 and three times the arrival rate in beat 3. It is seen that the fraction of calls answered by unit 1 in its own beat, Q_{11}, is smaller than the fraction of calls answered by unit 2 in its own beat, Q_{22}, which in turn, is smaller than the fraction of calls answered by unit 3 in its own beat, Q_{33}. This happens because more calls arrive in beat 1 than in beat 2 or 3 and thus units 2 and 3 are available to answer calls in beat 1 when unit 1 is busy.

A probabilistic assignment policy, $w_{k\ell_1\ell_2\ell}$, is input to the simulation model. This is determined by examining a particular beat configuration to be simulated. All other parameters being equal, Q_{12} in a run is less than Q_{12} in another run if w_{12} (probability of questioning unit 2 regarding its availability for dispatch if call arrives in beat 1 and unit 1 is busy) in the first run is less than the w_{12} in the second run. For example, in a three-beat district for a run with w_{12} equal to zero, Q_{12} (fraction of calls in beat 1 answered by unit 2) was 0.0188 whereas when w_{12} was 0.5 a Q_{12} of 0.0528 was observed. Q_{12} is non-zero when w_{12} is zero because even though unit 3 is always questioned next regarding its availability for dispatch ($w_{12} = 0$, $w_{13} = 1$) there are cases when both units 1 and 3 are busy and unit 2 is assigned.

Table 1. Simulated $Q_{k\ell}$ Three Beats, Poisson Arrivals, General Service Time Distribution.

Run No.	Utilization ρ_{av}	$\lambda^{(k)}$ k=1	k=2	k=3	Q_{11}	Q_{12}	Q_{13}	Q_{21}	Q_{22}	Q_{23}
9A1	.0638	3	2	1	.8925	.0448	.0627	.0331	.9256	.0413
9A2	.1276	6	4	2	.7800	.1100	.1100	.0557	.8734	.0709
9A3	.1914	9	6	3	.7674	.1103	.1224	.0918	.7762	.1320
9A4	.2552	12	8	4	.6834	.1418	.1748	.1245	.7094	.1660
9A5	.3190	15	10	5	.5882	.1947	.2170	.1392	.6833	.1775
9A6	.3828	18	12	6	.5492	.2100	.2408	.1755	.6182	.2063
9A7	.4466	21	14	7	.5112	.2411	.2477	.1997	.5592	.2411
9A8	.5104	24	16	8	.4900	.2379	.2721	.2138	.5279	.2582
9A9	.5742	27	18	9	.4677	.2832	.2491	.2237	.5125	.2638
9A10	.6380	30	20	10	.4408	.2543	.3050	.2670	.4471	.2859
9A11	.7018	33	22	11	.4420	.2871	.2710	.3002	.4147	.2851
9A12	.7656	36	24	12	.4064	.2870	.3065	.2778	.4358	.2864
9A13	.8294	39	26	13	.3862	.3172	.2966	.3037	.3845	.3118
9A14	.8932	42	28	14	.3620	.3284	.3096	.2854	.4136	.3011
9A15	.9570	45	30	15	.3824	.3098	.3078	.3602	.3280	.3119
9A16	1.0208	48	32	16	.3434	.3214	.3352	.3098	.3647	.3255
9A17	1.0846	51	34	17	.3312	.3229	.3459	.3304	.3481	.3215
9A18	1.1484	54	36	18	.3434	.3204	.3362	.3244	.3509	.3248
9A19	1.2122	57	38	19	.3497	.3237	.3266	.3351	.3560	.3090
9A20	1.2766	60	40	20	.3414	.3416	.3170	.3097	.3645	.3258
9A21	1.9140	90	60	30	.3398	.3485	.3117	.3383	.3455	.3163

Total service time distribution of calls occurring in beat 1 = $8e^{-30t} + \frac{32}{3}e^{-40t} + \frac{16}{3}e^{-20t} + 10e^{-50t}$

Total service time distribution of calls occurring in beat 2 = $6e^{-30t} + 10e^{-40t} + 5e^{-20t} + 15e^{-50t}$

Table 1. (Continued).

Run No.	ρ_{av}	$\lambda_{(k)}$ k=1	k=2	k=3	Q_{31}	Q_{32}	Q_{33}
9A1	.0638	3	2	1	.0194	.0097	.9709
9A2	.1276	6	4	2	.0735	.0343	.8922
9A3	.1914	9	6	3	.0842	.0471	.8687
9A4	.2552	12	8	4	.1188	.0668	.8144
9A5	.3190	15	10	5	.1607	.1059	.7335
9A6	.3828	18	12	6	.2019	.1341	.6640
9A7	.4466	21	14	7	.2471	.1672	.5858
9A8	.5104	24	16	8	.2510	.2162	.5328
9A9	.5742	27	18	9	.2443	.2443	.5115
9A10	.6380	30	20	10	.2619	.2376	.5005
9A11	.7018	33	22	11	.3127	.2518	.4354
9A12	.7656	36	24	12	.3051	.2542	.4407
9A13	.8294	39	26	13	.3092	.3110	.3798
9A14	.8932	42	28	14	.2947	.3221	.3832
9A15	.9570	45	30	15	.3449	.3218	.3333
9A16	1.0208	48	32	16	.3308	.3092	.3599
9A17	1.0846	51	34	17	.3256	.3073	.3671
9A18	1.1484	54	36	18	.2969	.3205	.3825
9A19	1.2122	57	38	19	.3058	.3606	.3336
9A20	1.2766	60	40	20	.3261	.3331	.3408
9A21	1.9140	90	60	30	.3625	.3462	.2913

Total service time distribution of calls occurring in beat 3 = $6e^{-30t} + 12e^{-40t} + 4e^{-20t} + 15e^{-50t}$

$W_{12}=0.5$, $W_{21}=0.5$, $W_{31}=0.5$; $E[t_{s(1)}] + E[t_{r11}] = 0.0328$, $E[t_{s(2)}] + E[t_{r22}] = 0.0314$,

$E[t_{s(3)}] + E[t_{r33}] = 0.0302$

A total 114 runs each lasting for 100 shifts (an elapsed time of 10.4 years) were analyzed. The average difference between analytical values (calculated from equations such as (18),(19),(20), (21), and (24)) and simulated values of $Q_{k\ell}$ was 5.5 percent.

LEGENDS FOR FIGURES

Figure Number		Page
1	Percent of Arrest in Relation to Overall Response Time[+]	3
2	A Dispatch Policy for a Two-Beat District	24
3	A Dispatch Policy for a Three-Beat District. Arrivals in Beat 1	32

END