OPERATIONS RESEARCH CENTER

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

A Response Time Estimator for Police Patrol Dispatching

by

Christian Schaack
Richard C. Larson

OR 126-84 May 1984

The research for this report was funded in part by the National Institute of Justice on grant number 83-IJ-CX-0065.

The opinions expressed in this report are those of the authors and do not necessarily reflect the views and opinions of the sponsor.

ABSTRACT

Several recent police studies have focused on the delay in police response for both high and low priority calls. There is evidence, especially for low priority calls, that citizen satisfaction can be significantly improved by providing knowledge about response delay to the caller while he is still on the phone. We develop a mathematical expression for a response time estimator that can be easily implemented on a police department's computer aided dispatch (CAD) system. This delay estimate for an incoming call is based on the call's priority and the number and status of other calls already in the system. The estimator is shown to be mathematically consistent in two special cases that can be solved exactly. Computational results show how the procedure developed can be helpful in providing the caller with useful information about the expected response delay.

TABLE OF CONTENTS

1. Background

Recent research projects like the Kansas City Response Time Study[5] and the Wilmington Split-Force Study[4] have shown that rapid police response is important only for a small minority of calls, approximately 10 to 15 percent of all calls for service at a police department. For the majority of calls, citizen dissatisfaction is largely determined by the discrepancy between expected response time and actual response time. In light of these results, the traditional response "A patrol car will be there as soon as possible" leaves much to be desired. A computer aided dispatch (CAD) system ought to give the caller some estimate of the response time.

The purpose of this study, conducted under a National Institute of Justice (NIJ) grant (number 83-IJ-CX-0065), was to investigate the feasibility of such a response time estimation scheme.

We developed a model that simulates emergency calls of several types and priorities, and we tested out potential estimators of the response time to a call. Each call was evaluated, given the number and types of the calls already in queue or in service when the new call came in. The following simple dispatch policy was used: dispatch patrol cars to higher priority calls first. Within a priority class, dispatch them first-come-first-served. Service times have a type-dependent general probability distribution, and the priorities of every type of call are assumed to be known by the dispatcher. Service is assumed non-preemptive (i.e., all calls currently in service complete service, even if higher priority calls arrive while they are being served.

Every time an emergency call was generated, we estimated the response delay it would incur. When the actual delay became known, we compared it to the estimate.

Despite the high variability in the system at the typical load factors at which police departments tend to operate, we feel that the estimator we derived can be useful in providing the caller with an idea of the response delay.

Section 2 below describes our model and assumptions in more detail. Section 3 derives analytic results for two special cases of the response time estimation problem. Based on these special cases, in Section 4 we derive an estimator for the more general case. We also describe our conputational experience with the estimator. Section 5 discusses the policy implications and conclusions from this phase of the study. The Appendix gives a brief description of the simulation program used to derive the computational results.

2. Description of the Model

The police department's dispatch center is modeled as a queueing system where calls for service of different types arrive at a Poisson rate. Associated with every call type is a call priority and a service time probability distribution.

If, when a call arrives at the dispatch center, there is a patrol unit available, it is dispatched to service the call. If no unit is available, the call is queued. When a patrol unit terminates service on a call, it is immediately assigned to another call if the queue of backlogged calls is not empty. If there is a backlog, high priority calls are assigned a service unit first. Within a priority, the call that has been queued longest gets serviced first. Service on a low priority call is not interrupted when a higher priority call arrives. In queueing theory jargon, this service discipline is referred to as non-preemptive, head-of-the-line, first-come-first-served (FCFS) within a priority.

We outht to point out that this service discipline can be improved on. In practice, it pays to hold patrol units in reserve, in anticipation of high priority calls in the near future, rather than to assign the last available units to low priority calls.

We assume that service time on a call has a general probability distribution depending only on the type of call being serviced. This fact enables one to model situations where travel time to and from an incident is negligible compared to on-scene service time, or where the patrol unit is required to return to its home base before its next assignment. The latter requirement may be inappropriate for police operations but applies to other emergency

services like ambulances. Therefore, we will assume for the purpose of this study, that travel times are negligible, thereby dispensing with the spatial aspect of the problem.

3. Analytic Results

In this section we derive some analytic results for two special cases of the general model described above: the case of a single patrol unit and the case of multiple patrol units using identical exponential service times for all call types. To avoid tedious repetitions of the definitions, we shall from hereon refer to the general model as the $M/G_i/m$ case, the single patrol unit model as the $M/G_i/1$ case, and the identical exponential service model as the M/M/m case. We shall use the following notation:

T ≡ number of call types

λ_i ≡ arrival rate of type i calls (for i=1..T)

$G_i(x)$ ≡ service time probability density function (pdf) of type i calls (for i=1...T)

$\overline{x_i}$ ≡ mean service time of type i calls

$\overline{x_i^2}$ ≡ mean squared service time of type i calls

$p(i)$ ≡ priority level of type i calls (for i=i...T) By convention, type i has a higher priority than j if $p(i) < p(j)$.

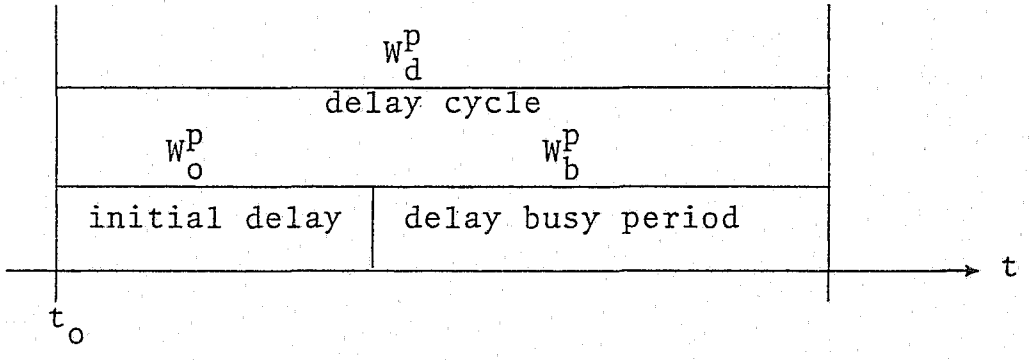
m ≡ number of patrol units (servers)

The purpose of this study is to derive an estimator for the expected delay incurred by an incoming call based on the number of calls (their types, priorities, etc.) currently in the system. All this information is available to the operator who takes the call via the CAD-systems utilized by most police departments.

Below, we derive closed form expressions for the expected delay for the prioritized $M/G_i/1$ and M/M/m systems. (Unfortunately, we believe that closed form results do not exist for the general $M/G_i/m$ case.)

The Single Server Case ($M/G_i/1$)

The analysis proceeds along a delay cycle approach (e.g., Kleinrock[1]). Suppose a call of priority p comes in at time t_o. For simplicity we shall refer to this call as call "p" (or customer "p").



The delay, W_d^p, incurred by "p" is equal to the sum of the initial delay, W_o^p, incurred because of calls already in the system at time to, plus the delay busy period, W_b^p, incurred because of higher priority arrivals during W_o^p: $W_d^p = W_o^p + W_b^p$.

W_o^p is the sum of the service times of higher or equal priority calls in queue at t_o plus the residual service time of the call being serviced at t_o, irrespectively of the latter's priority because the service discipline is non-preemptive.

The second term, W_b^p, is the sum of the sub-busy periods corresponding to the calls that arrive during W_o^p, the initial delay. The computation of the distribution of W_d^p will be performed in transform domain.

Let K_i be the random variable "number of type i calls arriving during W_o^p". Then, using an independence argument, we can write the Laplace-transform of the delay, conditioned on W_o^p and K_i as:

$$E\left[e^{-sW_d^p} | W_o^p=y,\ K_i=k_i,\ i=1\ldots T,\ p(i)<p\right] = e^{-sy} \prod_{\substack{i=1 \\ p(i)<p}}^{T} (P_i^T(s))^{ki}$$

where $P_i^T(s)$ is the Laplace transform of the sub-busy period associated with one arrival of type i during y.

Deconditioning on K_i, we find:

$$E\left[e^{-sW_d^p} | W_o^p=y\right] = e^{-sy} \prod_{\substack{i=1 \\ p(i)<p}}^{T} \left[\sum_{k_i=0}^{\infty} (P_i^T(s))^{k_i} \frac{e^{-\lambda_i y}(\lambda_i y)^{k_i}}{k_i!}\right]$$

$$= e^{-sy} \prod_{\substack{i=1 \\ p(i)<p}}^{T} e^{-(\lambda_i - \lambda_i P_i^T(s))y}$$

$$= e^{-(s+\sum_{\substack{i=1 \\ p(i)<p}}^{T} (\lambda_i - \lambda_i P_i^T(s)))y}$$

And finally, deconditioning on W_o^p, we find:

$$D_p^T(s) \equiv E\left[e^{-sW_d^p}\right] = I_p^T\left(s+\sum_{\substack{i=1 \\ p(i)<p}}^{T} (\lambda_i - \lambda_i P_i^T(s))\right) \qquad (3.1)$$

where $I_p^T(s)$ is the Laplace transform of the pdf of W_o^p. $I_p^T(s)$ is easily obtained by multiplying the Laplace transforms of the (residual) service times of the appropriate customers in the system at time t_o, just before call "p" arrives.

So the only unknowns that remain in our derivation are the P_i^T's. Since the sub-busy delay incurred by "p" because of a type i arrival during W_o^p can be decomposed into the service time of "i" and another sub-busy delay, it is readily seen that $P_i^T(s)$ obeys

an equation analogous to (3.1):

$$P_i^T(s) = G_i^T\left(s+\sum_{\substack{j=1\\ p(j)<p}}^{T}(\lambda_j-\lambda_j P_j^T(s))\right) \qquad \text{for} \quad \begin{vmatrix} i=1\ldots T \\ p(i)<p \end{vmatrix} \qquad (3.2_i)$$

where G_i^T is the transform of $G_i(x)$.

Equations (3.1) and (3.2_i) together determine the pdf of W_d^p, the delay incurred by "p".

In fact, equations (3.2_i) can be collapsed into a single equation -- (3.2) for the purpose of computing $D_p^T(s)$, by multiplying (3.2_i) by λ_i and summing over i:

Let $\alpha_p^T(s) \equiv \sum_{\substack{i=1\\ p(i)<p}}^{T} \lambda_i P_i^T(s)$,

then

$$\alpha_p^T(s) = \sum_{\substack{i=1\\ p(i)\ p}}^{T} \lambda_i G_i^T\left(s+\sum_{\substack{j=1\\ p(j)<p}}^{T}(\lambda_j-\lambda_j P_j^T(s))\right)$$

Setting

$$\Lambda_p^T(s) \equiv \sum_{\substack{i=1\\ p(i)<p}}^{T} \lambda_i G_i^T(s) \qquad \underline{\text{and}} \qquad \lambda_T^p \equiv \sum_{\substack{i=1\\ p(i)<p}}^{T} \lambda_i$$

we get:

$$\alpha_p^T(s) = \Lambda_p^T(s+\lambda_T^p-\alpha_p^T(s)) \qquad (3.2)$$

So, in summary, to find $D_p^T(s)$ we must solve the following two equations:

$$\begin{cases} D_p^T(s) = I_p^T(s+\lambda_T^p-\alpha_p^T(s)) & (3.1) \\ \alpha_p^T(s) = \Lambda_p^T(s+\lambda_T^p-\alpha_p^T(s)) & (3.2) \end{cases}$$

It is in general, difficult to get a closed form solution for $D_p^T(s)$, but it is easy to derive the first moments of W_d^p given the state of the system Ω from (3.1) and (3.2).

Differentiating (3.1) and (3.2) with respect to s and setting s to zero, we find that the conditional expected value of the delay is given by:

$$E\left[W_d^p|\Omega\right] = \frac{E\left[W_o^p|\Omega\right]}{1 - \sum_{\substack{i=1 \\ p(i)<p}}^{T} \lambda_i \overline{x}_i} \tag{3.3}$$

This equation says that the expected delay incurred by call "P", given Ω, (the state of the system upon arrival), is equal to the expected delay that "p" would incur if he only faced customers already in the system upon his arrival, multiplied by a constant factor accounting for potential incoming higher priority calls.

One can similarly derive the conditional variance:

$$Var\left[W_d^p|\Omega\right] = \frac{Var\left[W_o^p|\Omega\right]}{\left(1 - \sum_{\substack{i=1 \\ p(i)<p}}^{T} \lambda_i \overline{x}_i\right)^2} + \frac{E\left[W_o^p|\Omega\right] \sum_{\substack{i=1 \\ p(i)<p}}^{T} \lambda_i \overline{x_i^2}}{\left(1 - \sum_{\substack{i=1 \\ p(i)<p}}^{T} \lambda_i \overline{x_i}\right)^3} \tag{3.4}$$

The variance breaks up into two terms, the latter of which is due to the variability in the arrival process.

Higher order moments can be derived with the same ease. Let us now look at one of the rare cases for which (3.1)(3.2) can be solved in closed form:

Assume

$$G_i(x) = \mu e^{-\mu x}, \ \forall i=1...T. \quad \left(\text{Then } \overline{x_i} = \frac{1}{\mu} \equiv \overline{x}\right)$$

It is easy to derive, then, that

$$D_p^T(s) = \left(\frac{\lambda_T^p+\mu+s- \sqrt{(\lambda_T^p+\mu+s)^2-4\lambda_T^p\mu}}{2\lambda_T^p} \right)^N \qquad (3.5)$$

where N is the number of calls in the system at t_o, that will be served before "p". This result leads us very naturally to:

*the identical exponential service case with multiple servers (M/M/m)

Indeed, it is easy to see that the M/M/m system admits the solution:

$$D_p^T(s) = \left(\frac{\lambda_T^p+\mu+s- \sqrt{(\lambda_T^p+\mu+s)^2-4\lambda_T^p\mu}}{2\lambda_T^p} \right)^{N-(m-1)} \quad \text{For } N \geq m-1 \qquad (3.6)$$

From (3.6) we immediately deduce

$$E\left[W_d^p|\Omega\right] = \frac{(N-(m-1))}{m\mu\left(1-\frac{\lambda_T^p}{m\mu}\right)} \qquad (3.7a)$$

or

$$E\left[W_d^p|\Omega\right] = \frac{E\left[W_o^p|\Omega\right]}{1 - \sum_{\substack{i=1 \\ p(i)<p}}^{m} \lambda_i \bar{x}} - \frac{m-1}{m}\bar{x} \qquad (3.7b)$$

Unfortunately, all our attempts to generalize the $M/G_i/1$ model to multiple servers or the M/M/m model to different service rates for different call types have failed.

However, the results of this section, especially when we compare equations (3.3) and (3.7b), give us valuable insights into the complexity of the problem and support an estimator for the more general $M/G_i/m$ case.

4. Waiting Time Estimation and Computational Results

In police applications, service times for a particular incident typically have means of 20 to 30 minutes and coefficients of variation of around 0.5.

For this section, we chose values of the system parameters appropriate for the modeling of the operation of a police department. We restricted the discussion to two types of calls: type 1, high priority calls, and type 2, low priority calls. We also chose the ratio of high to low priority calls equal to 10 percent, a typical value for police operations.

Table 1 below summarizes typical values of system parameters for police operations.

Parameter	Value	Description
m	10	number of patrol units
$\overline{x_i}$	20 to 30 minutes	average service time
CV_i	0.5	coefficient of variation of the service time distribution
ρ	65% to 75%	system load factor
$\frac{\lambda_1}{\lambda_2}$	10%	ratio of high to low priority calls

Table 1

4.1 Estimating Waiting Times

The reason the delay cycle method doesn't generalize to multiple servers is that the waiting time of customer "p" is not a sum of independent random variables when m>1, which made the derivations in transform domain of Section 3 possible. With multiple servers, the ordering of the calls in queue becomes

important. This, however, makes analytic derivations intractable.

We shall briefly describe a first, in retrospect rather short sighted estimator, that we tried and discarded. It helped us understand some of the intricacies of the problem. This estimator works roughly as follows: When call "p" arrive, to compute its estimated delay, simulate the system as if service and arrival times were deterministic with respective values $\overline{x_i}$ and $\frac{1}{\lambda_i}$. While this estimator does a good job for the single server case, it behaves poorly for multiple servers. As an example, take an M/M/m queue with a single call type with service rate $\mu = \frac{1}{10}$ and with 10 servers. Given N, the number of calls in the system, an arriving call's expected waiting time is given by:

$$E[W/N] = \frac{N-(m-1)}{m\mu} \quad \text{for } N \geq m.$$

Let $\hat{W}_N$ denote the "deterministic" estimator described above. Table 2 compares $\hat{W}_N$ and E[W/N] for various values of N:

N	10	11	12	13	14	15	16	17	18	19	20	21	22
$\hat{W}_N$	10	10	10	10	10	10	10	10	10	10	20	20	20
E[W/N]	1	2	3	4	5	6	7	8	9	10	11	12	13

Table 2

Note the high relative error when $\hat{W}_N$ is used. Similar results can be shown to hold for M/M/m systems with several call types. The high relative error of this estimator makes it rather undesirable. We therefore rejected it. This experience shows that we cannot disregard the variability of the system (at

medium load factors, anyway).

The estimator we finally settled for is based on the results of Section 3. Equations (3.3) and (3.7b) suggest an estimate of the form:

(4.1)
$$\hat{W}_d^p = \frac{E\left[W_o^p \mid \Omega\right]}{m - \sum_{\substack{i=1 \\ p(i)<p}} \lambda_i \bar{X}_i} - b_p$$

where b_p is a constant depending on the priority p. We shall come back to the determination of b_p in Section 4.2.

From the start, the estimator defined by (4.1) has a definite advantage over the "deterministic" estimator: it is exact (i.e., $\hat{W}_d^p = E[W_d^p|\Omega]$) for one type of multiple server queueing system, the prioritized M/M/m system that yield equation (3.7b). While equations (3.3) and (3.7b) tell us that the estimator (4.1) is consistent with the special cases that we could tackle analytically, practical experience must show how well this estimator works in practice.

4.2 Computational Experience

We essentially tested out the general behavior of the estimator given by (4.1) under the following assumptions and parameter values:

- For ease of interpretation we confined ourselves to two call types.
- 10 servers
- Service times Erlang distributed of order 4 (the forth-order Erlang distribution fits the general pattern of the service distributions encountered in practice; it also exhibits a coefficient of variation of 0.5).

- $\frac{\lambda_1}{\lambda_2} = 0.1$: ten percent of the calls are high priority calls. (We also ran a few simulations with $\frac{\lambda_1}{\lambda_2} = 0.5$. Since they add no new insights, they are omitted here.)
- Load factors ranging from 65 to 85 percent.
- Mean service times of 20 and 30 minutes. In practice, these times are typical. Average service time for high priority calls may be slightly smaller than for low priority calls in practice (i.e. runs 6 to 10 in Table 3).

Table 3 summarizes the parameters distinguishing the various simulations we ran:

Run #	1	2	3	4	5	6	7	8	9	10
load factor ρ (in%)	65	70	75	80	85	65	70	75	80	85
(in minutes)	30	30	30	30	30	20	20	20	20	20
(in minutes)	30	30	30	30	30	30	30	30	30	30

Table 3

Since we didn't know how to determine b_p in (4.1), we proceeded as follows. For every set of parameters, we ran a Monte Carlo simulation generating 3,000-plus calls incurring positive waiting times. During a simulation run, each time a call had to be queued, we computed the first term of (4.1)

$$\frac{E[W_o^p|\Omega]}{m - \sum_{\substack{i=1 \\ p(i)<p}}^{T} \lambda_i \bar{x}_i}$$

(from hereon denoted as f_p), (i.e., $\hat{W}_d^p + b_p$).

This quality was paired with the actual waiting time when this became known, and dumped into a file. Figures 1 to 5 show plots of the waiting time, w, versus f_p, for p=2 (i.e., low priority calls) for runs 2, 3, 4, 6, and 10.

High priority calls typically get served very fast. In none of our runs did a high priority call wait more than 15 minutes, (as Figure 6 shows in the most unfavorable case, ρ=85%, $\bar{x}_1=\bar{x}_2=$ 30 minutes). Therefore, we shall concern ourselves exclusively with low priority calls for the moment. Indeed, these are the calls that incur the longer waiting times. It is also for these calls that citizen dissatisfaction is determined through the discrepancy between expected and actual delay. This is not so for high priority calls where fast response is primordial.

Therefore, we shall concentrate on the type 2 calls. Heteroscadicity of the data makes unweighted least squares regression of W on f a poor data exploration tool. Therefore, we proceeded as follows: We broke up the data into intervals of length 1.2 minutes along the f axis. We used these intervals as levels in an analysis of variance; that is, for a given interval of f, we computed the means and standard deviations of the waiting times of the data points contained in a vertical slice (on the plots in Figure 1 through 5) centered around f.

Figures 7 through 11 show plots of the group mean of the waiting times at level f versus f for runs 2, 3, 4, 6, and 10. Next we regressed the group-means on f. Table 4 summarizes the results of these regressions. The notation used is:

$$\text{group-mean}(f) = \alpha_o + \alpha_1 f.$$

Run #	1	2	3	4	5	6	7	8	9	10
α_o (in minutes)	-17.52	-17.17	-15.36	-17.40	-17.64	-14.82	-15.00	-14.64	-16.14	-16.38
α_1	1.001	0.975	0.891	0.975	0.975	0.923	0.920	0.901	0.965	0.958
R^2 (in %)	97.4	97.3	98.2	98.6	99.2	97.2	97.9	98.5	98.7	99.0

Table 4

Coefficients from regression of mean waiting time on estimated waiting time (all coefficients are significant). These results call for the following remarks:

(a) The slope α_1 is close to 1 (except for run 3), but seems to be consistently smaller than 1. A look at the plots (e.g., figure 7) shows a slight convexity of the plot. One expects the slope to be equal to 1 asymptotically, but not necessarily so at small waiting times. It is surprising that the slope is so close to 1 overall. In practice, one should in fact give the same weight to every level of (and not to every observation), thus giving relatively more weight to long waiting times since calls suffering long delays are more scarce, but more important. Figure 12 shows a plot of run 2 of the group mean for level f against f (one data point per level. Some levels had to be pooled to contain enough observations). The regression of the mean waiting time (given f) on f yields the coefficients: $\alpha_o = -17.85$

$\alpha_1 = 1.012$,

both very significant. Only one such analysis was performed because the data were difficult to get at, but we feel confident, in light of the other results, that 1 is indeed the value of α_1 when all levels are equally weighted.

(b) The intercept α_o seems to be insensitive to the load factor, at least for load factors in the range 65 to 85 percent as runs 1 to 5 and 6 to 10 in Table 4 show. Since $\alpha_1 \simeq 1$, we also have that $\alpha_o \simeq -b_p$. Therefore, our results seem to indicate that the correction factor b_p does not depend on system load. This in itself, is rather amazing, although equation (3.7b) shows that this is indeed so for the special case of the M/M/m system.

As a result, we think that there may be a way of computing b_p from the arrival rates and service distributions.

Figure 13 illustrates the output of the analysis of variance (ANOVA) performed on run 2. From this analysis, we get estimates of the standard deviation of the waiting time given f. Figure 14 plots the coefficient of variation for a given level f against the mean waiting time at the same level. Notice the downward sloping curve. The coefficient of variation gets smaller as the waiting time increases. This is not altogether unexpected, but is certainly a welcome confirmation of our intuition. This information can be used in practice to find confidence intervals for the estimated waiting time.

The last question that remains is how the estimator can be implemented in practice, and in particular, how b_p is determined. In actual police department implementation we suggest the following procedure. As a call for service of priority p comes in at time t, compute the first term of equation (4.1).

i.e., $f \equiv \dfrac{E\left[W_o^P \mid \Omega\right]}{m - \sum\limits_{\substack{i=1 \\ p(i)<p}}^{T} \lambda_i \bar{x}_i}$. The numerator $E\left[W_o^P \mid \Omega\right]$ is easily obtained by summing the expected residual service times of calls in service at time t and the expected service times of higher or equal priority calls in queue at time t (calls with priority $\leq$ p). The estimated waiting time at time t is given by (4.1), but originally b_p is unknown. (One can use a reasonable hunch for b_p during the first phase of the implementation, or wait until the system is calibrated before giving any response time estimates to callers.) The calibration of b_p is done as follows. As calls get served, one updates GM(f), the mean waiting time of calls for a given level of f (by levels we

understand small fixed intervals into which the values of f fall). When a few hundred calls have come in, one averages GM(f)-f for those levels that contain more than, say, 10 data points. This average $\overline{GM}(f)-\bar{f}$ yields the calibration factor b_p. In other words, the system is calibrated with data from real calls for service. After a few hundred calls, the value of b_p has settled down (this value of b_p can be periodically refined or updated). The system is now operational. Equation (4.1) provides the caller with a response time estimate*. The nice feature of the method outlined above is that the system fine-tunes itself as it gathers more data. To complement this estimate of the response time, one can similarly compute the standard deviation GS(f) of the waiting time for the levels of f to obtain confidence intervals on the estimator (4.1).

Finally, we would like to make one last comment. Our limited experience with the estimator (4.1) indicates that b_p might be rather insensitive to the load factor (as long as the ratio of high to low priority calls remains unchanged). This insensitivity may make frequent adjustments of b_p due to call frequency changes over a day unnecessary, but more research would be necessary to ascertain this fact.

---

* In fact, the estimate should be $\hat{W}_p^d = \max(0, f-b_p)$ instead of $f-b_p$ since it may occasionally happen that $f-b_p$ is negative.

5. Policy Implications and Conclusions

We conclude this paper with a few general comments and suggestions as to how to apply the results of this study in practice.

The premise for our recommendations that citizens experience dissatisfaction regarding police responses only when their expectations are inconsistent with the reality. Therefore, for high priority calls, it is unnecessary to give the caller a forecast of when a unit will respond to his call for help. Indeed, these calls are real emergencies. The standard response, "A patrol car will be there right away," fits this situation very well, especially in light of the short waiting times actually experienced by high priority calls (see Figure 6).

For low priority calls, we suggest the following strategy: For short expected waiting times it may still make sense to give the "We'll be there right away" response because of the large relative variance of our estimator in the range below 1.5 minutes (see Figure 14). Since most calls in this category would incur relatively short delays, this response is accurate enough given the nature of the call.

For the calls that are likely to incur longer waiting times, ($\hat{W}_d^p$ > 15 minutes), we suggest using $\hat{W}_d^p$ as a response time estimate. These callers are more likely to get upset when they experience long delays after they are led to expect a fast response. Estimating response time can be done by using $\hat{W}_d^p$ plus some multiple of the standard deviation to give the caller a conservative delay estimate. Alternately, the dispatcher can give the caller an interval of the form $[\hat{W}_d^p\ (1-C_v),\ \hat{W}_d^p\ (1+C_v)]$, where C_v is the

coefficient of variation. For example, the operator could say "A police car will be there in 30 to 45 minutes."

It is difficult to say which of these or possibly other alternative schemes would be best received by the public. The decision of how exactly to implement the delay forecast into day to day police operations lies with the policy makers at police departments. We cannot, without further surveying the public's preferences, recommend one alternative over another. However, we believe that giving just a point estimate $\hat{W}_d^p$ without accounting for the variance may be rather misleading and create unnecessary citizen dissatisfaction.

To conclude this paper, we will summarize our results. We modeled the dispatching operation of a police department by a prioritized non-preemptive head-of-the-line $M/G_i/m$ queueing system. We derived analytic results for the expected waiting time of a call for the special cases of a single server $M/G_i/1$ and a multiple server, identical exponential service M/M/m system. Based on these results we derived an approximation of the expected conditional waiting time for the general $M/G_i/m$ system. Finally, we analyzed the performance of our delay estimator in a number of typical simulation runs.

The estimator provides a useful method of calculating waiting times for low priority calls in a waiting range greater than 15 minutes. For high priority calls and for calls with waiting times of less than 15 minutes, an estimator is not necessary because the response time to these calls is quasi-immediate.

References

[1] Kleinrock, L., Queueing Systems Volumes I & II. John Wiley & Sons, Inc., New York, NY: 1975.

[2] Larson, R.C., Urban Police Patrol Analysis. Cambridge, MA: MIT Press.

[3] Larson, R.C., "Proactive Real-Time Management of Scarce Police Resources," A Proposal submitted to the National Institute of Justice, June 9, 1983.

[4] Tien, James M., R.C. Larson, et al, "An Evaluation Report: Wilmington Split Force Patrol Program," Public Systems Evaluation, Inc., Cambridge, MA: 1976.

[5] Van Kirk, M.L., "Kansas City Response Time Analysis Final Report," Vols. 1, 2, 3, Kansas City Police Department, Kansas City, MO: 1977.

Figures

Notation: On the following figures:

a * stands for 1 observation

a 2 through 9 represents 2 through 9 observations

a + stands for more than 10 observations

All times are expressed in minutes

Figure 1

Plot of Waiting Time vs. Uncorrected Estimator for Low Priority Calls (Run 2)

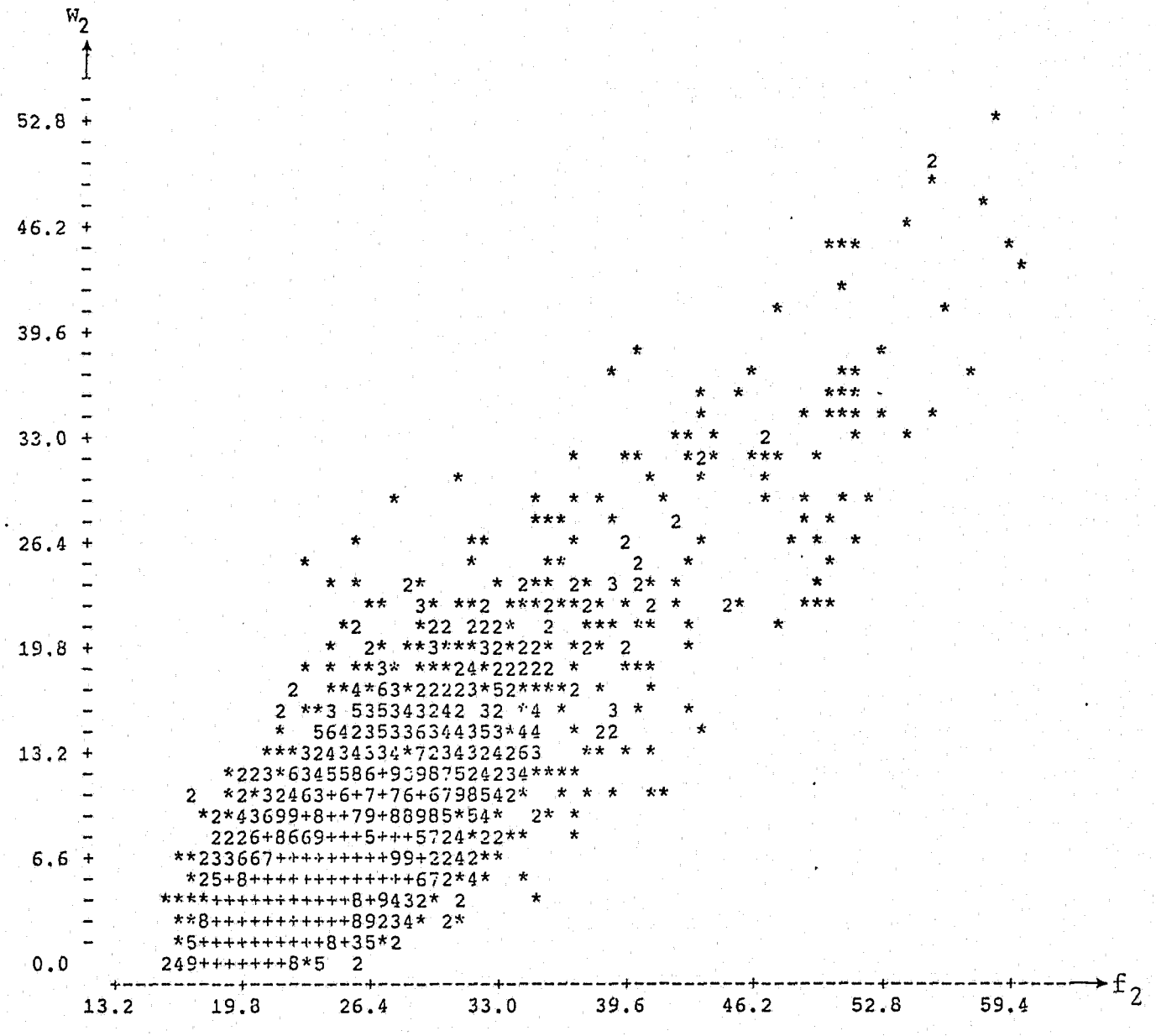


Figure 2

Plot of Waiting Time vs. Uncorrected Estimator for Low Priority Calls (Run 3)

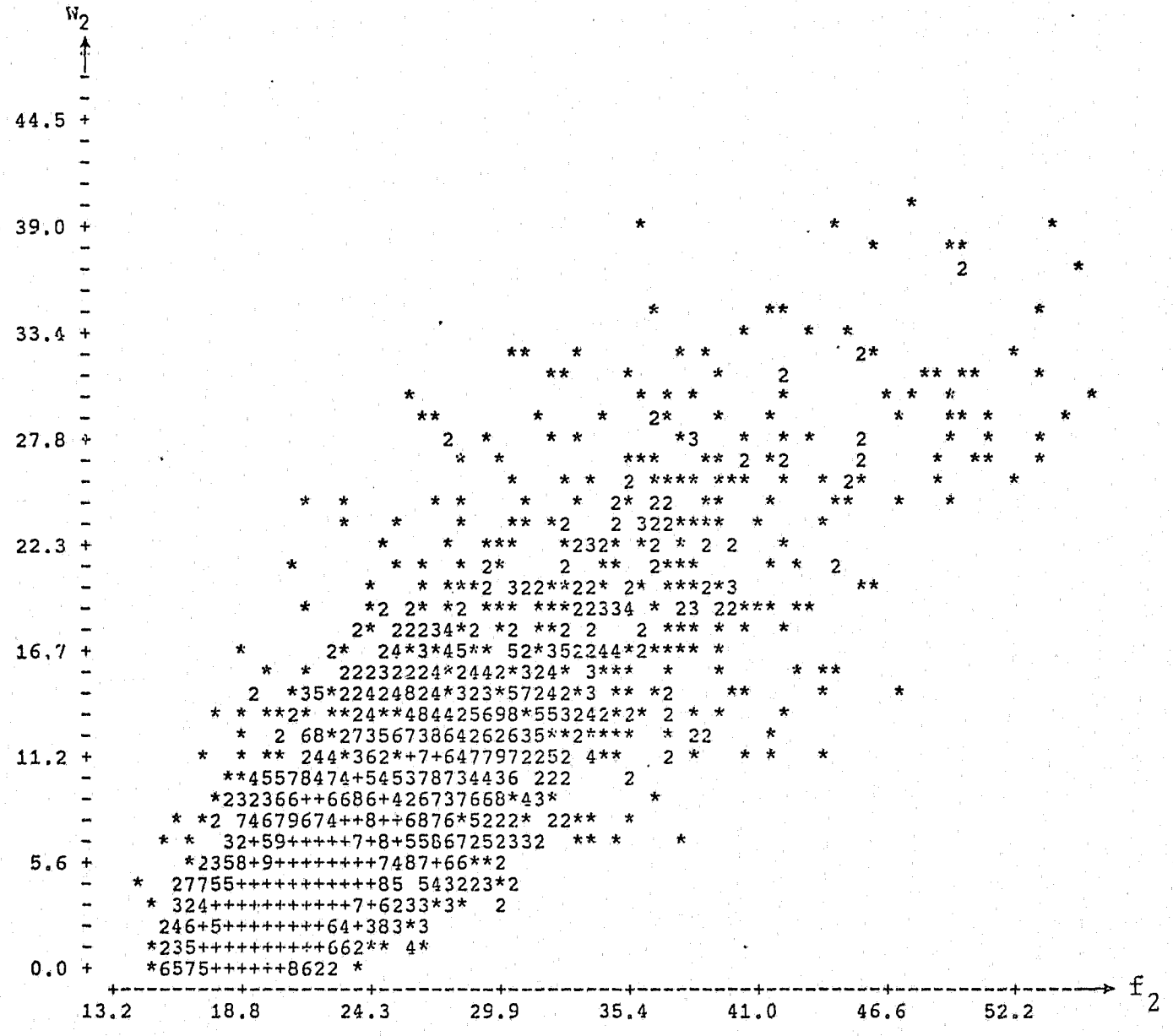


Figure 3

Plot of Waiting Time vs. Uncorrected Estimator for Low Priority Calls (Run 4)

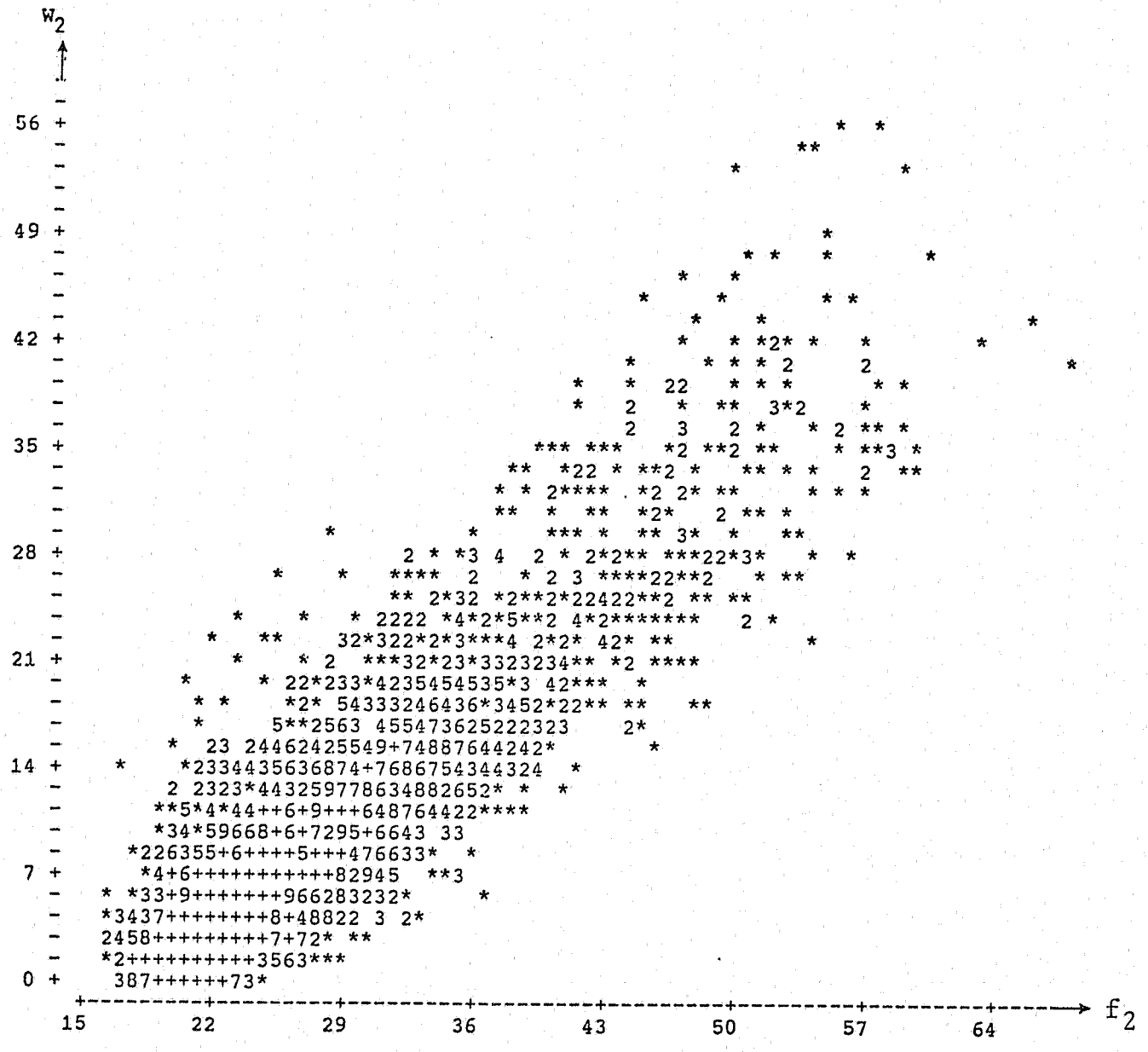


Figure 4

Plot of Waiting Time vs. Uncorrected Estimator for Low Priority Calls (Run 6)

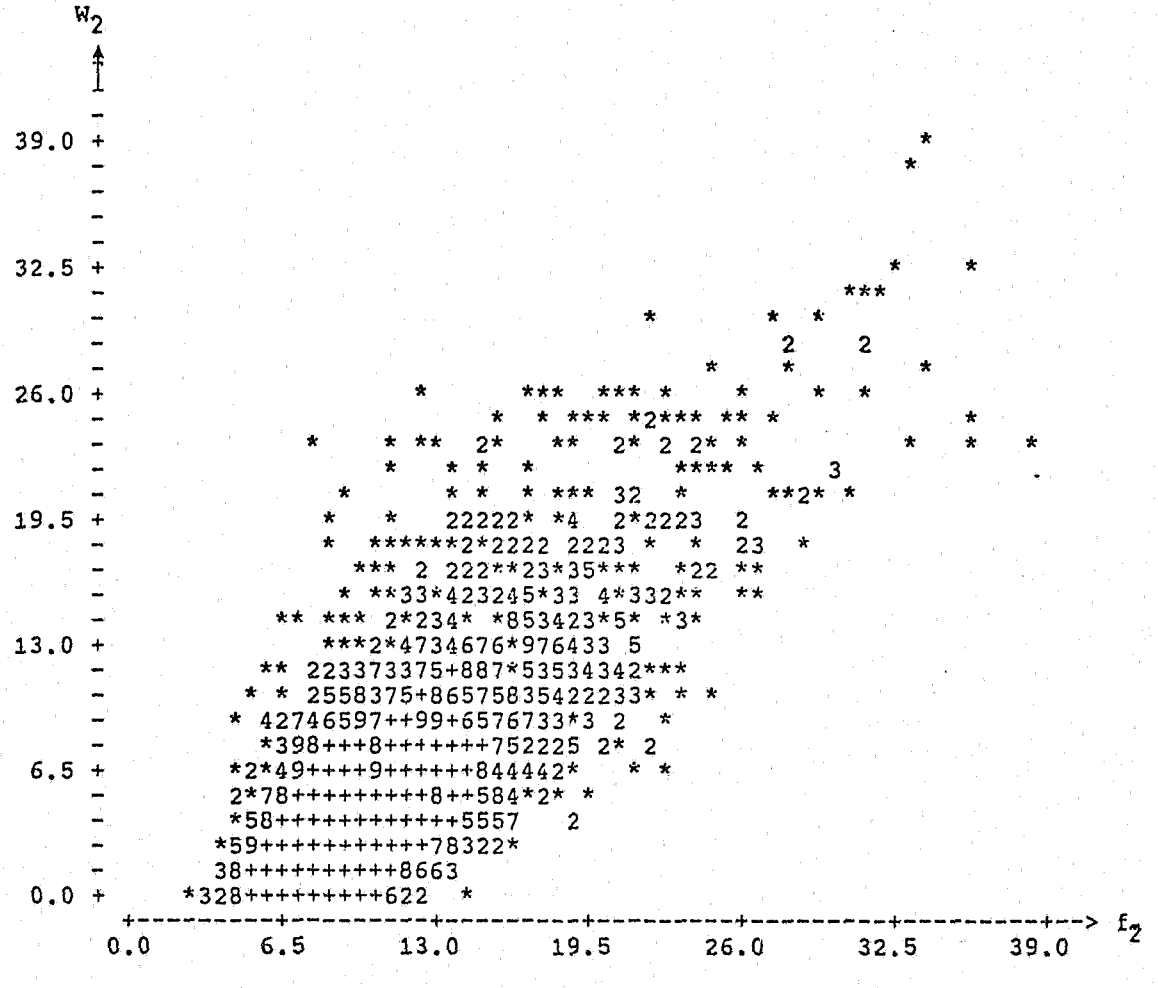


Figure 5

Plot of Waiting Time vs. Uncorrected Estimator for Low Priority Calls (Run 10)

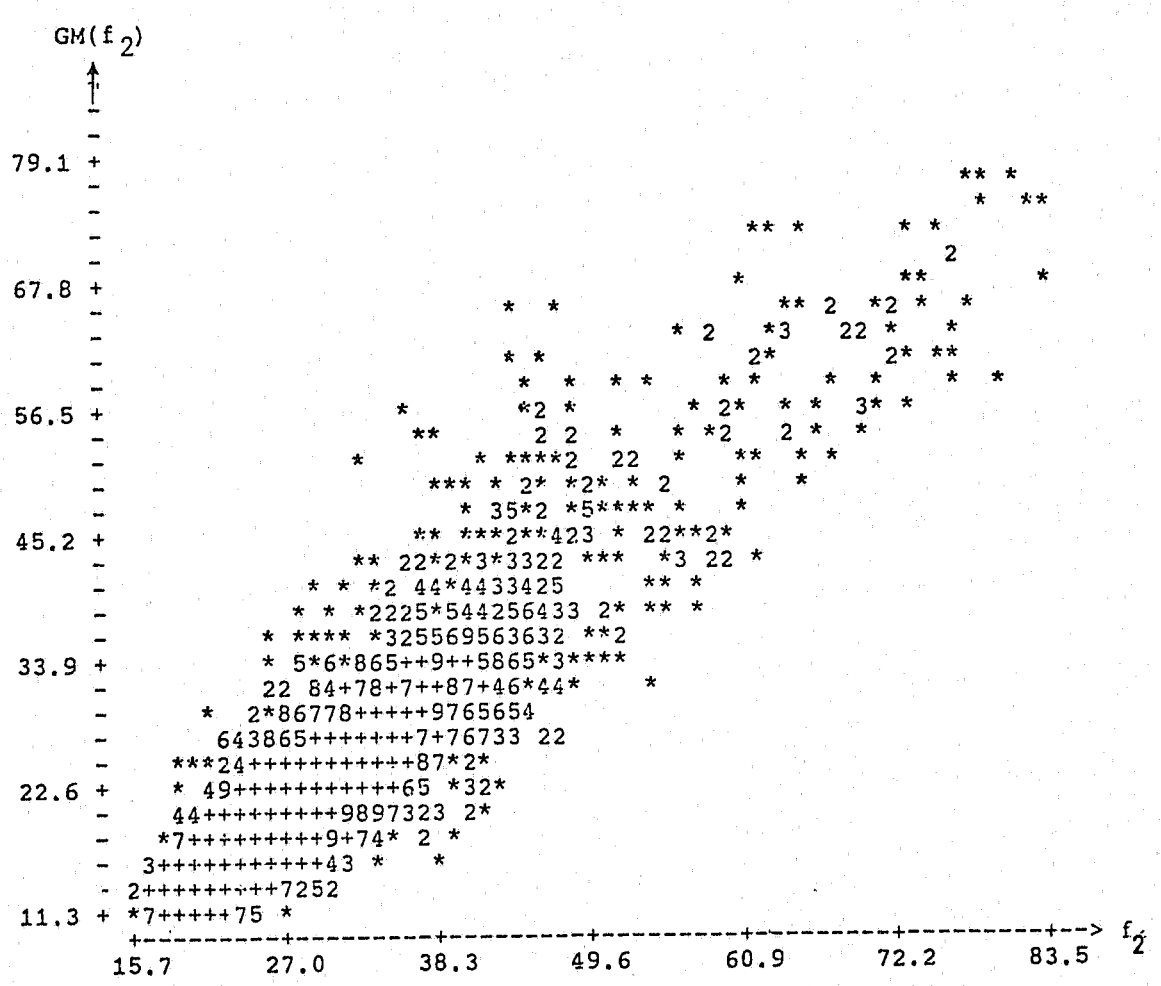


Figure 6

Plot of Waiting Time vs. Uncorrected Estimator for High Priority Calls (Run 5)

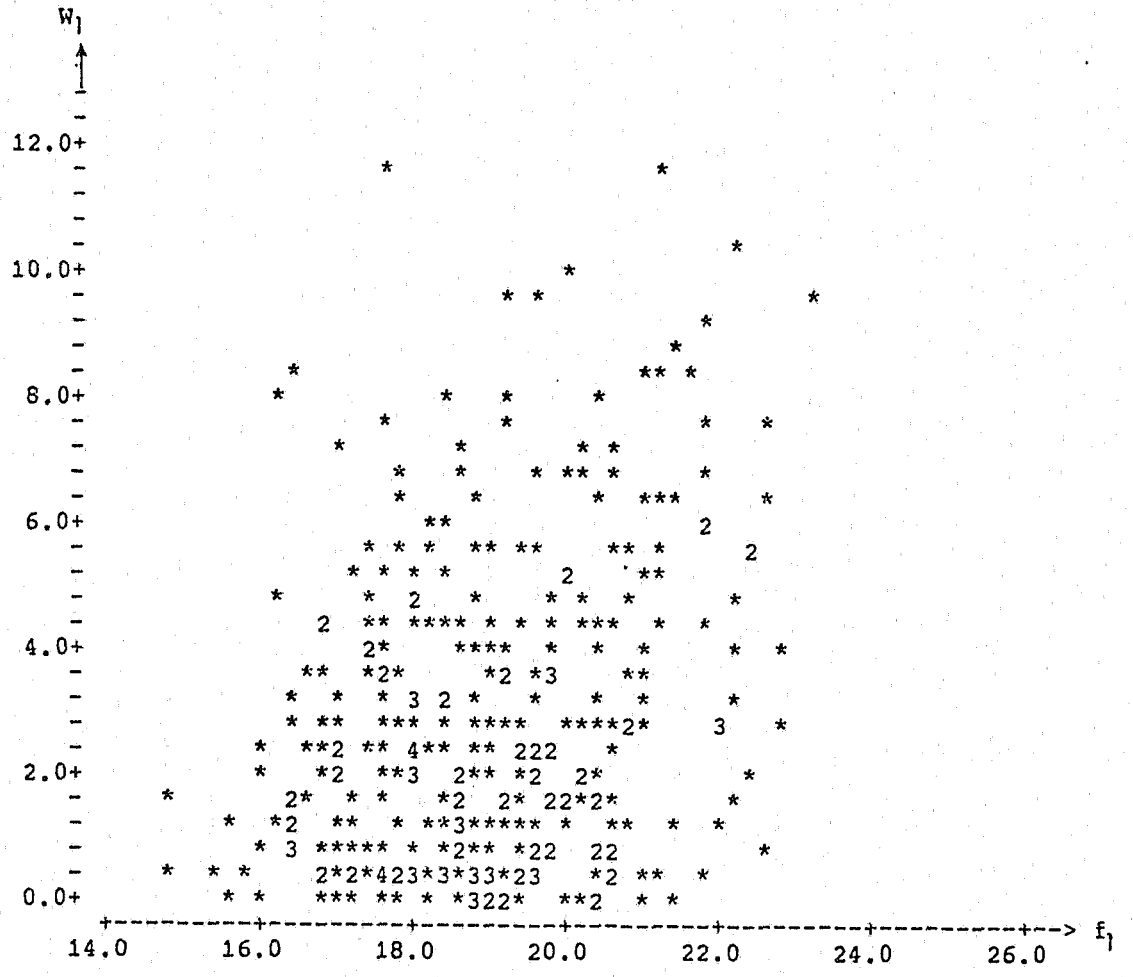


Figure 7

Plot of Group Mean Waiting Time (for f_2) vs.

Uncorrected Estimator, f_2 for Low Priority Calls (Run 7)

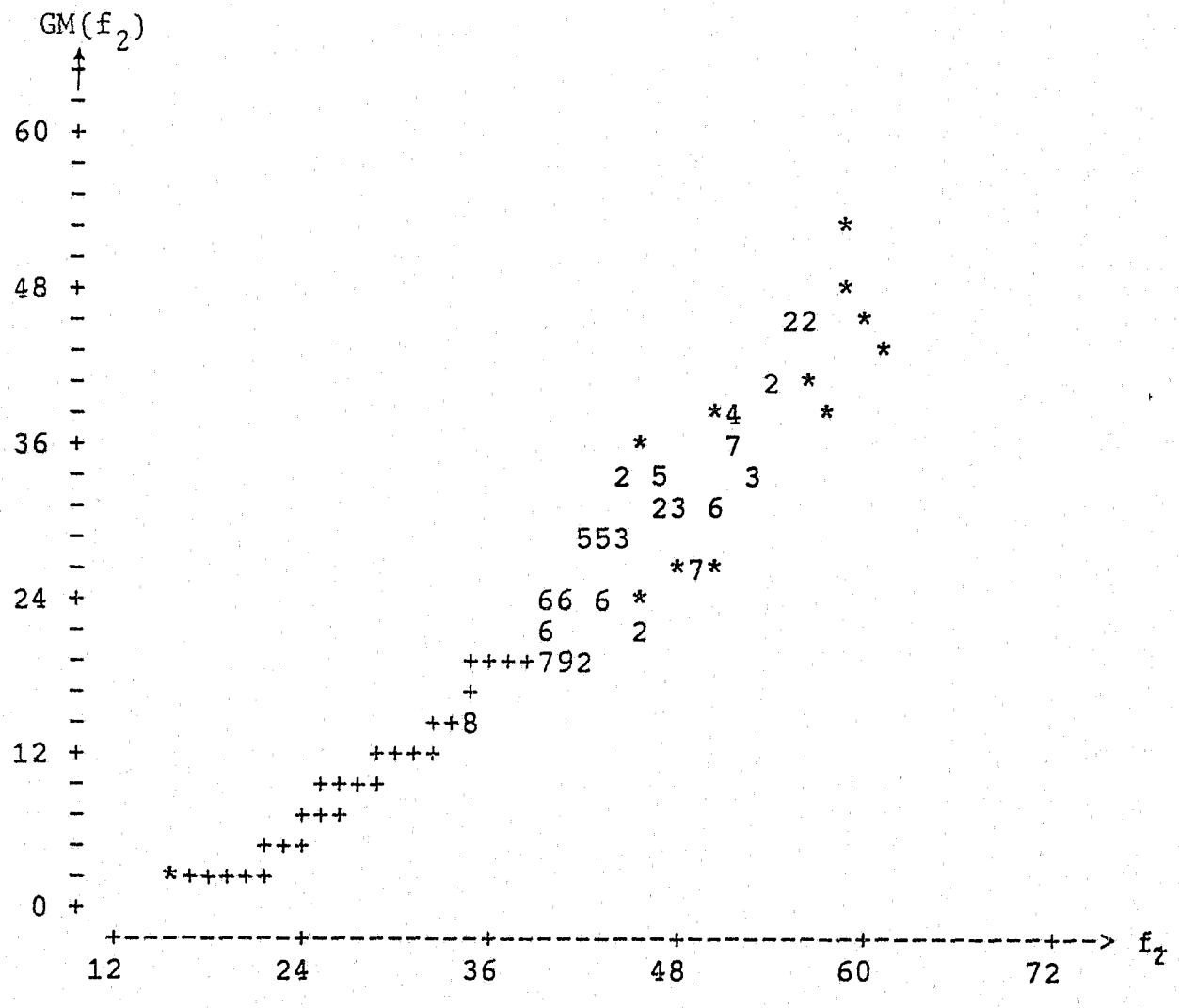


Figure 8

Plot of Group Mean Waiting Time (for f_2) vs.

Uncorrected Estimator, f_2 for Low Priority Calls (Run 3)

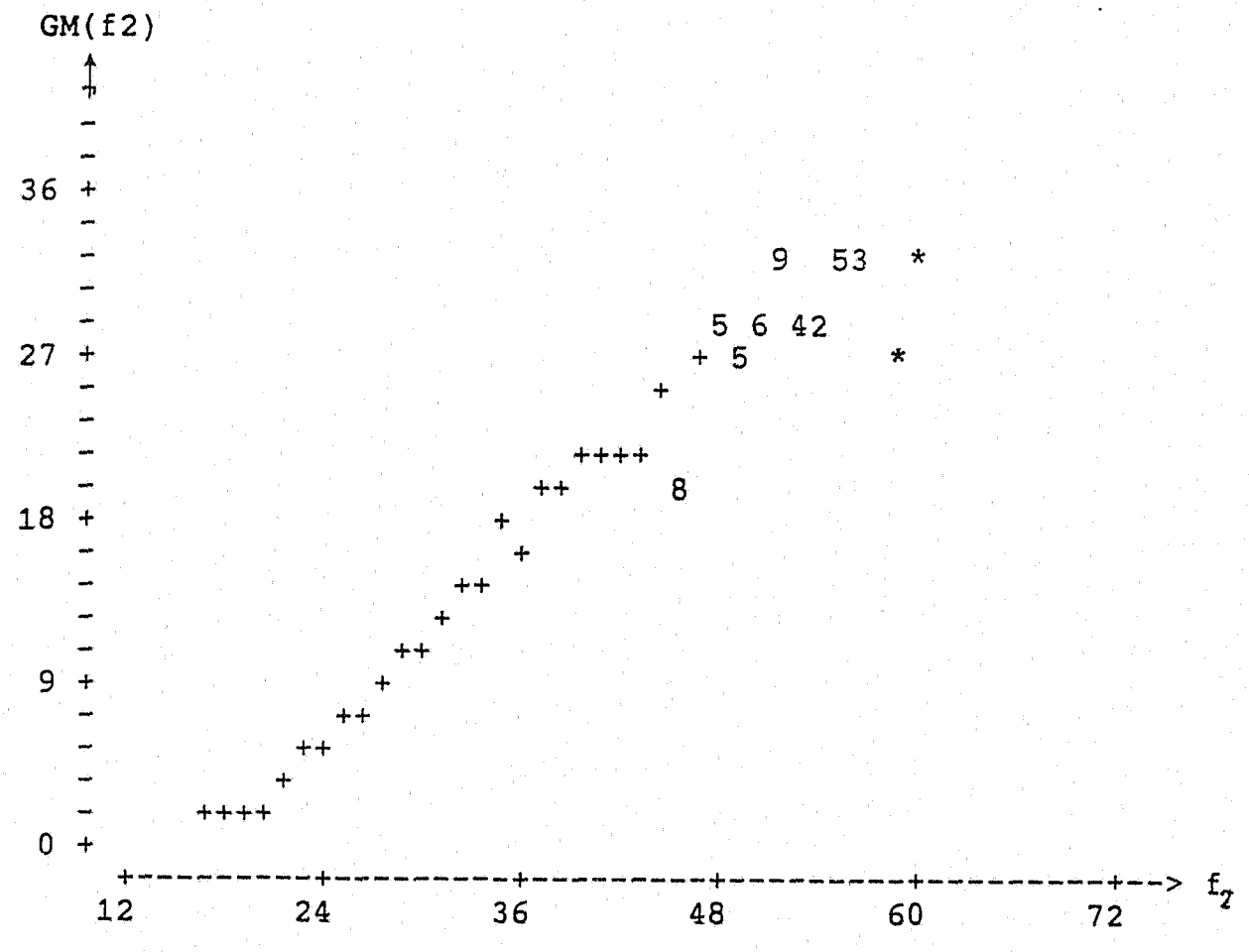


Figure 9

Plot of Group Mean Waiting Time (for f_2) vs.

Uncorrected Estimator, f_2 for Low Priority Calls (Run 4)

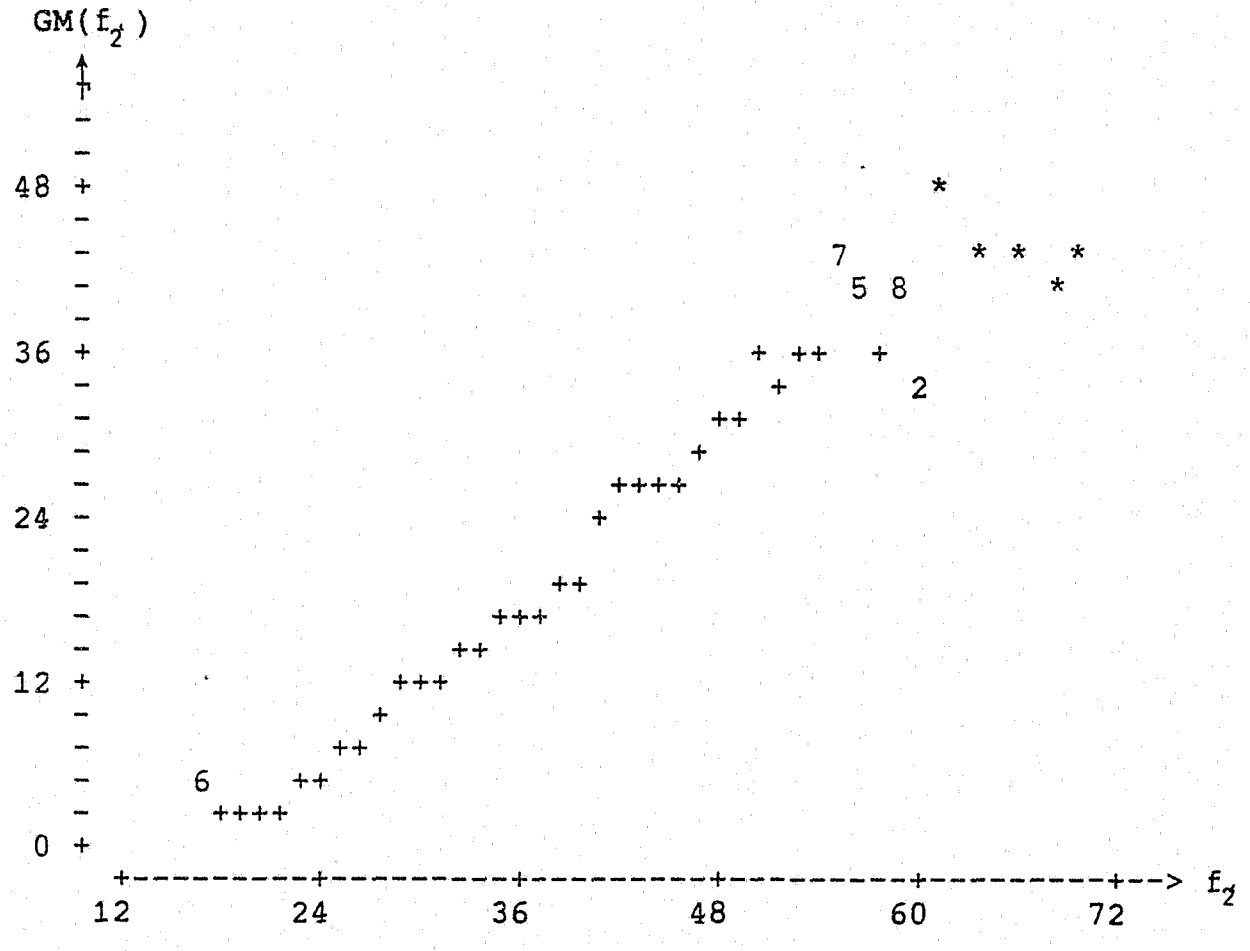


Figure 10

Plot of Group Mean Waiting Time (for f_2) vs.

Uncorrected Estimator, f_2 for Low Priority Calls (Run 6)

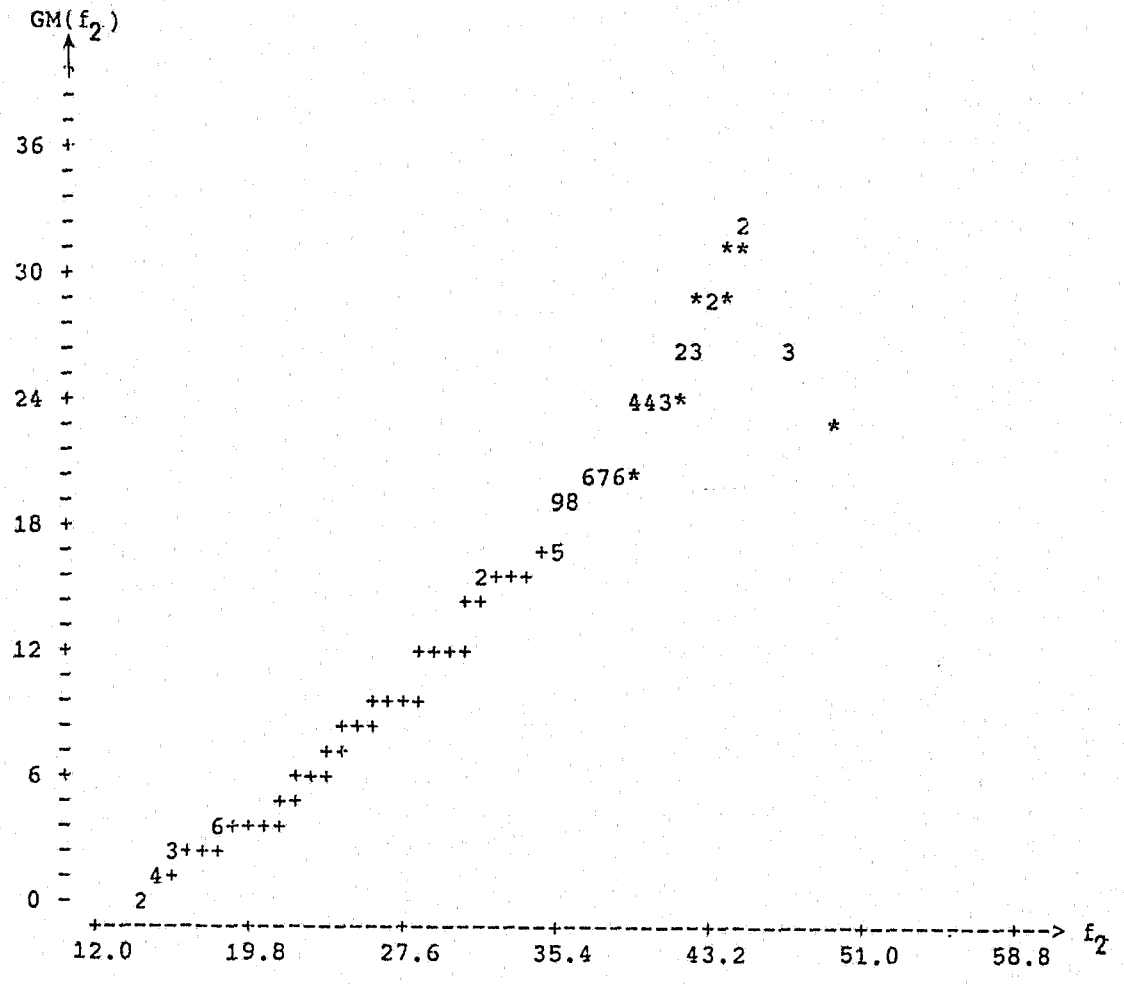


Figure 11

Plot of Group Mean Waiting Time (for f_2) vs.

Uncorrected Estimator, f_2 for Low Priority Calls (Run 10)

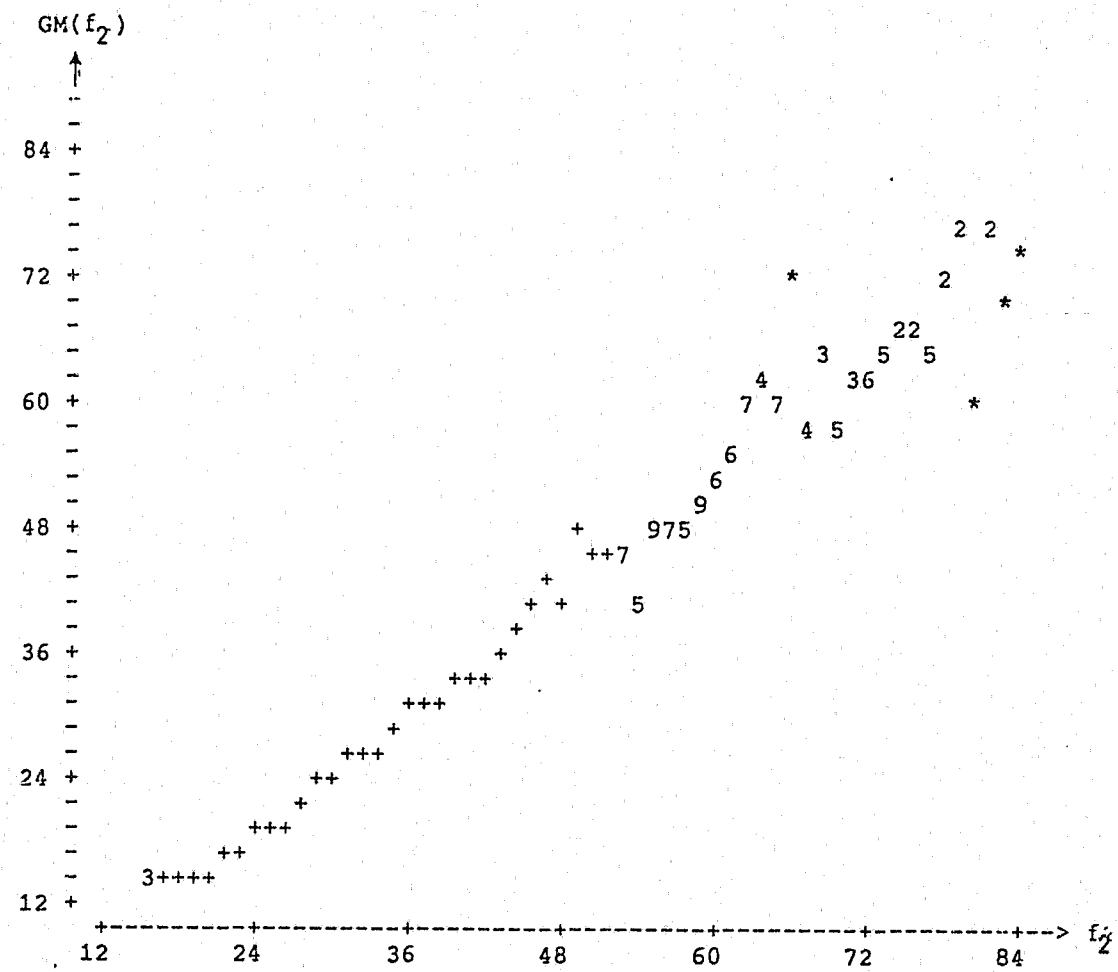


Figure 12

Plot of Group Mean Waiting Time vs. Level f_2

(One Data Point Per Level; Certain Levels Pooled)

(Run 2)

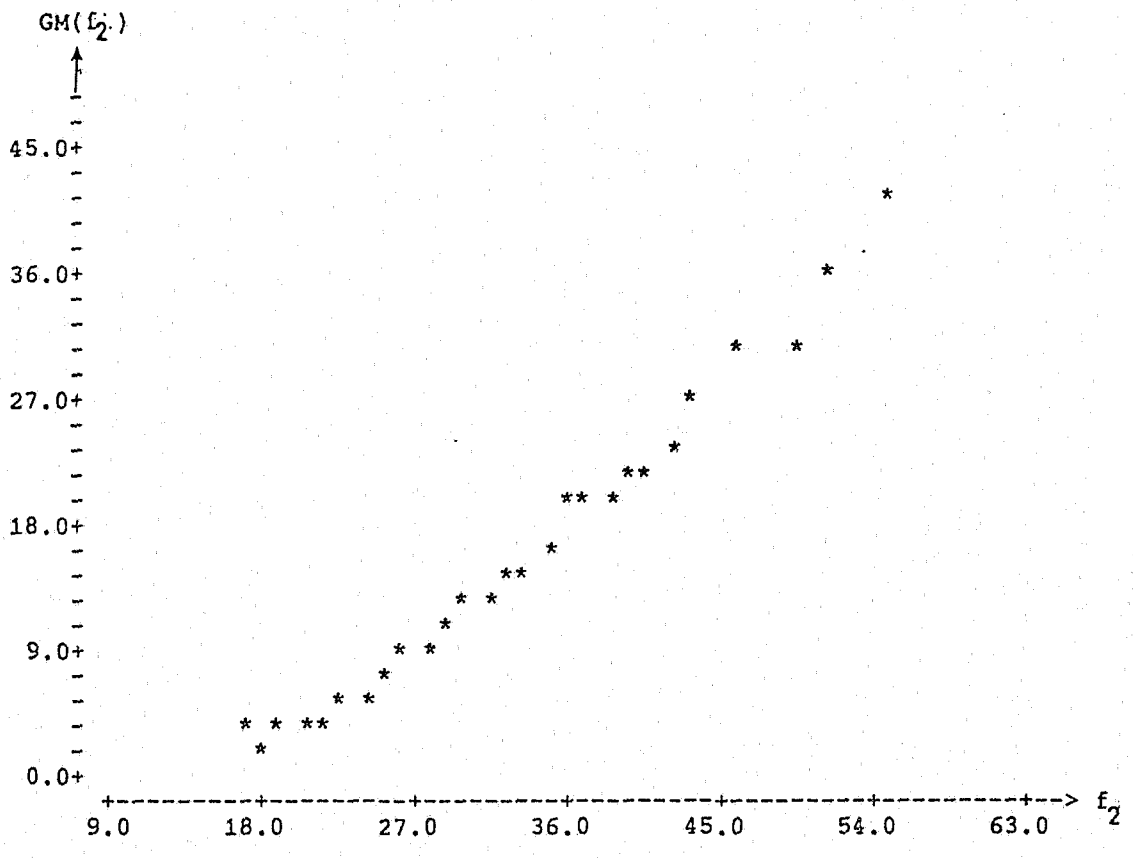


Figure 13

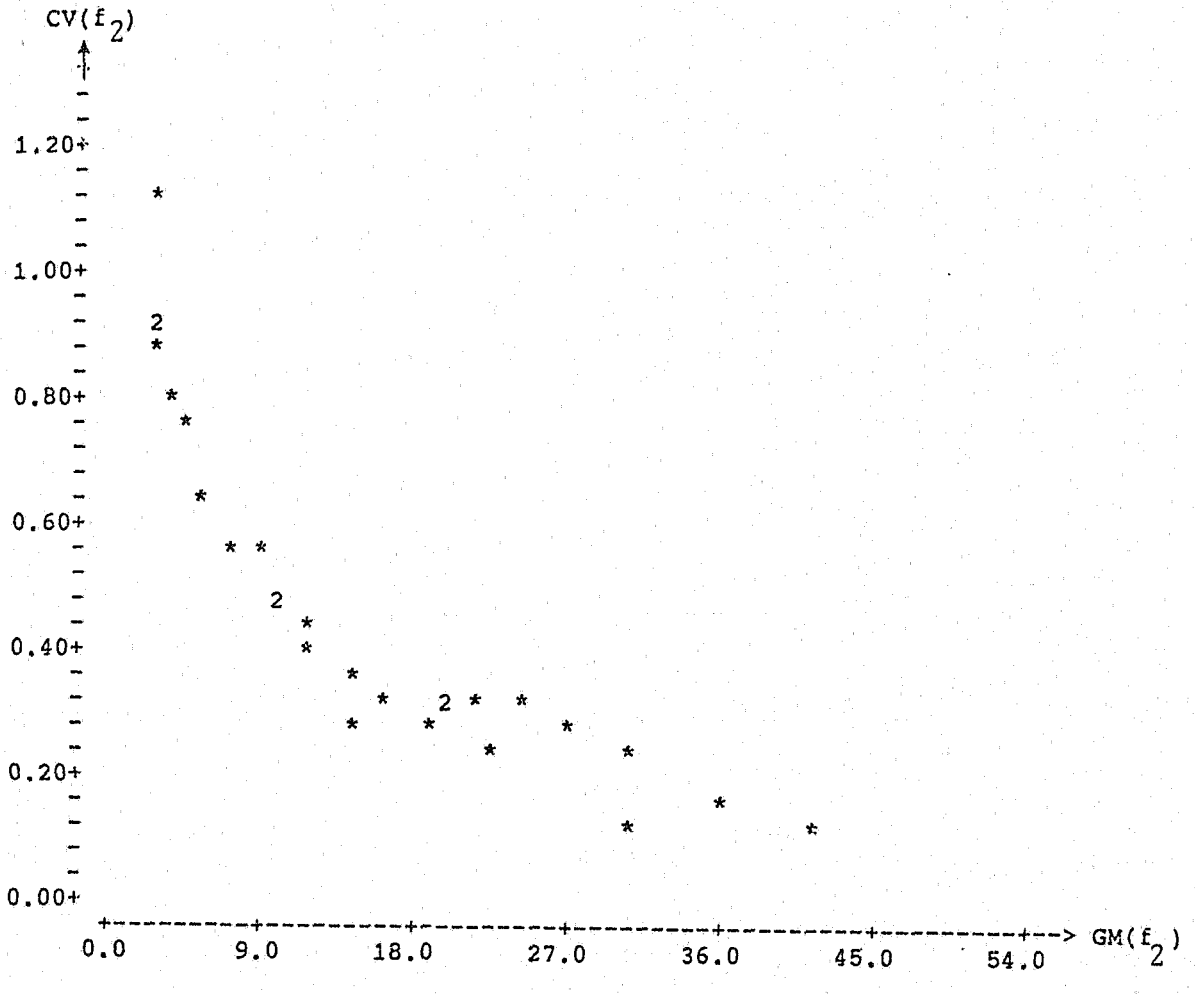
Summary of ANOVA for Low Priority Calls in Run 2

uncorrected estimator f_2	# of observ. in level f_2	mean waiting time for level f_2	standard deviation for level f_2
16.8	20	0.04738	0.05306
18.0	69	0.04145	0.03854
19.2	186	0.04583	0.04014
20.4	311	0.04715	0.04362
21.6	347	0.06196	0.05054
22.8	305	0.07862	0.06017
24.0	288	0.09677	0.06328
25.2	224	0.11898	0.06865
26.4	188	0.14367	0.08018
27.6	164	0.16023	0.07596
28.8	132	0.17029	0.07954
30.0	99	0.19562	0.07979
31.2	71	0.19637	0.08620
32.4	74	0.23878	0.08391
33.6	39	0.23965	0.06940
34.8	41	0.26779	0.08968
36.0	21	0.32133	0.08744
37.2	18	0.33685	0.10576
38.4	17	0.32817	0.10884

Figure 14

Plot of the Coefficient of Variation vs. the

Group Mean Waiting Time for Low Priority Calls in Run 2



Appendix

An Event-Paced Simulation Program Modelling Incoming Emergency Calls to a Police Dispatcher

This appendix gives a brief description of the PL1G program used to derive the results of this study. The simulations were run on the PRIME 850 at the Sloan School of Management at the Massachusetts Institute of Technology.

Description of Program

The program simulates calls for service arriving at a police department with Poisson rates. There are * t types of calls for service of various priorities, * s patrol units and * the service type i has an Erlang probability distribution with parameters (r_i, μ_i), i=1...t. The service discipline is non-preemptive with head-of-the-line priority structure.

The following is a list of the main variables used in the simulation program:

ct = # of call types

i = call type

m = # of servers (patrol units)

p = priority

Q(p) = priority-p queue

t1 = time horizon/end of simulation

t = time/clock

---

* In our simulation, we generate Erlang (r,λ) distributed random number from $e(r,\lambda) = \frac{1}{\lambda} \ln \left(\prod_{i=1}^{r} r_i\right)$, where r_i are uniformly distributed random numbers in the interval [0,1].

τ_i = type i interarrival time

τ_{ij} = service time of server j on call of type i, (independent of j)

tnc_i = time of next type i call

tss_j = time of service start of server j (if busy)

tsf_j = time of service finish of server j (if busy)

f_i^{Ω} = estimated unvorrected waiting time of a type i call given, Ω, the state of the system.

Figure A1 illustrates the flow of the simulation program

Figure A.1

Program Flowchart