OPERATIONS RESEARCH CENTER

working paper

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

THE DESIGN OF TIME-SERIES COMPARISONS UNDER RESOURCE CONSTRAINTS

by

Thomas R. Willemain

Nelson S. Hartunian

OR 109-81 November 1981

TABLE OF CONTENTS

Acknowledgements

1. Introduction: Advocate and Skeptic

2. Assumptions

3. The optimal fixed design

    3.1 Derivation of optimal fixed baseline

    3.2 Comparison with equal allocation design

    3.3 Use of fallible priors

4. The optimal flexible design

    4.1 Constructing the optimal switching rule

    4.2 Comparison with optimal fixed design and equal allocation design

    4.3 Use of fallible priors

5. Discussion

ACKNOWLEDGEMENTS

This work was supported by the National Institute of Justice, U.S. Department of Justice, under Grant Number 80-IJ-CX-0048 ("Adaptive Evaluation Designs") to the Operations Research Center, Massachusetts Institute of Technology. Professor Willemain's work was supported by a subcontract from the Operations Research Center, MIT to the John F. Kennedy School of Government, Harvard University. Computer time was donated by Smart, Hartunian and Willemain, Inc of Belmont, MA. Professor Richard C. Larson of MIT, serving as Principal Investigator, inspired, encouraged and improved this paper. Useful comments were received from Professors Richard C. Light and Herman B. Leonard of the John F. Kennedy School of Government, Harvard University. All statements in this paper are the responsibility of the authors and do not necessarily reflect the views of the sponsor.

1. Introduction: Advocate and Skeptic

Advocate is a police researcher with a new idea about reducing crimes committed in his city against residents of subsidized housing for the elderly. Skeptic is responsible for dispensing small grants for innovations in police services. Advocate, inspired by his idea, contacts Skeptic hoping to secure a small demonstration grant. Skeptic, mindful of her several sad experiences with "demonstrations" which established little, insists of some kind of controlled study of the innovation. Since there is only one such housing project in Advocate's city and the crime rate appears to have been fairly stable over time (though records are poorly kept), Advocate and Skeptic agree on an interrupted time-series study in which a period of careful baseline data collection will be followed by a trial of Advocate's ideas.

Writing a rough pre-proposal to Skeptic, Advocate wonders how long to let the baseline run. He feels he has a good idea of the current crime rate, despite the poor records, and wants to devote as much of his money as possible to the trial period. He proposes a brief baseline period. Reading the pre-proposal from Advocate, Skeptic senses an imbalance between the baseline and trial periods and suggests devoting equal allocations of funds to the two phases of the study. Since each day of the trial period will cost more than each baseline day, this suggestion requires that the baseline be longer than the trial.

Thinking about Skeptic's suggestion of equal allocations, Advocate wonders whether there might be an ideal baseline duration. After some analysis in which he assumes perfect knowledge of both the baseline rate and the experimental impact, Advocate discovers that there is an optimal baseline duration and that its use can improve the precision of estimates arising from the interrupted time-series comparison. In practice, Advocate realizes that he will not have perfect foreknowledge, but he is confident that his prior estimates of any unknown quantities will serve well enough.

Then Advocate pushes his idea even further. The purpose of the baseline period, he reasons, is to provide a reference crime rate estimate for later comparison. Why not make the duration of the baseline dependent on the baseline data as they come in? When the baseline has more or less confirmed his estimate of the existing crime rate, then he can switch over to the trial of his new program and make the best use of his limited funds. This will help minimize the "baseline tax" he must pay to try out his innovation. Advocate phones Skeptic to hear her reactions.

Skeptic is apalled. This sounds not only like bad science but also like a perfect opportunity to fudge the results in favor of the innovation. She points out that Advocate might (unconsciously, no doubt) begin with too high an estimate of the crime rate, wait opportunistically for just the right fluctuation in the number of crimes, then coast to a "successful" conclusion by exploiting regression to the mean masquerading as a real drop in the crime rate.

This paper describes a method for implementing Advocate's idea to Skeptic's satisfaction. Advocate has a Bayesian perspective; Skeptic is a frequentist. Skeptic will permit Advocate to use his Bayesian priors to manage the study by timing the switchover from baseline to trial. In return, however, Advocate will follow explicit rules for using his priors in combination with the baseline data as they appear, and will report his results in frequentist terms, using only the data obtained during the study. We will show that this methodology will provide better estimates of the change in the crime rate than would be obtained from a baseline period of predetermined duration, provided that Advocate's priors are sufficiently accurate.

2. Assumptions

We are interested in evaluating programs that change the rate of occurence of discrete events, such as serious crimes. We assume: (1) The underlying process we wish to change is a Poisson process in which events occur at a constant average rate λ per unit time. (2) The purpose of the experimental intervention is to change the rate of the Poisson process from λ to $\lambda + \delta$, so the purpose of the evaluation is to estimate the change in rate δ. (3) Evaluation resources are constrained. One unit of time of baseline operation costs one unit of resources, while one unit of time of experimental operation costs C units. The resource total available for the evaluation is R. For instance, R might be measured in days, in which case C=1 and the problem is to divide the R days between the baseline and experimental phases. Alternatively, R might be expressed in dollars per baseline day, with a day of experiment costing C times more than a day of baseline. (4) The evaluator has some (joint) prior estimates of λ and δ. These priors provide the basis for improving the evaluation design beyond that suggested by simple rules of thumb like "devote equal resources to both phases of the evaluation." However, while we will use the priors to design the evaluation, we will not use them directly in estimating the experimental effect δ (see assumption (5) below). Since the priors will inevitably be in error to some degree, we will test the sensitivity of our results to errors in the priors. (5) The evaluator's criterion for assessing the performance of an experimental design is the mean square error (MSE) in the resulting estimate of δ, where the mean is thought of as computed over many replications of the study. This is in keeping with the requirement in Section 1 that results be reported on a frequentist basis, even though Bayesian methods are used in conducting the study.

3. The optimal fixed design

We begin with the relatively straightforward problem of dividing limited evaluation resources between baseline and experimental phases of fixed duration. We call the prespecified division that minimizes the MSE in the estimate of experimental effect the "optimal fixed design." This is to be distinguished from fixed designs based on simple but suboptimal rules of thumb and from the flexible design of section 4, which does not pre-specify the duration of the baseline. We will first assume perfect prior knowledge, then relax this assumption in section 3.3.

3.1. Derivation of optimal fixed baseline.

Suppose the time series comparison consists of a baseline of duration T_b and an experimental phase of duration T_x. If a total of N_b events occur during the baseline and N_x events during the experiment, a natural estimate of the experimental impact δ is

$$\hat{\delta} = N_x/T_x - N_b/T_b \quad . \tag{1}$$

It is easy to show from the properties of the Poisson distribution that the mean and variance of the estimator are

$$E\ \{\hat{\delta}\} = \delta \tag{2}$$

$$V\ \{\hat{\delta}\} = (\lambda+\delta)/T_x + \lambda/T_b . \tag{3}$$

Thus $\hat{\delta}$ is an unbiased estimator of δ, and minimizing the MSE of the estimate (i.e., minimizing $E\ \{\hat{\delta}-\delta)^2\}$ is equivalent to minimizing its variance. This minimization will be constrained by the resources available for the evaluation. The problem can be written as

$$\underset{T_b,\ T_x}{\text{minimize}} \quad (\lambda+\delta)/T_x + \lambda/T_b \tag{4A}$$

$$\text{subject to} \quad T_b + CT_x \leq R \quad . \tag{4B}$$

The solution of this minimization problem is

$$T_b^* = R/[1 + (C(1+\delta/\lambda))^{1/2}] \tag{5}$$

$$T_x^* = (R/C)/[1+(C(1+\delta/\lambda))^{-1/2}] \quad . \tag{6}$$

Note that two factors control the relative durations of the two phases of the experiment: the cost of an experimental day relative to a baseline day C, and the proportional change in the Poisson rate δ/λ. When an experimental day is relatively expensive, it is obviously desirable to keep the experimental phase as brief as possible. It is also appropriate to limit the experimental period when the intervention reduces the Poisson rate, since the lower rate will have a smaller variance and require less observation for stable estimates.

3.2. Comparison with equal allocation design

We can contrast the optimal fixed design with a simpler approach that allocates half the resources to each phase of the evaluation. In this "equal allocation design"

$$T_b = R/2 \tag{7}$$

$$T_x = R/2C \quad . \tag{8}$$

Evaluating the MSE for both approaches using (3), we find for the optimal fixed design

$$MSE_{fix} = (\lambda/R)[1+(C(1+\delta/\lambda))^{1/2}]^2 \tag{9}$$

and for equal allocations

$$MSE_{equal} = (\lambda/R)2[1+C(1+\delta/\lambda)] \quad . \tag{10}$$

The proportional reduction in MSE afforded by the optimal fixed design is given by the ratio of (9) to (10). As the term $C(1+\delta/\lambda)$ approaches zero or infinity, this ratio approaches 0.5, so in theory the optimal fixed design could improve the MSE by 50% over the equal allocation design.

It appears, however, that the advantages of the optimal fixed design will be less dramatic in practice. If the scarce evaluation resource is dollars and - as might be expected - an experimental day is much more expensive than a baseline day, care in choosing the duration of the baseline will produce modest gains in performance. For instance, if the relative costs differ by an order of magnitude ($c=10$) and the experimental intervention changes the Poisson rate by 10 percent either way, the optimal fixed design will cut the MSE by about 20 percent. On the other hand, if every day costs the same or if the scarce resource is the total time available for data collection ($C=1$), then the optimal fixed design performs negligibly better even when the intervention changes the rate by 50 percent.

3.3 Use of fallible priors

To this point we have assumed that the evaluator has perfect knowledge of both the baseline rate and the change in that rate caused by the experimental intervention. In practice, the evaluator will have a joint prior distribution for these two parameters, which we denote by $f(\delta,\lambda)$. Logically, all pairs (δ,λ) are possible, provided

$$\lambda > 0 \tag{11A}$$

$$\delta > -\lambda \ . \tag{11B}$$

Thus the design problem changes from that given in (4) to

$$\underset{T_b,\ T_x}{\text{minimize}} \quad \int_{\lambda=0}^{\infty} \int_{\delta=-\lambda}^{\infty} [(\lambda+\delta)/T_x + \lambda/T_b]\ f(\delta,\lambda)\ d\delta d\lambda \tag{12A}$$

$$\text{subject to} \quad T_b + CT_x \leq R \ . \tag{12B}$$

Now the joint prior can be rewritten as the product of a conditional and an unconditional prior

$$f(\delta,\lambda) = g(\delta|\lambda)\ h(\lambda) \ . \tag{13}$$

If the prior assumes that the innovation has a constant proportional impact on λ, i.e.

$$E[\delta|\lambda] = F\lambda \tag{14}$$

then the solution to (12) is the same as the solution to (4) with F, the prior estimate of the proportional impact, replacing the unknown true value δ/λ. Expression (14) is probably a reasonable assumption to make considering the usual state of foreknowledge of experimental impact; in any case it is probably a good approximation for marginal improvements.

The result of (14) is that the evaluator will choose as the baseline duration

$$T_b = R/[1+(C(1+F))^{1/2}] \ . \tag{15}$$

Note that the problem of selecting a baseline of fixed duration requires only a single summary measure of the evaluator's expectation about proportionate impact. Every evaluator will have some rough sense of the value of F.

Nevertheless, if the chosen value of F differs too much from the true value δ/λ, using the prior will result in baseline durations that yield estimates inferior to those of the equal allocation design. How accurate must the choice of F be to guarantee improved performance?

The MSE associated with the prior F is given by (9) with F replacing δ/λ

$$MSE_{prior} = (\lambda/R)[1+(C(1+F))^{1/2}]^2 \ . \tag{16}$$

Using the prior will improve on the equal allocation design provided (16) is less than (10). This condition can be translated algebraically into the following condition on the prior F

$$\min\left\{\begin{matrix} C^{-1}-1 \\ C(1+\delta/\lambda)^2-1 \end{matrix}\right\} < F < \max\left\{\begin{matrix} C^{-1}-1 \\ C(1+\delta/\lambda)^2-1 \end{matrix}\right\} \ . \tag{17}$$

Figure 1 shows the range of acceptable priors when a day of experiment costs twice as much as a day of baseline (C=2). For many values of δ/λ, the range of acceptable values of F is rather wide. For instance, if the intervention will reduce the Poisson rate by 20% (δ/λ = -.2), then any value of F in the range -.50 to +.28 will suffice. The most severe restriction on the value of F occurs when the equal allocation design happens to coincide with the optimal fixed design; this occurs at $\delta/\lambda = C^{-1}-1$ and naturally requires a perfectly accurate prior (see Figure 1 where the true value δ/λ is -0.5).

4. The optimal flexible design

Our problem in the fixed design was to predetermine when to switch the evaluation from baseline to experiment. We outlined a method for making this decision and compared it to the simpler rule of equal allocation of resources to the two phases. Our problem in the flexible design is more complicated, since in this case the decision to terminate the baseline is not made once and for all before the evaluation begins. Rather, we imagine a series of equally-spaced decision points, which we will take to be days, and we must develop a way of making sequential decisions throughout the baseline. The operative question will always be: "Given the number of days that the baseline has continued and the number of events that have occured, should we extend the baseline one more day, or should we switch now to the experimental phase?"

We expect this more complicated design to provide better estimates of experimental impact in return for the greater complexity. An evaluator whose priors are accurate can selectively capitalize on chance, terminating the baseline early if it quickly confirms the prior.

As before, we will begin assuming perfect knowledge of both λ and δ, then assess the consequences of using fallible priors.

We assume that the evaluator will use as much of the available resources as possible. This precludes considering early termination of the experimental phase and exploiting another degree of freedom in the evaluation design. However, by assuming a fixed end point, we can use the method of backward recursion to construct the optimal rules for ending the baseline and to assess the MSE of the optimal flexible design.

4.1 Constructing the optimal switching rule

A "switching rule" tells the evaluator whether to switch after observing a given count of events over a given number of baseline days. The optimal switching rule minimizes the MSE of estimating the experimental impact. The best rule can be determined by a backwards recursion, which constructs the optimal sequence of decisions subject to a resource constraint, determining the optimal decision at one time from the optimal decision one time later.

Consider the evaluator's problem at the end of any baseline day. If the switch is made at that point, the MSE of estimate depends on the unknown number of events to be observed over the forthcoming experimental phase. On the other hand, if the evaluator extends the baseline by one day, the MSE will be minimized by subsequently making the best possible decisions, given the unknown number of events that will occur during the next baseline day. By averaging over the possible number of those events, the evaluator can compute the expected MSE arising from the decision to extend the baseline. Then the final choice for that day depends on comparing the expected MSE arising from extending the baseline against that arising from switching to the experimental phase.

Let $V(T_b, N_b)$ represent the lowest possible value for the MSE given that T_b days of baseline have produced N_b events (the MSE also depends on λ, δ, R and C, but these parameters are suppressed for simplicity). If the best choice conditioned on T_b and N_b is to extend the baseline one more day, during which some number m of events will occur according to a Poisson process with rate λ, then the value of $V(T_b, N_b)$ can be expressed in terms of the values of $V(T_b+1,*)$ as

$$V(T_b,N_b)=\sum_{m=0}^{\infty} V(T_b+1,N_b+m)\lambda^m \exp(-\lambda)/m! \quad . \qquad (18)$$

On the other hand, the best ~~choice~~ conditional on T_b and N_b may be to switch to the experimental phase. In this case the experimental phase would have duration

$$T_x = \lfloor (R-T_b)/C \rfloor \tag{19}$$

where $\lfloor x \rfloor$ signifies the largest integer less than or equal to x. If n_x events occur over the T_x days of the experimental phase, then the estimate of experimental impact will be

$$\hat{\delta} = n_x/T_x - N_b/T_b \quad . \tag{20}$$

The corresponding MSE will be

$$V(T_b, N_b) = E[\{\hat{\delta} - \delta\}^2] \tag{21}$$

$$= (\lambda+\delta)/T_x + (\lambda - N_b/T_b)^2 \tag{22}$$

To choose between extending and terminating the baseline we must determine which option has the smaller MSE

$$V(T_b, N_b) = \min \begin{cases} \sum_{m=0}^{\infty} V(T_b+1, N_b+m)\ \lambda^m \exp(-\lambda)/m! \\ (\lambda+\delta)/T_x + (\lambda - N_b/T_b)^2 \ . \end{cases} \tag{23}$$

Expression (23) provides the recursion relation.

The recursion can be started as follows. The longest possible baseline duration leaves only enough resources for a single experimental day (i.e., $T_x=1$, $T_b= \lfloor R-C \rfloor$). After this day there is no choice but to switch to the experimental phase. Hence the first values of the recursion are given by (22) specialized for this case

$$V(\lfloor R-C \rfloor, N_b) = (\lambda+\delta) + (\lambda-N_b/\lfloor R-C \rfloor)^2, \ 0 \le N_b < \infty \quad . \tag{24}$$

By starting with (24) and working backwards through the values of T_b using (23), we can determine the optimal switching rule for every pair of baseline conditions (T_b, N_b). Finally, since the evaluation begins at $(T_b=0, N_b=0)$ and must include at least one baseline day, we can determine the overall MSE of the optimal flexible design from

$$MSE_{flex} = V(0,0) = \sum_{m=o}^{\infty} V(1,m)\lambda^{m} \exp(-\lambda)/m! \quad . \tag{25}$$

Figure 2 illustrates the optimal switching rules for one particular set of parameters (R = 20, C = 2, λ = 1, δ = -0.1). Roughly speaking, the switching zone after T_b baseline days centers on the expected number of counts λT_b. However, several other features of the optimal switching rule are apparent in the figure and worth mentioning. First, the baseline can never exceed 18 days' duration. This is because there must be at least one day of experimental operation, and in this example one experimental day requires 2 of the 20 resource units available. Second, the switching zone is funnel shaped, flaring wide after beginning narrow. Intuitively, the criteria for terminating the baseline very early must be quite strict, and hence the switching zone be quite narrow, because one would ordinarily desire a baseline long enough to provide a stable estimate of λ. Only exact coincidence between expected and observed counts justifies terminating the baseline after but a few days. However, as the baseline continues and resources available for the experimental phase grow scarce, the criteria for terminating the baseline become less strict. Eventually there are so few resources remaining that the evaluator must be satisfied with even the most tenuous relationshop between the expected and observed number of baseline events. Finally, the switching rules are more stringent for odd numbered baseline days, as seen in the relatively narrow switching zones for days 9, 11, 13 and 15. This difference between odd- and even-numbered days results from the combination of two assumptions: that the baseline be an integer number of days and that an experimental day be twice as costly as a baseline day. Consider a baseline that has already extended 14 days. At this point there are $R-T_b=20-14=6$ resource units remaining, enough to support 3 experimental days. If the baseline is extended from 14 to 15 days, the 5 remaining resource units will support 2 experimental days but leave one resource unit unused. In contrast, extending the baseline from 14 to 16 days will convert the one unused resource unit into an additional baseline day while still supporting 2 experimental days. Hence it is more difficult to justify terminating the baseline after 15 than after 16 days, and the switching zone for 15 days is relatively narrow.

4.2 Comparison with optimal fixed design and equal allocation design

The optimal flexible design can achieve dramatic reductions in MSE compared to the equal allocation design. The optimal flexible design shown in Figure 2 produces much better estimates than do the equal allocation or optimal fixed designs, for which the MSEs are 0.280 and 0.274, respectively. The MSE for the optimal flexible design is 0.153, about 45% below that produced by the others. To get a more general assessment, we analyzed 84 cases chosen to span a wide range of values of parameters R, C, λ and δ; we found that the optimal flexible design reduced MSE by anywhere from 20% to 70%, with a median of roughly 30%. Figure 3A shows the histogram of the savings achieved by the optimal flexible design in the 84 cases analyzed.

The optimal flexible design also performs better than the optimal fixed design, although the difference is naturally not as great in this case; Figure 3B gives a histogram of the savings achieved by the optimal flexible design relative to the optimal fixed design. Comparisons not shown here reveal that, in general, the additional gain achieved by moving from the fixed to the flexible design is greatest when the fixed design has the least advantage over the equal allocation design.

It is safe to conclude that the theoretical potential of the flexible design is great: to reduce MSEs by one-third or more for the same total resource commitment is impressive. It remains to be seen, however, whether these theoretically achievable gains can be realized in practice. The next section considers the effects of the fallible priors on the gains associated with the flexible design.

4.3 Use of fallible priors

Again we consider whether the gains available in theory are vitiated by fallible priors. Two changes are required to evaluate the performance of a flexible design managed without perfect knowledge. First, we must integrate (23) over the joint prior for δ and λ to determine the switching rules perceived to be optimal by the evaluator. Note that this prior must now be conditioned by the experience observed over the baseline. Second, these switching rules must be evaluated using the true values of λ and δ in (23) -- but choosing not necessarily the smaller of the right hand sides, but rather the one corresponding to the switching rule determined with the fallible prior. If the joint prior distribution is inaccurate and/or imprecise, there may be a large difference between the MSE of the chosen decision and the MSE of the better decision.

The first task is to integrate (23) over the joint prior for δ and λ. We again write the joint priors as a product, this time conditioning on both the baseline data and the evaluator's prior.

As is customary when working with Poisson processes, we will take the prior for λ to be gamma distributed. We will denote the parameters of this distribution by T_o and N_o. It is well known that after observing N_b events over T_b baseline days, the updated prior will be gamma with parameters $T_o + T_b$ and $N_o + N_b$.

Formally, we can modify (13) to read

$$f(\delta,\lambda|N_o,T_o,N_b,T_b) = g(\delta|\lambda, N_o,T_o,N_b,T_b)\cdot h(\lambda|N_o,T_o,N_b,T_b) \qquad (26A)$$

where

$$h(\lambda|N_o,N_b,T_o,T_b) = \frac{(T_o+ T_b)^{N_o + N_b}}{(N_o+ N_b-1)!} \lambda^{N_o+N_b-1} \exp[-\lambda(T_o+ T_b)] \quad . \qquad (26B)$$

Note that the updated prior $h(\lambda|N_o,T_o,N_b,T_b)$ has mean

$$E[\lambda|N_o,N_b,T_o,T_b] = (N_o+ N_b)/(T_o+ T_b) \qquad (28)$$

and variance

$$V[\lambda|N_o,N_b,T_o,T_b] = (N_o+N_b)/(T_o+T_b)^2 \quad . \tag{29}$$

We again take the conditional distribution of δ, $g(\delta|\lambda, N_o,T_o,N_b,T_b)$, to have mean proportional to λ, so that

$$E[\delta|\lambda, N_o,N_b,T_o,T_b] = F(N_o+N_b)/(T_o+T_b) \quad . \tag{30}$$

Now the MSE associated with the decision to terminate the baseline after N_b events in T_b days is found by integrating (22) over (26), giving

$$MSE_{switch} = \frac{N_o+N_b}{T_o+T_b}\left[\frac{1}{T_o+T_b} + \frac{1+F}{T_x}\right] + \left[\frac{N_o+N_b}{T_o+T_b} - \frac{N_b}{T_b}\right]^2 \quad . \tag{31}$$

Finally, the MSE associated with extending the baseline by one day is found by integrating (18) over (26). This results in changing the Poisson distribution to a negative binomial distribution

$$MSE_{extend} = \sum_{m=o}^{\infty} V(T_b+1,N_b+m)\,\frac{(m+N_o+N_b-1)!}{m!(N_o+N_b-1)!}\left[\frac{1}{1+T_o+T_b}\right]^m\left[\frac{T_o+T_b}{1+T_o+T_b}\right]^{N_o+N_b} \cdot \tag{32}$$

The evaluator will terminate the baseline if (31) is less than (32); otherwise, he will extend the baseline one more day. Making this comparison for all possible pairs (T_b,N_b) constitutes determining the switching rules by which the interrupted time-series study is managed.

Once the switching rules have been determined, we can assess the MSE associated with the flexible design by using these switching rules in place of the optimal switching rules in the procedure of Section 4.2.

To see how well fallible priors perform in the flexible design, we evaluated several, some "fuzzy" and some "firm". The fuzzy priors assumed that the evaluator stated his prior for λ in terms equivalent to only one day of baseline operation (T_o=1), whereas the firm priors assumed the equivalent of ten days' experience

(T_0=10). In all cases, the expected values of the prior estimates for λ and δ/λ were made to range widely around the true values in order to expose the link between accuracy of prior and performance of design. (Recall that only the mean value, not the variance, need be quoted for the prior estimate of experimental impact, so the notion of fuzziness applies only to the prior for the baseline rate.)

The results of these analyses are shown in Figures 4 and 5. In both figures, the correct values of baseline rate and proportional change are indicated by the "x" in the center. Results are given in terms of the ratio of MSE in the flexible design to MSE in the equal allocation design. In Figure 4, the priors for the baseline rate are fuzzy. The consequences of using these relatively non-commital priors are two. First, there is a fairly large tolerance for error in the prior, as reflected in the wide range of errors over which the MSE is still lower with the flexible design than with the equal allocation design. Second, this tolerance comes at the cost of very minor improvements in performance, with no more than a 10% improvement recorded over the range examined and typical improvements being only 3 or 4%. By contrast, the priors in Figure 5 are firm, and the results are opposite those obtained with a fuzzy prior. With the firm priors, only quite accurate estimates of the baseline rate will lead to improved performance. However, the improvements are larger, usually more than 10% and reaching as much as 26%. Unfortunately, it is also true that when the error is large with a firm prior, the degradation in performance is greater than it is with a fuzzy prior.

These results illustrate a fundamental tradeoff in the use of prior information to manage an interrupted time series. If the evaluator relies strongly on prior information, any resulting gains or losses will be large. Relying weakly on the priors minimizes the possibilities of either large losses or large gains. Thus the tradeoff is between risk and return.

Finally, we must remark on a curious aspect of the results shown in Figures 4 and 5: better performance can be had with less accurate priors. In all cases, a better estimate of the baseline rate pays dividends, but in many cases performance is improved with a less accurate prior for the experimental impact (F). For instance, in Figure 4, the relative performance of the flexible design is best in the semi-circular band visible in the lower half of the figure. away from the correct values in the middle.

The explanation of this apparently perverse phenomenon lies in the ability of a deficiency in one prior to compensate for a deficiency in the other. Recall expression (31) for the MSE associated with switching to the experimental phase. Note that this MSE is linearly related to F, the prior estimate of (fractional) experimental impact defined in (30). Therefore, choosing a lower F makes it more likely that the baseline will be terminated early. Now an accurate but fuzzy prior will lead to overly conservative management of the baseline, missing opportunities to terminate early. It follows that an underestimate of F will provide some useful compensation for using an accurate but uncertain prior estimate of the baseline rate λ. Conversely, if the prior for λ is firm but inaccurate, an overestimate of the experimental impact will tend to prevent inappropriate early termination of the baseline. In practice, of course, it will not be possible to exploit this type of compensation because the evaluator will be using his best estimates and will not know his errors (if he did, he would adjust his priors accordingly).

5. Discussion

To our knowledge, this is the first treatment of the problem of designing an interrupted time series subject to resource contraints. The body of literature on sequential estimation is similar in spirit but inapplicable because our problem involves events that occur after the decision to stop the baseline has been implemented. The paper that is most similar is the treatment by El-Sayyad and Freeman* of the problem of estimating a single Poisson rate subject to sampling costs. Their use of Bayesian methods and dynamic programming inspired much of our work. However, they did not treat the problem of estimating a change in a Poisson rate over both a baseline and experimental phase, nor did they consider a constraint on resources.

We have shown that it is possible to make notably better estimates of the change in Poisson rate by reacting to baseline data as they appear. Reductions in MSE of 30 percent were common in the cases we examined. We have also given an algorithm for computing the rules by which an evaluator should react to the baseline data in making the decision about switching from baseline to experiment.

Throughout we have used a productive but perhaps puzzling combination of Bayesian and frequentist perspectives. We have permitted the evaluator to use his priors for the baseline rate and for the experimental impact in deciding when to terminate the baseline. Nevertheless, we have evaluated the resulting estimates in frequentist terms, free of any direct influence of the evaluator's priors. We feel this separation is important for maintaining the credibility of flexible designs, which are by definition quite susceptible to the preconceptions of the evaluator.

---

*G.M. El-Sayyad and P.M. Freeman, "Bayesian sequential estimation of a Poisson process rate", Biometrika (1973) 60: 289-296.

One of our major concerns has been the effect of misconceptions on the evaluator's part. We have shown that use of accurate and strongly held priors leads to better estimates than would be achieved in an equal allocation design. Inaccurate but strongly held priors lead to worse estimates, and weakly held priors lead to lesser differences. In practice, an evaluator will not know how accurate his priors are, but our results should be helpful for probing the risks involved in using firm priors.

This work should be seen as an early step in an evolution toward evaluation methods well-adapted to the kinds of changing environments that can ruin studies that require constant conditions. Our method can forestall premature decisions about the duration of the baseline phase in an interrupted time series study, but it cannot cope with structural changes in the nature of the random process being studied or in the experimental intervention that seeks to change its rate. The next steps will be more difficult, but this first step may be an indication of how much we stand to gain by evolving a technology of flexible evaluation.

ACCEPTABLE VALUES OF PRIOR ESTIMATE OF EXPERIMENTAL IMPACT (C=2)

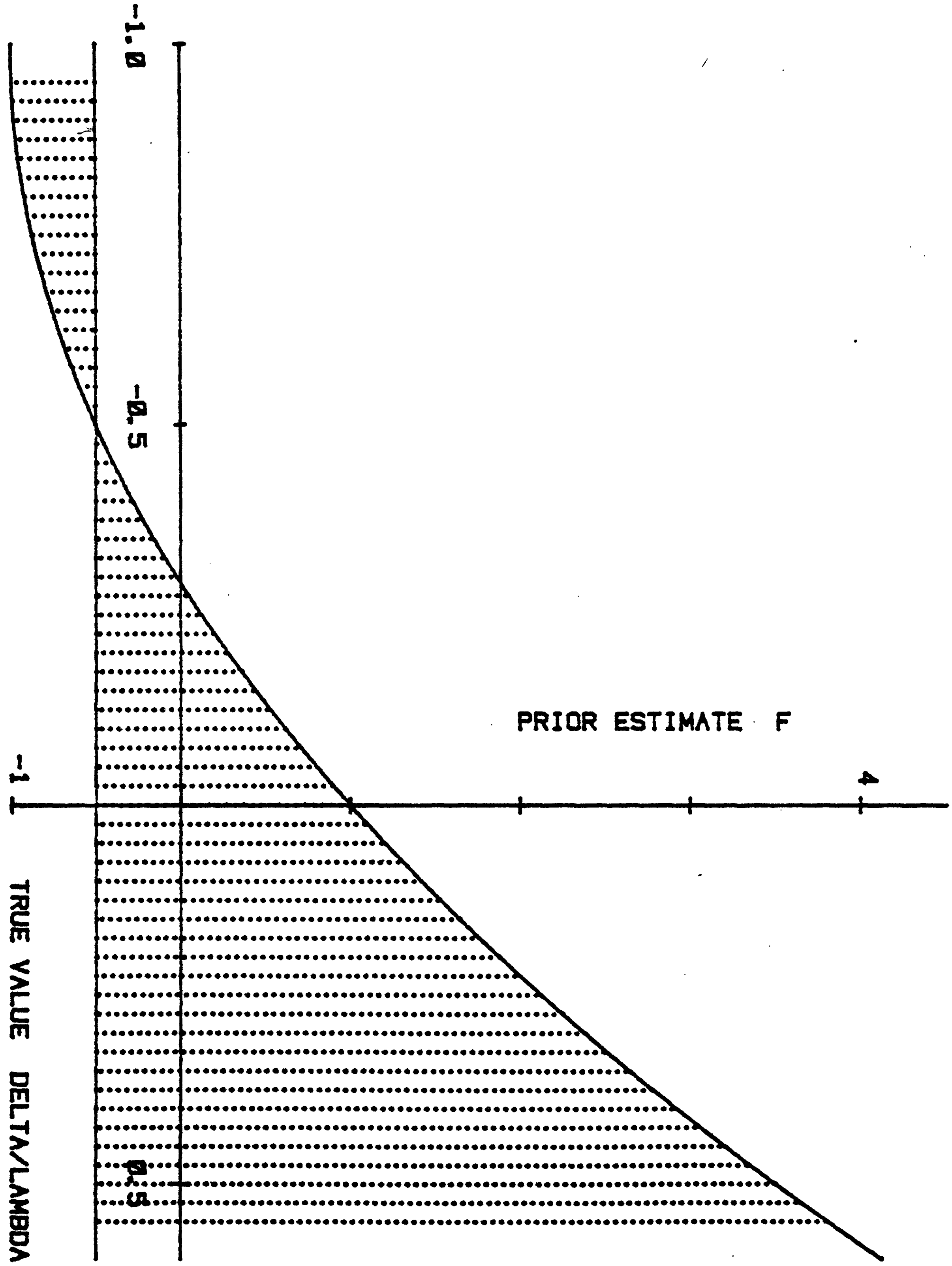

Figure 1

SWITCHING ZONE FOR OPTIMAL FLEXIBLE DESIGN

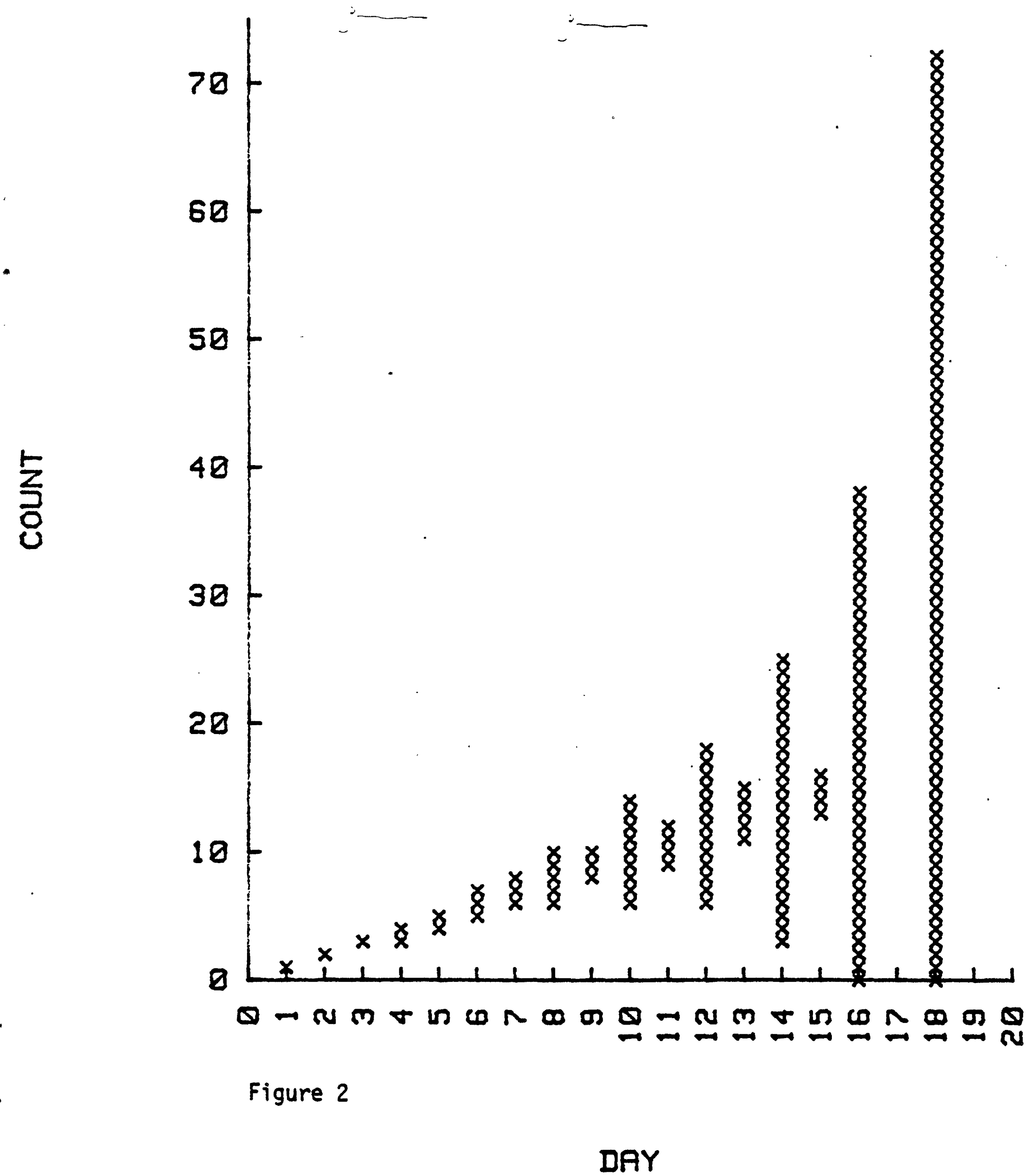

Figure 2

BASELINE RATE = 1 event per day

CHANGE IN RATE = -0.1 event per day

RESOURCE TOTAL = 20 units

COST PER EXPERIMENTAL DAY = 2 units

ADVANTAGE OF OPTIMAL FLEXIBLE DESIGN OVER THE EQUAL ALLOCATION AND FIXED DESIGNS

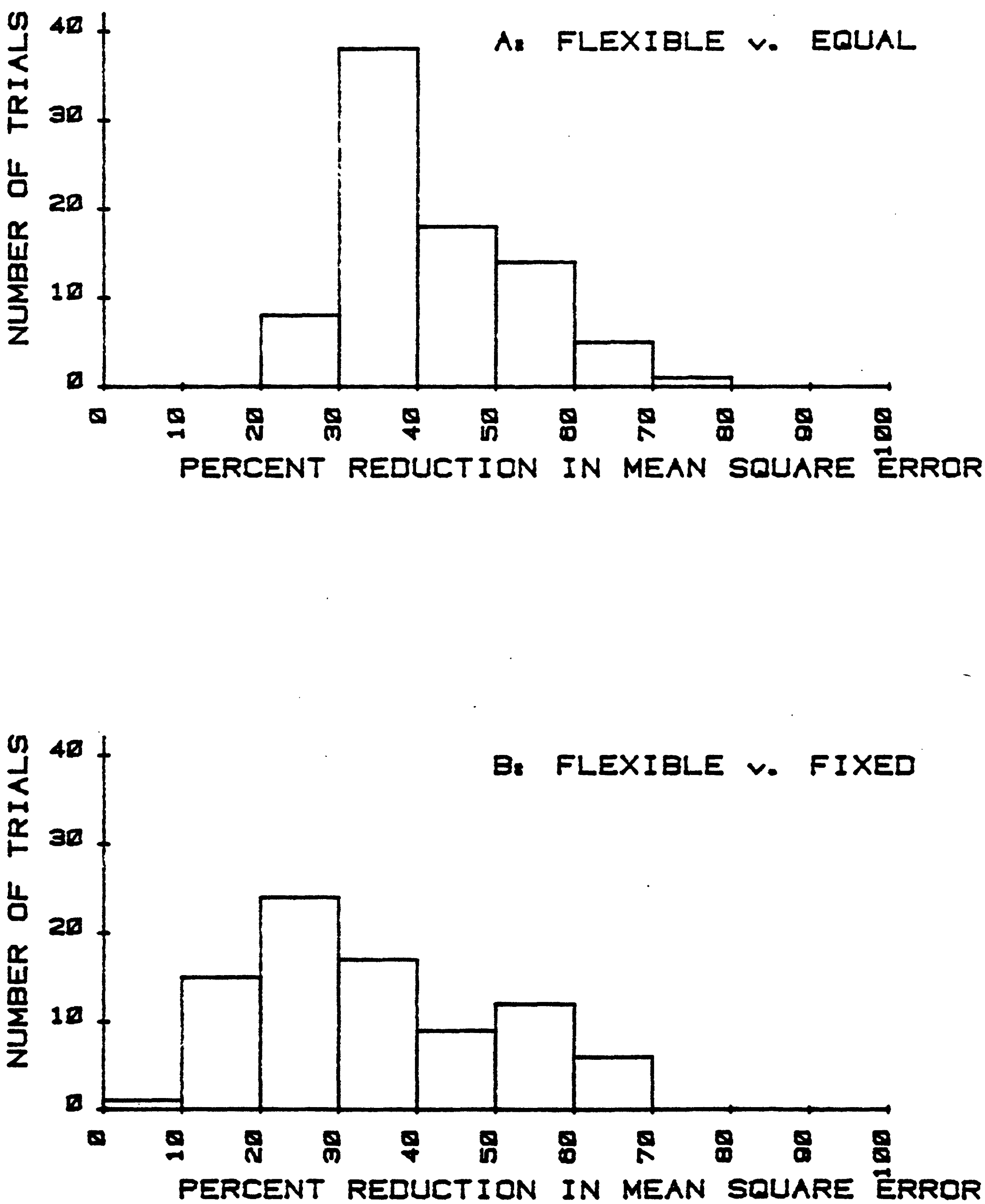

Figure 3

FUZZY PRIOR

KEY : RATIO MSE FLEX / MSE EQUAL

*	>1.20	.	0.95-1.00
+	1.10-1.20	−	0.90-0.95
−	1.05-1.10	+	0.80-0.90
.	1.00-1.05	*	0.75-0.80
		#	<.75

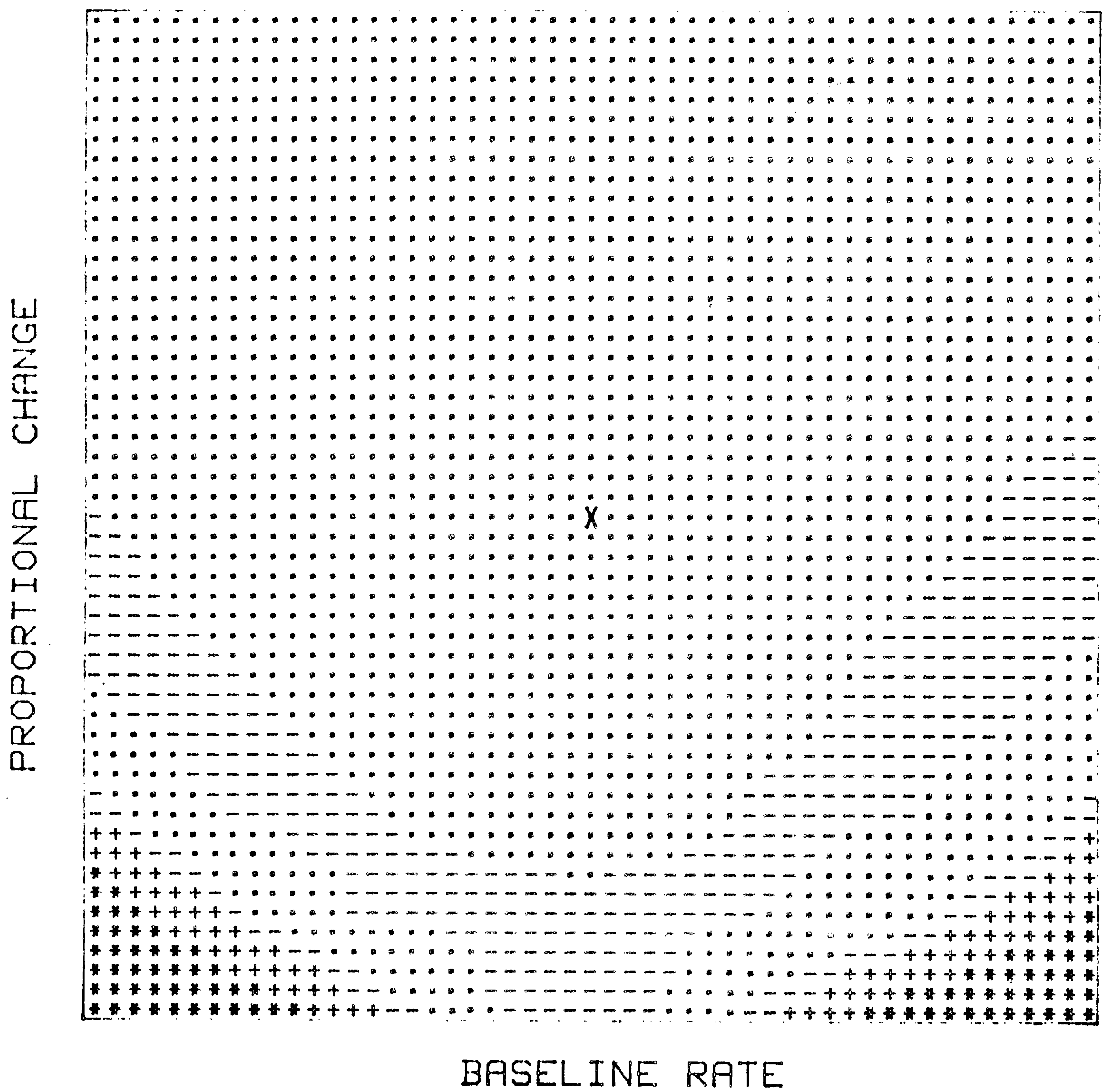

Figure 4

FIRM PRIOR

KEY : RATIO MSE FLEX / MSE EQUAL

*	>1.20	.	0.95-1.00
+	1.10-1.20	-	0.90-0.95
-	1.05-1.10	+	0.80-0.90
.	1.00-1.05	*	0.75-0.80
		#	<.75

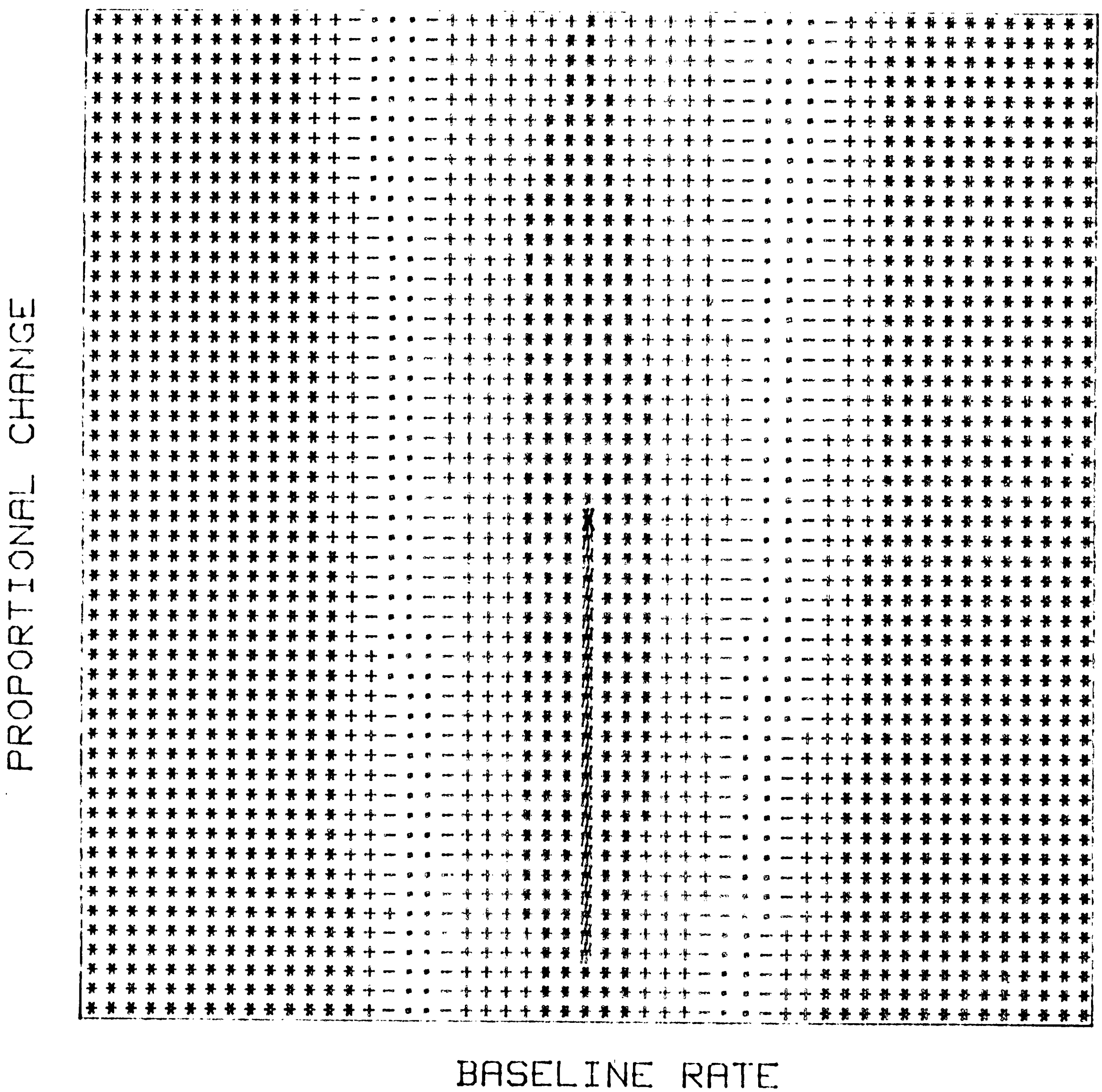

Figure 5