faculteit der economische wetenschappen

RESEARCH MEMORANDUM

[redacted]

TILBURG UNIVERSITY
DEPARTMENT OF ECONOMICS
Postbus 90135 - 5000 LE Tilburg
Netherlands

SMALL-JOBS-FIRST:

A combined analysis queuing, simulation, and regression

F. Keyzer, J. Kleijnen,
E. Mullenders, A. van Reeken.

Katholieke Hogeschool
Tilburg, Netherlands.

February 1979

CONTENTS

SMALL-JOBS-FIRST:

A combined queuing, simulation, and regression analysis[1)]

ABSTRACT

Key-punching in our university computing center is modeled as a queuing simulation with two servers (typists) and three priority classes: small, medium, and large jobs. The 90% quantiles of the waiting times per job class are estimated for different borderlines X and Y between the three job classes. An overall criterion function is formulated, quantifying tradeoffs among the waiting times of each job class. Several regression models are investigated, expressing the quantiles as functions in the class limits X and Y. Optimal class limits are determined applying a numerical search algorithm. The resulting optimal limits have been implemented.

1. INTRODUCTION

We investigate the following practical problem. The computing center of our university provides a key-punching service to its users. There are two typists (servers) available. It is known from queuing theory, that expected waiting time is minimized if the jobs (customers) are served in the order of their service times; see Cobham (1954), Jaiswal (1968). The computing center's management decided that a complete ranking of jobs would be impractical. To provide a "reasonable" turnaround time the following priorities were established. Small jobs (S jobs) have priority over medium (M) jobs, which in turn are key-punched before large (L) jobs. Within each of the three priority classes a first-come-first-served rule holds. The priority rules are not preemptive. Classification of jobs is possible indeed, since the key-punching time for a job can be predicted accurately enough from the code sheets to be key-punched. Originally, intuition was used by management to chose the borderline X between "small" and "medium" jobs, and the limit Y between "medium" and "large" jobs. The purpose of our study is to derive "optimum" X and Y

values. We emphasize that similar priority rules can be used in many other practical queuing situations: computer operating systems, automobile repair shops, etc.

From the above problem formulation it follows that we formulate the system as a queuing system with two servers and three priority classes, priorities being based on the lengths of the service times. This problem has not yet been solved analytically, even though in our case the arrival and service times are exponentially distributed (Poisson processes with parameter $\lambda = 0.033$ for arrivals and $\mu = 0.021$ for service times; the service times are minimally 10 minutes since smaller jobs are key-punched by the users themselves). During our investigation it soon turned out that management is not so much interested in the average queuing time, as in helping "as many people as fast as possible". It was agreed that we should therefore study the 90% quantile, i.e. if x denotes queuing time, then the 90% quantile Q is such that

$$P(x \leq Q) = 0.90 \qquad (1.1)$$

So there is only a 10% chance that customers have to wait longer than Q. Note that x is defined as waiting time excluding key-punching itself. We use discrete-event simulation to estimate the quantile Q for various X and Y values[2)].

2. PREVIOUS RESEARCH

In a first report, namely Coppus et al. (1976), 19 combinations of the X and Y limits were simulated, each combination yielding one observation for the 90% quantiles of the S, M, and L jobs: Q^S, Q^M, Q^L respectively. For each of these three quantiles a quadratic function was fitted to the 19 observations, e.g.,

$$Q_i^S = \alpha_0 + \alpha_1 X_i + \alpha_2 Y_i + \alpha_{11} X_i^2 + \alpha_{12} X_i Y_i + \alpha_{22} Y_i^2 + e_i \qquad (2.1)$$
$$(i = 1,\ldots,19)$$

The statistical tests applied to the resulting regression equations showed that X and Y did affect the quantiles, but the quadratic model did not correctly specify these effects. In Van den Bogaard and Kleijnen (1977) the following alternative regression models were investigated:

(1) Replace Y by (Y-X) in eq. (2.1). Unfortunately, the individual regression effects remained insignificant.

(2) Replace X and Y by p_1 and p_2, the probability of a job falling in class 1 (small jobs) and 2 (medium jobs) respectively. Again the regression effects were insignificant.

(3) Next some models were tried with the ratio X/Y as the explanatory variable. The results remained unsatisfactory.

(4) Finally, some models linear in X and Y were investigated. For instance,

$$\hat{Q}^s = 49.44 + 0.63X - 0.03Y$$
$$(t : 10.79 \quad 9,73 \quad - 1.48) \qquad R^2 = 0.86 \qquad (2.2)$$

where the numbers in brackets denote the t-statistics of the corresponding regression parameters. This model suggested that Q^s is sensitive to X, but not to Y. Indeed the model with a single explanatory variable yielded

$$\hat{Q}^s = 43.66 + 0.63X$$
$$(t : 17.70 \quad 9.36) \qquad R^2 = 0.84 \qquad (2.3)$$

We shall come back to this result in the present paper[3].

The above research demonstrated that each of the three job classes has conflicting optimal X and Y values (see also note 3). With hindsight it is obvious that the smallest X value is optimal, considering only small jobs: if X decreases then (independently of Y) there are fewer competing small jobs, so that queuing times decrease in this job class; see also eq. (2.3). Obviously the minimization of waiting times in a system with separate job classes, requires <u>tradeoffs</u> among the waiting times per class. Therefore we should not minimize waiting times per class independently. Together with the computing center's management we decided

to reformulate our problems as follows:

$$\text{Minimize } z = W^S (Q^S/B^S) + W^M (Q^M/B^M) + W^L (Q^L/B^L) \qquad (2.4)$$

where the weights W denote the relative number (fraction) of jobs per class, and the waiting time quantile is expressed relative to the mean service time per class. Eq. (2.4) formalizes a "nice" optimization problem, for instance, a decrease of X leads to the following results: lower W^S, lower quantile Q^S, lower average service time B^S, higher W^M, etc. Note that simulation is needed to estimate the quantiles Q, not the weights W and conditional mean service times B. For, if s denotes the individual service times then

$$W^S = P(s < X) = \int_a^X \mu\, e^{-\mu s}\, as = e^{-\mu a} - e^{-\mu X} \qquad (2.5)$$

and

$$B^S = E(s \mid s < X) = \{(a+\frac{1}{\mu})e^{-\mu a} - (X+\frac{1}{\mu})e^{-\mu X}\}\{e^{-\mu a} - e^{-\mu X}\}^{-1} \qquad (2.6)$$

while similar results apply for medium and large jobs, i.e. the weights W and conditional mean service times B can be expressed as explicit functions of X and Y.[4)]

3. THEORETICAL CONSIDERATIONS

The previous two reports Coppus et al. (1976) and Van den Bogaard & Kleijnen (1977) - and our experience in general - strongly suggests that regression analysis without a theoretical basis leads to very questionable results. Therefore we made a qualitative theoretical analysis before "grinding our simulation results through the regression mill". Note that Coppus et al. (1976) had only 19 observations available; Van den Bogaard & Kleijnen (1977) augmented this to 71 observations. In the present report we shall use these observations and later on we shall increase the data to 91 observations.

The regression model is a metamodel, i.e., it explains how the outputs (Q) of the simulation model react to changes in the simulation in-

puts (X,Y). The metamodel should explain the complicated simulation model as parsimonously as possible. The selection of explanatory variables and the functional form of the metamodel should be based on theory and common sense. See Kleijnen (1979).

Assumption 1: Q^S depends on X and not on Y

The waiting times of small (S) jobs are affected by the borderline X between S and M (medium) jobs, but not by the limit Y between M and L (large) jobs. For S jobs have priority over all other jobs, and are not influenced by the priority rules among these remaining jobs. This assumption neglects the non-preemptive character of our priority rules. Assumption 1 is corroborated by various regression results such as eqs. (2.2) and (2.3).[5)]

Assumption 2: Q^L depends on Y and not on X

Large jobs have to wait until all small and medium jobs have been served; their waiting times are not influenced by the subdivision into S and M jobs. This assumption is again corroborated by various regression models (not shown here).

Assumption 3: Q^M takes over the role of Q^S when X approaches zero

For medium (M) jobs waiting times depend on both X and Y. However, as X approaches zero (or more precisely X approaches the lower limit for service times a = 10 minutes), no S jobs remain and M jobs acquire the highest priority. So if $X \downarrow a$ then M jobs depend on their upper-limit Y just like S jobs depended on their upper-limit X. In general we have

$$Q^S = f^S(X) \qquad (0 \leq X) \tag{3.1}$$

$$Q^M = f^M(X,Y) \qquad (0 \leq X \leq Y) \tag{3.2}$$

A special case is X = a so that eq. (3.2) becomes

$$Q^M = f^M(a,Y) = g^M(Y) \qquad (0 \leq Y) \qquad (3.3)$$

Eq. (3.1) and eq. (3.3) both specify how the waiting times of the top-priority jobs depend on their upper-limit. Hence we assume

$$f^S(X) = g^M(Y) \quad \text{for} \quad X = Y \qquad (3.4)$$

Assumption 4: Q^M takes over the role of Q^L when Y approaches infinity

A similar reasoning as for assumption 3 can be made for M and L jobs. If Y becomes large, say Y = b, then eq. (3.2) becomes

$$Q^M = f^M(X,b) = h^M(X) \qquad (0 \leq X) \qquad (3.5)$$

The equation

$$Q^L = f^L(Y) \qquad (0 \leq Y) \qquad (3.6)$$

and eq. (3.5) both specify how the waiting times of the lowest priority jobs depend on their lower limit. Hence we assume

$$f^L(Y) = h^M(X) \quad \text{for} \quad X = Y \qquad (3.7)$$

Assumption 5: The functional relationships between the quantiles and the class limits may be approximated by low degree polynomials[6]

As we know from mathematics, a nicely behaving function may be represented through its Taylor series. We assume that this series may be cut off after the first or second derivatives. Hence we approximate the true functions by first or second degree polynomial regression models. e.g., eq. (3.2) is approximated by

$$\hat{Q}^M = \beta_0 + \beta_1 X + \beta_2 Y + \beta_{11} X^2 + \beta_{12} XY + \beta_{22} Y^2 \qquad (3.8)$$

We shall estimate the β coefficients through regression analysis. The resulting estimates will be tested for significance using the traditional

t-statistics.[7] Likewise assumptions 1, 2 and 5 yield

$$\hat{Q}^S = \alpha_0 + \alpha_1 X + \alpha_2 X^2 \qquad (3.9)$$

$$\hat{Q}^L = \gamma_0 + \gamma_1 Y + \gamma_2 Y^2 \qquad (3.10)$$

4. EMPIRICAL METAMODELS

Empirical results suggest that the interaction effect β_{12} in eq. (3.8) is unimportant. In other words, we compute the regression model (3.8) and find that $\hat{\beta}_{12}$ is insignificant.[8]

Originally very low R^2 values are found for L jobs, e.g. $R^2 = 0.63$ for eq. (3.10). On inspection of the scatter diagram this low R^2 is explained by the presence of some wild observations: outliers. These outliers occur because waiting times of L jobs show high variances; see Coppus et al. (1976; figure 6). The higher Y becomes, the fewer L jobs remain, and the higher their variance becomes. In simulation experiments - as opposed to other empirical work - we can check whether an outlier is indeed caused by chance. We simply repeat the X,Y combination with a new stream of random numbers. In this way we indeed check a number of suspicious observations on Q^L. All new observations fall within the "cloud" of observations. Obvious outliers are eliminated. In this way the total number of observations increases from 71 to 91.[9]

We accept an estimated regression model when it meets the following statistical criteria:

(1) It provides a good explanation of the changes in Q as X and/or Y vary: "high" R^2.

(2) The estimated individual effects $\hat{\beta}$ are significantly different from zero.

These two statistical requirements, together with the assumptions 1, 2 and 5 of the preceding section, yield purely quadratic models, i.e., in the eqs. (3.8) through (3.10) the first degree effects are insignificant, so that they are eliminated:

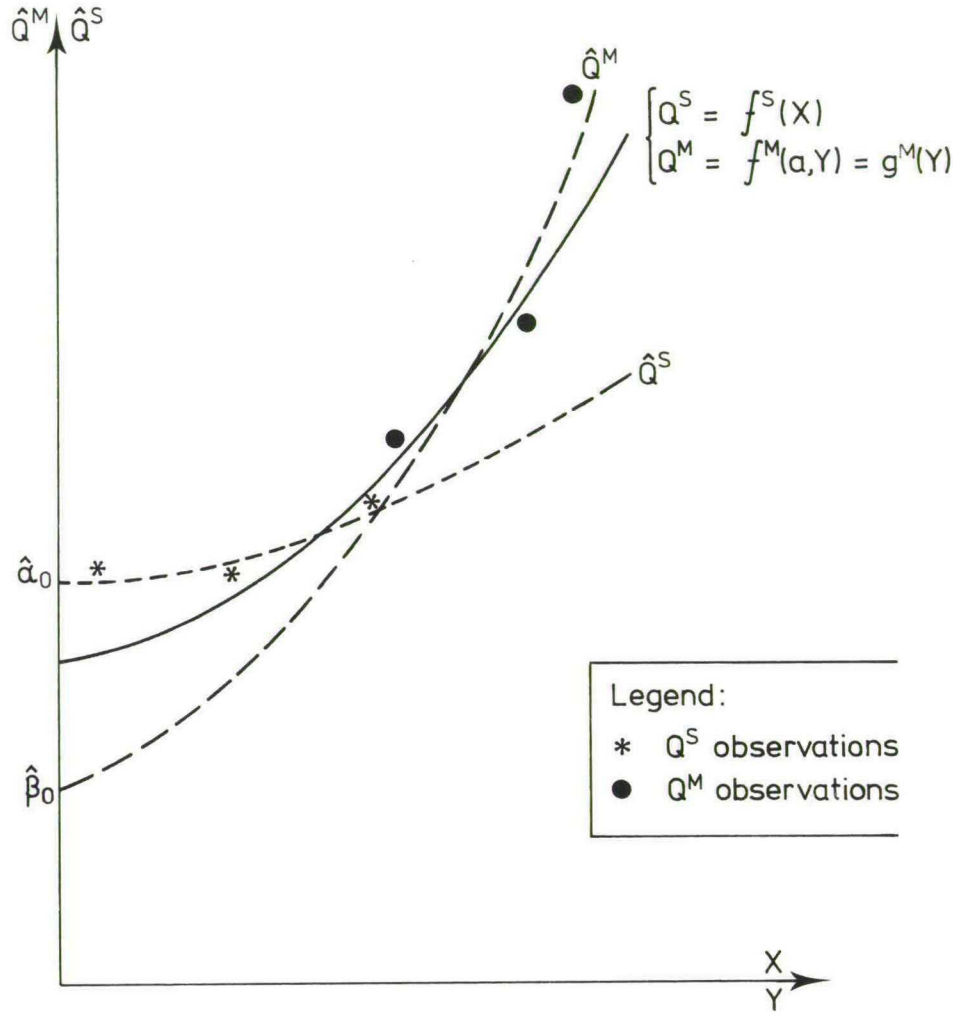

Fig.1 The intercept estimates in $\hat{Q}^S$ and $\hat{Q}^M$

$$\hat{Q}^S = 56.07 + 0.0087\ X^2$$
$$(t: \quad 89.1 \qquad 17.43) \qquad R^2 = 0.82 \qquad (4.1)$$

$$\hat{Q}^M = 30.5 + 0.0323\ X^2 + 0.0092\ Y^2$$
$$(t: \quad 3.9 \qquad 6.75 \qquad 31.19) \qquad R^2 = 0.95 \qquad (4.2)$$

$$\hat{Q}^L = 584.9 + 0.1564\ Y^2$$
$$(t: \quad 2.3 \qquad 16.0) \qquad R^2 = 0.74 \qquad (4.3)$$

This type of model also agrees with assumption 3 of section 2: $\hat{\beta}_{22} = 0.0092$ and $\hat{\alpha}_2 = 0.0087$ are not significantly different (at $\alpha = 0.05$). Alternative models fail at this point.[10]

We further note that the intercepts $\hat{\alpha}_0$ and $\hat{\beta}_0$ of eqs. (4.1) and (4.2) are significantly different, violating assumption 3. However, Q^S is estimated for small X values, whereas Q^M is estimated for large Y values (which take over the X role; see eq. 3.4). We assume that $\hat{\alpha}_0$ is a more reliable estimate than $\hat{\beta}_0$; see the corresponding t-statistics and also fig.1. We also note that the coefficients $\hat{\beta}_{11}$ and $\hat{\gamma}_2$ of eqs. (4.2) and (4.3) are significantly different, as are $\hat{\beta}_0$ and $\hat{\gamma}_0$, violating assumption 4.

The dual role of eq. (4.2) - metamodel for Q^M and/or Q^S when substituting X = a - leads to the following idea: When fitting the Q^M model take into account the Q^S observations, or more generally, fit the Q^S, Q^M and Q^L models simultaneously. This can be formalized as follows.[11]

Assumption 6: Per job-class the quantile Q depends on its left and right class limits L and R

More specifically this assumption together with the above results, yields table 1, where a corresponds to the minimum service time and b corresponds to the maximum service time (bigger key-punching jobs are served outside the computing center).

Left	Right	$\hat{Q} = \hat{f}(L,R)$
a	X	$\hat{Q}^S = \delta_0 + \delta_1 a^2 + \delta_2 X^2$
X	Y	$\hat{Q}^M = \delta_0 + \delta_1 X^2 + \delta_2 Y^2$
Y	b	$\hat{Q}^L = \delta_0 + \delta_1 Y^2 + \delta_2 b^2$

Table 1: Single, simultaneous model.

Comparison with eqs. (3.9) and (3.10) shows the identities $\alpha_0 \equiv \delta_0 + \delta_1 a^2$, $\alpha_2 \equiv \delta_2$, $\gamma_0 \equiv \delta_0 + b^2 \delta_2$, etc. The coefficients δ are estimated using all Q^S, Q^M and Q^L observations simultaneously:

$$Q = \delta_0 \underline{I} + \delta_1 \underline{Z}_1 + \delta_2 \underline{Z}_2 \tag{4.4}$$

with the vectors

$$\begin{aligned} \underline{Q}' &= (Q^S_1,\ldots,Q^S_{71},\ Q^M_1,\ldots,Q^M_{71},\ Q^L_1,\ldots,Q^L_{71}) \\ \underline{Z}_1' &= (a^2,\ldots,a^2,\ X^2_1,\ldots,X^2_{71},\ Y^2_1,\ldots,Y_{71}) \\ \underline{Z}_2' &= (X^2_1,\ldots,X^2_{71},\ Y^2_1,\ldots,Y^2_{71},\ b^2,\ldots,b^2) \end{aligned} \tag{4.5}$$

$\underline{I}$ being a vector with 71 elements equal to one. For a = 10 and b = 900 eq. (4.4) results in:

$$\begin{pmatrix} \hat{Q}^S \\ \hat{Q}^M \\ \hat{Q}^L \end{pmatrix} = 67 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 0.15 \begin{pmatrix} 10^2 \\ X^2 \\ Y^2 \end{pmatrix} + 0.00076 \begin{pmatrix} X^2 \\ Y^2 \\ 900^2 \end{pmatrix}$$

$$(t\ :\ 1.3 \qquad 23.75 \qquad 3.73 \qquad)\ R^2 = 0.90 \tag{4.6}$$

Though R^2 is high, examination of the residuals shows that the model systematically overestimates near the origin.[12] We assume that in the estimation of the intercept the influence of Q^M and Q^L observations far away from the origin, is to be blamed. Therefore we proceed to a new scheme.

Assumption 7: The $\hat{Q}^M$ model (4.2) provides the best estimates of the reaction coefficients δ_1 and δ_2 in eq. (4.4); the $\hat{Q}^S$ model provides the best estimates of the intercept δ_0

So as our starting point we take the Q^M model (4.2) (with high R^2) with $\hat{\delta}_1 = 0.0323$ and $\hat{\delta}_2 = 0.0092$. We substitute $\hat{\delta}_1$ and $\hat{\delta}_2$ into the $\hat{Q}^S$ model of table 1. Each of the 71 Q^S values yields an estimated intercept. Their average is $\hat{\delta}_0 = 52.33$.[13)]

Next we apply a similar procedure to determine an "effective' upperbound for the Q^L function. The factor b = 900 in Table 1 is an "absolute" upperbound, i.e. jobs with service times longer than 900 are not accepted.[14)] Hence we introduce:

Assumption 8: The "effective" upperbound for Q^L is not the absolute limit b = 900, but the upperbound under which virtually all jobs remain

To estimate the effective upperbound, we substitute $\hat{\delta}_1$ and $\hat{\delta}_2$ (from $\hat{Q}^M$) and $\hat{\delta}_0$ (from $\hat{Q}^S$) into the $\hat{Q}^L$ function of table 1. Each Q^L observation corresponds with an effective upperbound, its average being 563.15. Note that the probability of service times higher than 563.15 is virtually zero.

Upon substitution of b = 560 into eq. (4.5), simultaneous estimation via eq. (4.4) yields

$$\begin{pmatrix} \hat{Q}^S \\ \hat{Q}^M \\ \hat{Q}^L \end{pmatrix} = 47.9 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 0.123 \begin{pmatrix} 10^2 \\ X^2 \\ Y^2 \end{pmatrix} + 0.00400 \begin{pmatrix} X^2 \\ Y^2 \\ 560^2 \end{pmatrix} \qquad (4.7)$$

$$(t \ : \quad 0.61 \qquad 13.99 \qquad 5.43 \ ) \quad R^2 = 0.82$$

The corresponding residuals now show an acceptable pattern, i.e., no systematic over or underestimation occurs.[15)]

5. OPTIMIZATION OF THE CRITERION

The criterion function was shown in eqs. (2.4) through (2.6). The quantiles Q^S, Q^M, Q^L were approximated by simple functions in X and Y. Estimating these functions per quantile separately yielded eqs. (4.1) through (4.3). Simultaneous estimation resulted in eq. (4.7). Substituting these functions for the quantiles into eq. (2.4), and substituting the functions in X and Y for the weights W and the conditional mean service times B (eqs. 2.5 and 2.6) results in a criterion that is an explicit function in X and Y. Unfortunately, after taking derivates $\partial z/\partial X$ and $\partial z/\partial Y$, and setting these to zero, we cannot derive an explicit solution for the optimal values X_0 and Y_0. Therefore we resort to a computerized iterative search procedure (starting at X = 30 and Y = 180, the values in current use). The individual Q models of eqs. (4.1)-(4.3) yield X_0 = 44.83 and Y_0 = 177.19. The simultaneous model of eq. (4.7) results in X_0 = 47.46 and Y_0 = 183.65. These results have been immediately implemented by the computing center: Y has been maintained at 180 minutes, and X has been increased from the intuitively chosen 30 minutes to 45 minutes.

Note that we did not investigate the effects of changes in the arrival intensity λ and the service intensity μ. The effects of μ on the weights W and mean conditional service times B follow from eqs. (2.5) and (2.6). The effects of λ and μ on the quantiles Q require additional research.

6. CONCLUSION

In many practical queuing systems priorities are introduced so that small jobs (short service times) are served first. The resulting model cannot be solved analytically so that simulation is often used. Interpreting the voluminous simulation output data requires regression analysis. Choosing the appropriate regression model should be based on at least a qualitative theoretical analysis. We have illustrated the above philosophy with a real-life case for which practical results have been derived.

REFERENCES

1. Cobham, A., Priority assignment in waiting line problems. JOURNAL OF OPERATIONS RESEARCH SOCIETY OF AMERICA, 2, 195[illegible], pp. 70-76.

2. Coppus, G., M. van Dongen and J.P.C. Kleijnen, Quantile estimation in regenerative simulation: a case st dy. PERFORMANCE EVALUATION REVIEW, 5, no. 3, Summer 1977, pp. 5-15. (Reprinted in SIMULETTER, 8, no. 2, Jan. 1977, pp. 38-47.)

3. Crane, A. and J. Lemoine, AN INTRODUCTION TO THE REGENERATIVE METHOD FOR SIMULATION ANALYSIS. Springer-Verlag, Berlin, 1977.

4. Draper, N.R. and H. Smith, APPLIED REGRESSION ANALYSIS. John Wiley & Sons, Inc., New York, 1966.

5. Jaiswal, N.K., PRIORITY QUEUES. Academic Press, New York, 1968.

6. Kleijnen, J.P.C., STATISTICAL TECHNIQUES IN SIMULATION. (In two volumes.) Marcel Dekker, Inc., New York, 1974/1975.

7. Kleijnen, J.P.C., Generalizing simulation results through metamodels. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, vol. SMC-9, no. 2, Feb. 1979, pp. 93-96.

8. Meyers, R.H., RESPONSE SURFACE METHODOLOGY. Allyn and Bacon, Boston, 1971.

9. Scheffé, H., THE ANALYSIS OF VARIANCE. John Wiley & Sons, Inc., New York, 1959.

10. Van den Bogaard, W. and J.P.C. Kleijnen, MINIMIZING WAITING TIMES USING PRIORITY CLASSES: A CASE STUDY IN RESPONSE SURFACE METHODOLOGY. Department of Business and Economics, Katholieke Hogeschool, Tilburg (Netherlands), June 1977.

NOTES

1. This research was performed by F. Keyzer and E. Mullenders, graduate students in the Department of Econometrics, as a project for their simulation course, under the guidance of J. Kleijnen, Senior Research Associate in the Department of Economics, and with the cooperation of A. van Reeken, manager of the Computing Center at Tilburg University.

2. We can estimate the quantiles with known statistical accuracy. To compute confidence intervals we analyze the simulation using its "regenerative" property, i.e., when the system becomes empty (both servers idle), a new history starts independently of the past simulated history; we refer to Crane and Lemoine (1977) for a general exposé of renewal analysis in simulation, and to Coppus et al. (1976) for the technical details of this analysis when applied to our problem.

3. Van den Bogaard and Kleijnen (1977) continued with an approach different from Coppus et al. (1976) and the present paper, i.e., Response Surface Methodology (RSM) was applied. That is, <u>locally</u> a linear model $\hat{Q} = \hat{\alpha}_0 + \hat{\alpha}_1 X + \hat{\alpha}_2 Y$ is fitted. The signs of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ tell whether X and Y should be increased or decreased in order to minimize Q. The relative changes in X and Y depend on the ratio $\hat{\alpha}_1/\hat{\alpha}_2$, so-called steepest descent method. At the "bottom of the valley", the linear model becomes a bad guide for determining the X, Y effects. In that stage of experimentation a quadratic model for <u>local</u> search is introduced. Such a model also reveals the character of the optimum region: is there a unique minimum, a saddle point or a ridge? For details on RSM we refer to, e.g. Meyers (1971) and Kleijnen (1975). In our application a problem was that the path of steepest descent for large jobs differed very much from the paths for medium and small jobs.

4) $$W^S = P(S < X) = \int_a^X \mu e^{-\mu p} dp = e^{-\mu a} - e^{-\mu X}$$

$$W^M = P(X < s < Y) = \int_X^Y \mu e^{-\mu p} dp = e^{-\mu X} - e^{-\mu Y}$$

$$W^L = P(s > Y) = \int_Y^b \mu e^{-\mu p} dp = e^{-\mu Y} - e^{-\mu b}$$

$$B^S = E(s|s < X) = E(s).(P\{s < X\})^{-1} = \int_a^X p.\mu\, e^{-\mu p}\, dp.\left[\int_a^X \mu\, e^{-\mu p}\, dp\right]^{-1} =$$

$$= \{(a+\tfrac{1}{\mu})e^{-\mu a} - (X+\tfrac{1}{\mu})e^{-\mu X}\}.[\{e^{-\mu a} - e^{-\mu X}\}]^{-1}$$

$$B^M = \int_X^Y p.\mu.e^{-\mu p}\, dp = \left[\int_X^Y \mu.e^{-\mu p}\, dp\right]^{-1} = \{(X+\tfrac{1}{\mu})e^{-\mu X} - (Y+\tfrac{1}{\mu})e^{-\mu Y})\}.$$
$$.[e^{-\mu X} - e^{-\mu Y}]^{-1}$$

$$B^L = \int_Y^b p.\mu.e^{-\mu p}\, dp.\left[\int_Y^b \mu.e^{-\mu p}\, dp\right]^{-1} = \{(Y+\tfrac{1}{\mu})e^{-\mu Y} - (b+\tfrac{1}{\mu})e^{-\mu b}\}.$$
$$.[e^{-\mu Y} - e^{-\mu b}]^{-1}$$

5. Although the purpose of the first paper in this series, namely Coppus et al. (1976), was to estimate the variances of the quantiles, we did not have these variances available for all 71 observations, at least not in an easily usable form. Therefore our regression models are based on Ordinary Least Squares implying a common unknown variance σ^2, and the regression parameters' significance is measured through t-statistics using an estimate of σ^2 based on the Mean Squared Residuals; see regression textbooks such as Draper & Smith (1966).

6. On hindsight it might be interesting to investigate approximations based on exponentials:
   (1) Exponential functions show a behavior comparable to the second-degree polynomials used in this report.
   (2) Exponential behavior is often met in queuing theory.
   (3) Exponentials <u>might</u> lead to an explicit solution for the optimal X and Y upon differentiation of the overall criterion z which comprises a number of exponentials.

7. If a first-order model were used, then we would test the signs of the estimated coefficients: one-sided instead of two-sided test. For instance, for Q^S we would hypothesize that $\alpha_1 > 0$ (or $\partial Q^S/\partial X > 0$); see eq. (3.9) with $\alpha_2 = 0$.

8. The scatter diagrams show that for fixed X parallel curves in the Q, Y plane result. For some X, Y, Q combinations very irregular patterns were found.

9. More precisely, first 5 suspected observations are simulated again, each replicated twice. These 10 new observations show that the suspected observations are outliers indeed. Next 15 more suspected observations are simulated again. However, these observations need not be rejected.

10. The best linear unbiased estimate (BLUE) of, say, α_1 in eq. (3.9) is $\hat{\alpha}_1$. So , even though $\hat{\alpha}_1$ is insignificant; we might retain this value in eq. (3.9). This is a moot issue in statistics. In our case, however, assumption 3 forces the issue.

11. An alternative model also applying simultaneous estimation is as follows. Each X, Y combination yields an observation on Q^S, Q^M and Q^L respectively. Hence we represent each Q^S, Q^M or Q^L observation in a three-dimensional space with axes X, Y and, say, Q. Then Q^S is assumed to depend on X only so that its observation vectors are projected onto the X, Q plane. Analogously we project each Q^L vector onto the Y, Q plane. Table 1 is then replaced by:

Original		New		
X	Y	X'	Y'	Q
X	Y	X	0	Q^S
X	Y	X	Y	Q^M
X	Y	0	Y	Q^L

and in eq. (4.5) we substitute

$$\underline{Z}_1' = (X_1^2,\ldots,X_{71}^2,\ X_1^2,\ldots,X_{71}^2,\ 0,\ldots,0)$$

$$\underline{Z}_2' = (0,\ldots,0\quad,\ Y_1^2,\ldots,Y_{71}^2,\ Y_1^2,\ldots,Y_{71}^2)$$

However, the results of this model are very bad. Moreover assumption 3 can be satisfied only if $\delta_1 = \delta_2$, since $\partial Q^S/\partial X = 2\delta_1 X$ and

$[\partial Q^M/\partial Y]_{X=a} = 2\delta_2 Y$.

12. Compare also $\hat{\alpha}_0 = 56.07$ with $\hat{\delta}_0 + \hat{\delta}_1 . a^2 = 67 + (0.15)(100) = 82$.

13. This approach could have been improved by an iterative procedure: Re-estimate $\hat{Q}^M$ under the condition $\delta_0 = 52.33$; use the resulting estimates of δ_1 and δ_2 in $\hat{Q}^S$, etc.

14. In practice a special class of jobs is submitted to the computing center, requiring excessive key-punching time, namely more than 900 minutes. These jobs were not included in the arrival and service distributions.

15. As in note 12 we compare $\hat{\alpha}_0 = 56.07$ with $\hat{\delta}_0 + \hat{\delta}_1 a^2 = 52.33 + (0.03 3)$ (100) = 55.56, a much better result indeed.