UNIVERSITY
HOLIEKE
UNIVERSITEIT
BRABANT

POSTBOX 90153
5000 LE TILBURG
THE NETHERLANDS

[redacted]

DEPARTMENT OF ECONOMICS

RESEARCH MEMORANDUM

[redacted]

SIMULATION AND OPTIMIZATION IN PRODUCTION PLANNING: A CASE STUDY

Jack P.C. Kleijnen

FEW 308

SIMULATION AND OPTIMIZATION IN
PRODUCTION PLANNING: A CASE STUDY

Jack P.C. Kleijnen

Department of Information Systems
School of Business and Economics
Catholic University Brabant (KUB)
5000 LE Tilburg
Netherlands

Version 2: January 1988
Version 1: September 1987

1) Version 1 was titled: "A DSS for production planning: a case study including simulation and optimization".

CONTENTS

SIMULATION AND OPTIMIZATION IN PRODUCTION PLANNING: A CASE STUDY

Jack P.C. Kleijnen
Department of Information Systems and Auditing
School of Business and Economics
Catholic University Brabant (KUB)
5000 LE Tilburg
Netherlands

ABSTRACT

This paper reports on a practical decision support system (DSS) for production planning, developed for a Dutch company. To evaluate this DSS, a simulation model is built. Moreover, the DSS has 15 control variables which are to be optimized. The effects of these 15 variables are investigated, using a sequence of 2^{k-p} experimental designs. Originally 34 response variables were distinguished. These 34 variables, however, can be reduced to one criterion variable, namely productive machine hours, which is to be maximized, and one commercial variable measuring lead times, which must satisfy a certain side-condition. For this optimization problem the Steepest Ascent technique is applied to the experimental design outcomes. The resulting Response Surface Methodology is developed theoretically. In practice a number of complications arise.

Keywords: heuristics, regression analysis, multiple criteria

1. INTRODUCTION: PROLOGUE AND OVERVIEW

This paper presents a case study concerning a decision support system (DSS) for production planning in metal tube manufacturing. For proprietary reasons it should suffice to characterize the company as follows. The factory makes different types of product, on order. The major initial problem was the lead times: a drastic reduction seemed possible (maybe 50%; in § 7 we shall indeed realize a 62.8% reduction). First the company investigated Material Requirements Planning (MRP-I) and Manufacturing Resource Planning (MRP-II) but found this type of approach not suitable for its job-shop production process. Next a team of operations researchers started to develop a DSS especially for this company. This DSS should yield daily production orders (some details are given in Section 2). It was too risky to implement the DSS without further testing and fine tuning. Therefore this OR team developed a simulation program (in SIMULA). Fine tuning concerned 15 parameters or control variables of the DSS. Preliminary sensitivity analysis with the simulated DSS had just started. A major technical problem was that one simulation run took 6 hours on the company's mainframe, provided the simulation program is executed at night when no other jobs are run. Hence, sensitivity analysis as initially designed, would require about 30 runs or 180 hours of computing time. That was a prohibitive amount of computer time. Therefore I was invited to apply special statistical techniques to this problem; also see Kleijnen (1974/ 1975, 1987).

This case study illustrates practical problems such as lack of data, time pressures, and compromises to be made when modeling complex systems in an organizational context. We further show how a set of 34 responses can be reduced to only two responses (see § 4). The study also demonstrates the use of techniques, namely experimental design, regression analysis, and steepest ascent. These techniques are standard for the expert in the various fields; nevertheless, in practice operations researchers are often unfamiliar with techniques such as 2^{k-p} designs (see Section 3). Moreover, we add a novel idea to the steepest ascent technique in situations with multiple responses (see § 6).

This paper is organized as follows.

§ 2 "A production planning system" describes the job shop and the DSS, emphasizing commercial and production goals.

§ 3 "The original simulation and experimental design" presents an (inferior) one-factor-at-a-time design, a large set of simulation responses, and the original regression model.

§ 4 "Reconsidering the problem" reduces the original 34 responses to only two responses; the production manager is interested in maximizing response 1 without violating an upper limit for response 2, a commercial variable.
§ 5 "Basic experimental design and results" uses a 2^{14-10} design for a local first-order model in the first stage of experimentation, which results in better control settings and in estimated local first-order effects (which will guide the second stage of experimentation); no control variables are eliminated at this stage!
§ 6 "Multi-variate optimization: theory" applies the steepest ascent technique to the estimated local first-order model for response 1 (see § 4), while considering the linear constraint on response 2, to determine the maximum step size along the steepest ascent path (determining the step size in this way, seems novel in Response Surface Methodology or RSM).
§ 7 "Practical multi-variate optimization" does not determine the maximum step size (since the linear constraint of § 6 could not be quantified soon enough); instead it uses heuristics to determine the step size: a second 2^{14-10} experiment is executed which results in improved control settings; further exploration was stopped because the project was aborted.
§ 8 "Epilogue: simulation methodology" criticizes the "shadow" or "parallel" running approach which gives unfair comparisons between the simulation model's output and the human planner's output; this section briefly discusses validation, optimization, and sensitivity and robustness analyses.
§ 9 "Conclusions" summarizes the paper.

2. A PRODUCTION PLANNING SYSTEM

The company has 6 machines on which different classes or types of products can be produced. (Together these classes comprise at least 700 different products.) When a machine switches to a different class of products, major costs are incurred, i.e., major adjustments to a machine must be made and during a sizable period of time no production is possible. (Switchover costs within a class are minor.) To minimize these production losses, it is desirable to have long production runs. Such a policy, however, would yield long lead times. Therefore it is necessary to balance commercial and production goals. The OR team developed a heuristic Production Planning System (PPS), including 15 control variables or parameters x_j (with j = 1,...,15). For example, x_1 is a "penalty for producing class-2 products on the next best machine"; obviously this penalty can be manipulated to improve the DSS performance. For this paper, the DSS is a black box. We can indeed treat the DSS as a black box, since our methodology (2^{k-p} designs and Steepest Ascent) does not depend on specific knowledge about the DSS and the corresponding production system. (Of course, actual values resulting from the standardized design do depend on the specific system; see Table 3 later on.) Another reason for treating the DSS as a black box is the proprietary character of the system. Moreover, this paper would become too long, were the details of the DSS heuristics explained. These heuristic were the result of much effort by a number of operations researchers.[1)]

So from a technological viewpoint there are many different products (at least 700) which can be grouped into different "product" classes such that switchover costs are minor within a product class and major between classes. From a commercial viewpoint there are five different "order" classes; for example, class-1 orders are emergency or rush orders, i.e., a customer must be supplied "immediately". An individual order may comprise different products.

3. THE ORIGINAL SIMULATION AND EXPERIMENTAL DESIGN

The OR team selected the following simulation approach (which we shall criticize in § 8). The DSS was programmed and fed with the historical data of July 1986 through October 1986. For those 4 months detailed data are available on orders (several thousands), changes in orders (30% of the orders is revised), machine breakdowns, etc. By definition, one simulation run implies constant values for the 15 DSS parameters, during those 4 months. The original experimental design for sensitivity analysis uses the one-factor-at-a-time method:

Run 1: Fix the 15 control variables at their base values (say) x_j^b with j = 1,...,15. (These base values were suggested by the developers of the DSS using "common sense". Common sense implies subjectivity so that there are good reasons indeed to perform sensitivity analysis. The values x_j^b will be displayed in Table 3 for 14 of the original 15 control variables.)

Run 2: Increase variable x_1 by 20% and keep all other 14 variables at their base values.

Run 3: Decrease variable x_1 by 20% and keep $x_{j'} = x_{j'}^b$, with $j' = 2,\dots,15$.

Run 4: Increase x_2 by 20% and keep all other variables at their base values ($x_1 = x_1^b$, $x_3 = x_3^b$, ..., $x_{15} = x_{15}^b$).

And so on. Altogether this approach would take $1 + 2 \times 15 = 31$ runs. It is well-known in the experimental design literature that the one-at-a-time method is inferior, compared to factorial designs. (Nevertheless operations researchers often apply this inferior design, as this case study illustrates.) So only $2^{15-11} = 16$ runs suffice to estimate the individual effects of 15 variables; see Kleijnen (1987) and table 1. Moreover, optimization takes <u>several</u> rounds of experimentation and analysis, as we shall see in later sections, so that efficient designs become even more desirable.

The original idea was to evaluate each run using the following aspects (verbatim):

"a. Average and spread of realized lead times, for orders of types 1, 2 and 3.

b. Average and spread in lead time inaccuracy (= absolute value of realized lead time minus promised lead time) for orders of types 1, 2, 3, 4 and 5.

c. Utilization degree = production hours/(production hours + idle time + switchover time) × 100%.

d. Switchover degree = switchover time/(production hours + idle time + switchover time) × 100%"

For each aspect (say) y, the OR team wanted to fit a <u>regression</u> model. They assumed that the following first-order approximation would be adequate in the first stage of the investigation (also see the last paragraph of § 6):

$$y_i = \beta_0 + \sum_{j=1}^{15} \beta_j x_{ij} + e_{ij} \quad (i = 1,\dots,31) \tag{3.1}$$

where the regression parameter β_j denotes the effect of the control variable x_j; β_0 is the overall response; the OR team assumed that the classical assumptions hold, i.e., the errors e_{ij} are Normally and Independently Distributed with mean zero and constant variance σ^2:

$$e_{ij} \sim NID(0,\sigma^2) \ . \tag{3.2}$$

Ordinary Least Squares (OLS) yield the estimators $\hat{\beta}$. If $\hat{\beta}_j$ is significant, according to the classical t test, then the OR team would follow up with a more extensive experimental design exploring only the significant variables x_j. Actually we shall not test $\hat{\beta}_j$ for significance; hence we shall not need the assumptions of eq. (3.2) (see the text below eq. 5.3).

4. RECONSIDERING THE PROBLEM

The preceding section lists many aspects thought to be relevant for the evaluation of the DSS. We count 34 aspects, namely the average realized lead time of class-1 orders, the spread of the preceding variable, the average for class-2 orders, ..., the spread of realized lead times for class-5 orders, the average promised lead time of class-1 orders, ..., the utilization degree of machine 1, ..., the utilization degree of machine 6, the switchover degree of machine 1, ..., the switchover degree of machine 6.

Obviously, managers cannot select a system accounting for 34 aspects. (Miller, 1956, wrote a famous article with the revealing title "The magical number seven plus or minus two: some limits on our capacity for processing information".) Therefore we proposed to the client to reconsider the original problem formulation and to try and reduce the number of criteria drastically. The preceding section (sub c and d) mentions the "utilization degree" and the "switchover degree" per machine. We can derive, as follows, that each machine has its own average contribution (during "utilization") to gross profits. Each machine is technically more suited for certain products: not all products can be made on all machines, and if a product can be made on more than one machine then those machines are not equally good. Moreover, profits margins differ over products.

Accounting data are indeed collected on the average gross-profit contribution per machine. In this way utilization and switchover degrees for each of the 6 machines (together 2 × 6 = 12 variables) can be combined into a single variable per simulation run, namely profit contribution (say) y.

The preceding section lists - besides utilization and switchover degrees - "averages" and "spreads" of "realized" and "promised" lead times (the absolute difference between realized and promised times is the lead time inaccuracy), for each of the five order classes. Theoretically these many aspects of lead time can be translated into financial terms, since a reduction in (for example) realized lead times leads to more orders (goodwill effect). In practice, it is hard to quantify the financial consequences of (say) reducing the actual lead time from 27 days to 26 days. In our view it is management's job to specify a maximum for acceptable lead times. (Analogy: inventory control theory assumes that the financial consequences of out-of-stocks can be specified, whereas in practice management specifies an acceptable service percentage.)

We are still confronted with lead times for five order classes. Actually lead times are not critical for class-4 and -5 orders (by definition). As the outcome of several discussions with the client, we decided to focus on orders in class 2, one reason being that class-2 orders form the "major" part of the order portfolio. (In inventory control there is the 20-80 rule: 20% of the articles account for the "major" part, namely 80% of the sales volume.)

In order to reduce the number of aspects further, we observe that lead time inaccuracy is negligible, according to historical data. Therefore we concentrate on promised lead times. (We ignore realized lead times when performing sensitivity analysis and optimization of the DSS; yet the simulation does report realized lead times and lead times for orders in classes other than class 2.)

A final step concerns the distinction between the average and the spread of ℓ, lead times promised for class-2 orders. These two measures can be easily combined into <u>quantiles</u>, i.e., we use the 90% quantile (say) z:

$$P(\ell \leq z) = 0.90 \quad (4.1)$$

To estimate z we sort all individual lead times ℓ which are promised to class-2 customers during one simulation run; z is the value exceeded by only 10% of these individual lead times. This procedure yields an asymptotically unbiased estimator, whether the observations ℓ are correlated or not (they are correlated, since they come from a single run). The autocorrelation would become a serious problem, if we were to estimate the variance of the estimated quantile; see Kleijnen (1987, p. 82) for a detailed discussion. Actually we do not need $\widehat{var}(\hat{z})$ (also see note 4 later on).

Note that the selection of the 90% instead of the 95% or 99% quantiles, is quite arbitrary. (Quantiles rather than averages are also used in the optimization of priority class queues in a computer center case study; see Kleijnen, 1987, pp. 214-216.)

<u>Summary</u>: We succeeded in reducing the original 34 evaluation aspects of the DSS to only 2 variables. One variable y is the total profit contribution by the 6 machines, and should be maximized. The other variable z is the 90% quantile of promised (approximately equal to realized) lead times of class-2 orders (the most important type of orders). So the production manager should try to maximize y without violating a maximum value for the quantile of lead times, to be quantified by the marketing manager. All other aspects are also measured in the simulation, but they do not explicitly control the optimization of the DSS.

Table 1: 2^{14-10} experimental design $\mathbf{D} = (d_{ij})$.
(+ means +1 and - means -1; 5 = 12 means $d_{i5} = d_{i1}\, d_{i2}$, etc.)

run	1	2	3	4	5 = 12	6 = 13	7 = 14	8 = 23	9 = 24	10= 34	11= 123	12= 124	13= 134	14= 234
1	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2	-	+	+	+	-	-	-	+	+	+	-	-	-	+
3	+	-	+	+	-	+	+	-	-	+	-	-	+	-
4	-	-	+	+	+	-	-	-	-	+	+	+	-	-
5	+	+	-	+	+	-	+	-	+	-	-	+	-	-
6	-	+	-	+	-	+	-	-	+	-	+	-	+	-
7	+	-	-	+	-	-	+	+	-	-	+	-	-	+
8	-	-	-	+	+	+	-	+	-	-	-	+	+	+
9	+	+	+	-	+	+	-	+	-	-	+	-	-	-
10	-	+	+	-	-	-	+	+	-	-	-	+	+	-
11	+	-	+	-	-	+	-	-	+	-	-	+	-	+
12	-	-	+	-	+	-	+	-	+	-	+	-	+	+
13	+	+	-	-	+	-	-	-	-	+	-	-	+	+
14	-	+	-	-	-	+	+	-	-	+	+	+	-	+
15	+	-	-	-	-	-	-	+	+	+	+	+	+	-
16	-	-	-	-	+	+	+	+	+	+	-	-	-	-

5. BASIC EXPERIMENTAL DESIGN AND RESULTS

At the outset the OR team considered 15 control variables which were to be investigated in a first experiment of 31 simulation runs (see § 3). Note that experimental design theory speaks of "factors" instead of "control variables".[2)]

Upon closer examination we find that 2 of the 15 factors can be combined into a single factor (to explain this, we would have to explain the DSS heuristics; see § 2). To optimize the remaining 14 variables x_j (j = 1,..., 14), we apply Response Surface Methodology (RSM); also see Kleijnen (1987, pp. 202-206). So we start with a local first-order approximation (also see eq. 3.1):

$$y = \beta_0 + \sum_{1}^{14} \beta_j x_j + e . \qquad (5.1)$$

RSM assumes that in the first stages of experimentation with the simulation model, a first-order model is good enough to guide the search for better responses (also see § 6). § 3 gave a one-factor-at-a-time design with 31 runs. Actually the 15 regression parameters β in eq. (5.1) can be estimated without bias, using a classical 2^{14-10} design, which takes only 16 runs (and a single run requires 6 hours of computer time, so that the savings are substantial). (Moreover, if the errors were independently and identically distributed with zero mean, then a 2^{14-10} design would be optimal; for example, $\hat{var}(\hat{\beta}_j)$ would be minimal. We do not use this particular error specification in our analysis. Also see Kleijnen, 1987, pp. 334-337.) The design matrix **D** is displayed in Table 1 (readers familiar with experimental design do not need Table 1: a 2^{14-10} design is fully specified, once we give the 10 generators 5 = 12, 6 = 13,...,14 = 234 which are also listed in Table 1).

To obtain the matrix of independent variables **X** corresponding to eq. (5.1), we (arbitrarily) associate the levels +1 and -1 of **D** in Table 1 with the actual "low" and "high" values of the control variables in the DSS, i.e., in the first experiment (comprising 16 runs; more experiments will follow) +1 in Table 1 corresponds to the base values (which were specified using "common sense"; see § 3) and -1 corresponds to 20% higher values (taking 20% is rather arbitrary: the control variables have not much intuitive meaning so that it is difficult to specify a "high" value). Finally D is augmented with a column of 16 one's corresponding to β_0 and we compute the OLS estimator

$$\hat{\beta} = (X'X)^{-1}X'y \tag{5.2}$$

where the vector y equals $(y_1, ..., y_i, ..., y_{16})$ and y_i denotes the total number of productive hours of the 6 machines in run i. In the preceding section we introduced the "profit contribution". However, it turned out to be impossible to obtain the accounting data and to incorporate them in the simulation program, at short notice (lack of data is a well-known problem in OR implementation). Obviously productive hours and profit contribution are closely related: both responses eliminate idle time and switchover time, but profit contribution also accounts for different contributions per machine.

We also observe z_i, the 90% quantile of promised lead times for order category 2 in run i, and we estimate γ the effects of the control variables on z:

$$\hat{\gamma} = (X'X)^{-1}X'z \ . \qquad 5.3)$$

We do not eliminate factors with small $\hat{\beta}$ and $\hat{\gamma}$ effects: in RSM we fit a first-order model only locally, and we use the estimated first-order effects only to determine the direction of our search for better control settings of x_j (see figure 1 later on). As we move in stages through the experimental area $(0 \le x_j < \infty)$, the local first-order effects change. We do not eliminate factors, because a factor non-significant in one stage, may become significant in a later stage. (A significance test would use the Student t statistic which requires the estimators $\hat{var}(\hat{\beta}_j)$ and $\hat{var}(\hat{\gamma}_j)$ and the error specification of eq. 3.2.)[3)]

Note that reducing the number of factors from 15 to 14 does not decrease the required number of runs: the way incomplete factorial designs are constructed, implies that the number of runs must be a multiple of four exceeding the number of first-order effects; see Kleijnen (1987, pp. 301-303). So the number of runs remains 16 (= $2^{14-10} = 2^{15-11}$). The degrees of freedom increase from 16-16 to 16-15, but we do not use these degrees of freedom to estimate var(e) = σ^2 in (3.2), as we explained at the end of the preceding paragraph.

For confidentiality reasons we do not display the responses y_i and z_i (i = 1,...,16). However, we do give some comments on the results and we do display the changes in y and z caused by changes in the control variables x_j, that is, we do display the local sensitivity estimates $\hat{\beta}_j$ and $\hat{\gamma}_j$ in Table 2.

(i) Run 1 of the design in Table 1 corresponds to the base run (the common sense setting of the control variables). Other settings yield more productive hours (higher y) while resulting in lower lead times (lower z). For example, run 2 increases y by 0.7% and decreases z by 13.4%; run 4 increases y by 1.6% and decreases z by 9.5%. So our design identifies control settings which "dominate" the base setting.

(ii) Some factors have favorable (local) effects on both responses, y and z; see Table 2. For example, factor 1 increases y (because $\hat{\beta}_1 > 0$) and decreases z (since $\hat{\gamma}_1 < 0$). Factor 4 shows $\hat{\beta}_4 < 0$ and $\hat{\gamma}_4 > 0$ so that it is attractive to decrease x_4. In run 2 (see (i) above) these factors have the good settings: $d_{21} = -1$ and $d_{24} = +1$ (see Table 1).

(iii) To evaluate the effect of factor j we should consider, not the unit effects $\hat{\beta}_\gamma$ and $\hat{\gamma}_j$, but the products $\hat{\beta}_j x_j^b$ and $\hat{\gamma}_j x_j^b$ (where x_j^b denotes the base run value of factor j; see the third column in Table 3). The reason is that the factors have different scales and ranges; also see Bettonvil and Kleijnen (1988) and Kleijnen (1987, pp. 141-142).

Table 2: Local sensitivity estimates $\hat{\beta}$ and $\hat{\gamma}$.
(x_j^b denotes the base run value.)

Control variable x_j	Productive hours y		Lead time z	
j	$\hat{\beta}_j$	$\hat{\beta}_j\ x_j^b$	$\hat{\gamma}_j$	$\hat{\gamma}_j\ x_j^b$
1	0.52	62.40	-0.054	-6.48
2	-39.30	-117.90	-1.504	-4.51
3	0.65	78.00	0.072	8.64
4	-18.07	-0.90	150.583	7.53
5	-128.96	-64.48	-16.519	-8.26
6	0.00	0.00	-0.102	-29.38
7	-0.22	-132.00	-0.006	-3.60
8	13.88	20.82	2.963	4.44
9	-1.53	-38.25	1.311	32.78
10	1.39	139.00	0.072	7.20
11	0.03	9.00	0.037	11.10
12	527.23	158.17	8.485	2.55
13	-9.27	-46.35	-6.351	-31.76
14	-0.46	-55.20	-0.145	-17.40

6. MULTI-VARIATE OPTIMIZATION: THEORY

Optimization of simulated systems is a well-known problem area, i.e., such optimization may be desired in practice but there is no standard mathematical technique to optimize a non-linear, possibly stochastic, system with multiple responses. Kleijnen (1987, pp. 202-206) surveys different techniques such as RSM and coordinate search, and complications due to side conditions and multiple responses. Hoerl (1985, p. 190) states "... multiple responses ... is basically an unsolved problem ...". In the present case study we needed a fast solution due to time constraints, and we developed the following approach which turns out to work (see the results at the end of § 7).[4)]

§ 4 shows that we wish to optimize y, the total profit contribution by the 6 machines, under the restriction that z, the 90% quantile of promised lead times for order category 2, does not exceed a prespecified limit (say) zmax. For practical reasons, we redefined y as the total number of productive hours; see § 5. Quantifying the commercial limit zmax is more difficult. The idea is that a higher zmax results in a higher maximum for y, and that at the end of our investigation management selects an attractive combination of zmax and max (y). (Analogy: in practical inventory control, management selects a combination of service percentage and inventory investment.)

The mathematical problem becomes (also see Figure 1 later on): maximize

$$\hat{y} = \hat{\beta}_0 + \sum_1^{14} \hat{\beta}_j \, x_j \qquad (6.1)$$

under the restriction[5)]

$$\hat{z} = \hat{\gamma}_0 + \sum_1^{14} \hat{\gamma}_j \, x_j \leq \text{zmax} \; . \qquad (6.2)$$

Because eqs. (6.1) and (6.2) are fitted only locally, we know that these two equations do not hold over the whole area of interest $(0 \leq x_j < \infty)$. Therefore it makes no sense to apply Linear Programming to eqs. (6.1) and (6.2) (also see note 7 later on). Instead, we proceed as follows.

The sign of $\hat{\beta}_j$ shows whether x_j should be increased or decreased to maximize y. Actually the path of <u>steepest ascent</u> implies

$$\frac{\Delta x_j}{\Delta x_1} = \frac{\hat{\beta}_j}{\hat{\beta}_1} \quad (j = 1,\ldots,14) \; , \qquad (6.3)$$

that is, this path is perpendicular to the hyperplane (6.1). The step size along this path must be selected arbitrarily and depends on the scaling of the independent variables x_j. To test the goodness of this path we propose to ask the following two questions:[6)]

(i) Does y indeed increase?

(ii) Does z indeed remain below zmax?

Figure 1 illustrates the situation for only 2 control variables. We emphasize that the steepest ascent path is based on local and estimated values $\hat{\beta}_j$. In Figure 1 the local experiment in the first stage is the subdomain represented by the rectangle ABCD. The iso-production line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ is shown only for that subdomain (because this line holds only locally). The illustration implies that the condition $z = \hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 a_2 \leq z\ max$ is not violated by any of the observed responses z_i corresponding to A, B, C, D. If the local estimates hold far outside the subdomain, then the greatest step along the steepest ascent path takes us to P, the intersection of the steepest ascent path and the restriction.[7)] Actually Q, the starting point of the steepest ascent path, is selected arbitrarily, i.e., several parallel paths could have been drawn in Figure 1. For example, if C shows the highest production then a better heuristic seems to start from C which leads to P'. The difference between P and P', however, is not really important, because P and P' are both computed from observations far away from P and P' (namely A, B, C, D); so we must repeat the (first) experiment in the neighborhood of P and P' (not shown in Figure 1). The second experiment should reveal whether indeed the simulation response y increases (see question i above) and whether $z \leq zmax$ (see question ii). We can start this second local experiment with a first run with the control variables fixed to the values corresponding to P (or, maybe better, P'). The following situations are possible, where the first experiment comprised n runs (n = 4 in Figure 1 but $n = 2^{14-10}$ in the actual case) and n+1 corresponds to P:

(i) $y_{n+1} > \max_{1 \leq i \leq n} (y_i)$ and $z_{n+1} \leq zmax$.

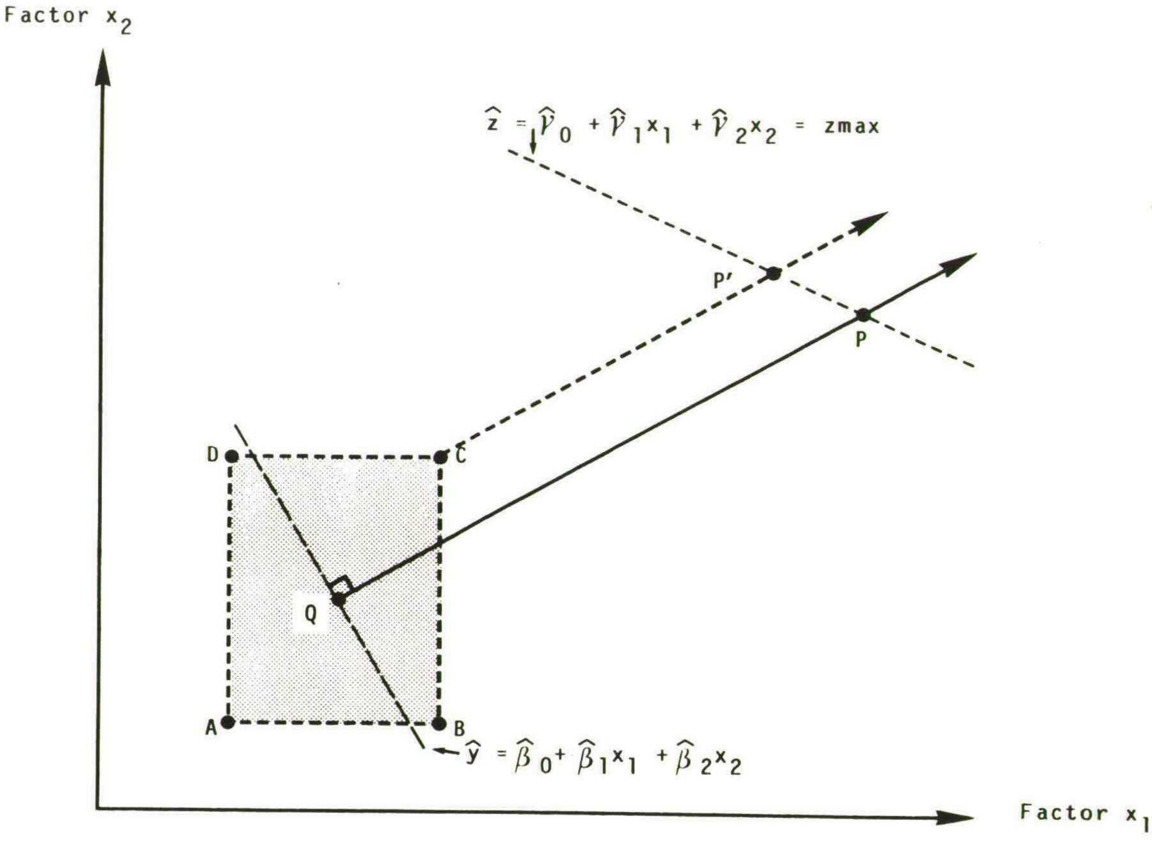
Then we continue to experiment around P and execute a new 2^{14-10} design: see Table 1 where run 1 now corresponds to P.

(ii) $y_{n+1} < \max (y_i)$ and $z_{n+1} > zmax$.

Then the local approximations do not hold outside the subdomain of the first experiment. We might try a point (say) halfway between Q and P.

(iii) $y_{n+1} > \max (y_i)$ and $z_{n+1} > zmax$.

Figure 1. Steepest Ascent Path with One Restriction.



Since the commercial restriction is violated, we have to back up on the steepest ascent path. If z_{n+1} is only slightly higher than zmax, then we back up only a little. We may use linear interpolation, defining S_0 and S_n to be the old and new step-sizes (i.e. S_0 is the distance between P and Q) and defining z_Q to be the z value corresponding to Q:

$$\frac{S_n}{S_0 - S_n} = \frac{zmax - z_Q}{z_{n+1} - zmax} . \qquad (6.4)$$

iv) $y_{n+1} < \max(y_i)$ and $z_{n+1} \leq zmax$.
We may proceed as in situation (ii).

Note that as we move into the optimal area, the first-order approximation of eq. (5.1) or (6.1) becomes less adequate so that we have to switch to a second-order approxiation. This fine-tuning requires the estimation of interactions $\beta_{jj'}$ ($j' = 2,\ldots,14$ and $j' > j$) and purely quadratic effects β_{jj}. See Kleijnen (1987, pp. 202-208, 312-316) (also see note 4 again).

7. PRACTICAL MULTI-VARIATE OPTIMIZATION

This project was performed under a very strict time schedule: each simulation run took 6 hours and results were needed within a few weeks for presentation to top management who had to decide if the project (in its current form) was to be continued. The theoretical approach of the preceding section requires specification of zmax, the commercially acceptable maximum value for the 90% quantile of promised lead times in order category 2. This value did not become available within the time constraints mentioned above. Therefore we modified the theoretical approach as follows.

We have available the results of the first local experiment; again see Table 2. So we can compute the steepest ascent path for y (productive hours) as required by the theoretical approach; see eq. (6.3). We decide to start our search along this path starting at the midpoint of the first experiment; see Q in Figure 1.[8] The (first) step size along this path, is always determined heuristically in RSM (see the comment below eq. 6.3). We may try to make this step size as big as seems "possible", which leads to the heuristic developed around eq. (6.4.). However, the latter heuristic requires quantification of zmax, which turned out to be impractical. Now we try a step size such that it is not as big as possible but it does change the control variables "sizably". We try the following two mutually related heuristics:

(i) Select a step size such that one control variable is (roughly) doubled while the other variables are less than doubled: see x_{12} in Table 3 (column 4).

(ii) Fix the control variable as in (i), but further increase the step size such that one other variable becomes (roughly) halved: see x_4 and x_5 in Table 3 (column 5).

Table 3 shows that the other 11 control variables do not change substantially, when applying the steepest ascent technique to the estimated response plane of the first local experiment. How do these heuristics affect the responses?

Upon applying heuristic (i), the productive hours indeed exceed the values in the first experiment except for 2 combinations, namely y_5 and y_{12}; also z_5, the lead time quantile for order category 2, is smaller. Heuristic (ii) gives better results: its productive hours y exceed the hours in the first experiment, except for y_{12}, but z_{17} is substantially smaller than z_{12} (y_{17} exceeds y_5 and z_{17} is only marginally larger than z_5; see heuristic i). Therefore we perform a second experiment around the setting of heuristic (ii) in Table 3. In other words, run 1 of experiment 2 is identical to run 17 of the total experiment. This second experiment again uses the first-order approximation of eq. (6.1) and hence the 2^{14-10} design of Table 1. Now row 1 of Table 1 corresponds to the base run of experiment 2, which is specified by the last column of Table 3 (heuristic ii). Again a minus sign in Table 1 (d_{ij} = -1) means that the corresponding control variable increases with 20%, for example, x_1 becomes 132 × 1.2 = 158.4.

Table 3: Control variables x_j in base run (x_j^b) and along steepest ascent path.

Control x_j j	Effect $\hat{\beta}_j$	Value of control variable j in Base run	Heuristic (i)	Heuristic (ii)
1	0.52	132	132.0003	132.0001
2	-39.30	3.3	3.28035	3.2214
3	0.65	132	132.0003	132.0001
4	-18.07	0.075	0.065965	0.03886
5	-128.96	0.55	0.48552	0.29208
6	0.00	316.8	316.8	316.8
7	-0.22	660	659.9999	659.996
8	13.88	1.65	1.65694	1.6778
9	-1.53	27.5	27.44992	27.4469
10	1.39	110	110.0007	110.003
11	0.03	330	330	330
12	527.23	0.33	0.5936	0.6
13	-9.27	5.5	5.4954	5.4815
14	-0.46	132	131.9998	131.9991

The second experiment yields the following results.

(i) The second experiment is performed in the neighborhood of the new base run (see point P in Figure 1) so that some y-values are higher than y_{17} (namely runs 26, 30, 31) (9 z-values are smaller than z_{17}).

(ii) Compared to the base run of experiment 1 (the initial common sense setting) only 3 out of 16 y-values are <u>not</u> higher, namely y_{27}, y_{28} and y_{32}. Though the steepest ascent path does not increase y for these three combinations, it does happen to decrease z (z_{27}, z_{28} and z_{32} are smaller than z_1). So RSM does yield improval control settings; also see the results (iii) through (v).

(iii) The <u>maximum</u> y-value is y_{26} (run 10 of experiment 2); y_{26} is 2.6% higher than y_1. And z_{26} happens to be only 0.4% higher than z_1 so that, if z_1 is acceptable, then z_{26} is too.

(iv) Other combinations improve y less but they do improve z drastically. For example, run 15 of experiment 2 improves the base run's y_1 by 1.9% while z decreases by 12.5%.
(v) Run 5 of experiment 2 gives the minimum lead time quantile: z_1 is reduced by 62.8%. And y_{21} is still 1.3% higher than y_1.

The improvements of y in the second experiment are smaller than we had hoped for. Several explanations are possible. Maybe RSM is not an effective optimization technique for this case study (local hills?); also see Kim and Blake (1987). Maybe the intuitively selected setting for the control variables x_j is close to the optimum. The intuitive setting, however, does not give good delivery times; for example, run 5 of experiment 2 decreases z_1 by 62.8% (while its y is still 1.3% higher than y_1; see result v above). And it was the delivery times that initiated this DSS (see § 1). We might explore the dual problem formulation, namely, minimize the lead time quantile z while keeping productive hours y at y_1 or, better, while keeping y at its historical value. We might also compute the new local estimates $\hat{\beta}$ and $\hat{\gamma}$ for y and z respectively, and continue searching in a third experiment. Unfortunately, these steps were not realized, because the project was aborted, mainly because of lack of personnel needed to develop and implement the DSS.

8. EPILOGUE: SIMULATION METHODOLGY

Our approach emphasized the importance of obtaining historical data on lead times in order to evaluate the simulation output z, the 90% quantile of promised lead times for order category 2. Upon studying these historical data, some people in the organization concluded that lead times realized by the person responsible for production scheduling, are better than the lead times realized by the model! This conclusion, however, is based on the simulation originally followed by the OR team; this approach was called shadow or parallel running (a term often used in the information systems field).

Their simulation model represents the "factory" (6 machines) and the DSS which use historical orders as input. The output consists of lead times, idle times, switchover times, and so on. In the preceding paragraph this output was compared to the historical output of the human planner. But this comparison, is unfair, in our opinion! For example, in practice the production capacity is higher and more flexible than it is in the simulation model; hence the human planner can realize better lead times. Actually there are a number of complications in practice which are not accounted for in the model; of course the human planner did respond to these complications in reality. Therefore a fair comparison of the model and the human planner requires a different simulation approach, namely the following approach (which we think is standard).

The simulation model still represents the factory, and one variant represents the heuristic production planning system, as above. The second variant, however, represents the human planner! This new model variant can indeed be built, if it is possible to make the human decision rules explicit. (These rules may be represented by a few lines of code or by a complete Expert System; see IntelliCorp, 1986.) If the human decision-making process can not be formalized, then a gaming variant can be built, i.e., the human planner has to make decisions in a simulated factory. This approach yields fair comparisons, whereas the preceding approach does not!

We observe that the simulation was fed with historical orders. This is an accepted methodology for validating a simulation model. So in the second variant (presented in the preceding paragraph) the simulation model is fed with historical input, and gives simulated output which can be compared to the historical output, in order to check whether the model of the factory and the human planner is realistic. After validation and optimization, the sensitivity analysis should concentrate on changes in the order stream and in the factory, in order to check the robustness of the model versus the human planner; for example, can the model cope with a labor strike (the simulation model already includes historical machine breakdowns)?

This paper concentrates on optimizing the heuristic production planning system. The original idea, however, was to use this system as a

Decision Support System. In other words, the human planner does not compete with a model but is assisted by a model, which in this case comprises a heuristic module and a simulation module for what-if questions. So the original project looked like this:

(i) Develop a heuristic production planning module.

(ii) Evaluate and "optimize" (improve) this heuristic module (this is the topic of our paper).

(iii) Let the human plannner do his job assisted by the "optimized" production planning system: the computer generates many more alternative plans than the human expert can contemplate in the time available for planning; the DSS can also screen-out alternatives which are clearly inferior. (Remember that our optimization considered only 2 responses.) Is the performance of this interactive system "better" than the performance of the human planner alone?

As we mentioned before, the project was aborted before the end of step (ii).

9. CONCLUSIONS

At the outset of this case study, we had a DSS with 15 control variables and a great many, namely 34, response variables. We reformulated the problem such that only 2 response variables remained, namely y, the number of productive hours (which excludes idle times and switchover times), and z, the 90% quantile for promised lead times of order category 2. We wished to maximize y since it directly affects profits, and originally we wished to keep z below some commercially acceptable limit, zmax. Unfortunately, in the few weeks of this project we could not obtain a "hard" value for zmax. Nevertheless we could proceed as follows.

The 15 control variables could be reduced to 14. At the outset of the project these 14 variables x_j had intuitively selected base values x_j^b $(j = 1,...,14)$. Our first experiment investigated the 14 control variables in only 16 runs (a 2^{14-10} design), increasing each variable by 20%. This experiment showed that other settings of the control variables can indeed increase y and at the same time decrease z. Moreover, this experiment gave the estimated, local steepest-ascent path.

Next we heuristically selected a step size along this steepest ascent path. In that neighborhood we performed a second 2^{14-10} experiment, again changing each control variable by 20%. Several combinations in the second experiment were better than the initial base run, that is, y_1 was smaller and z_1 was higher. The maximum increase in y was 2.6% while the corresponding z remained virtually equal to z_1. For practical reasons we could not continue our (steepest ascent) search for better control settings; neither could we implement our (suboptimal) solution. Nevertheless this paper demonstrates statistical design and analysis techniques which are standard for statistical experts but not for Operations Researchers. This statistical methodology is simple and effective, i.e., it leads to control settings better than the intuitively selected (base) setting. We also indicated an extension of RSM methodology that seems novel (see § 6). Our case study illustrates practical problems such as lack of data (see the variables zmax and profit contribution), time pressure, and organizational politics.

REFERENCES

BETTONVIL, B. and J.P.C. KLEIJNEN, Measurement scales and resolution IV designs. Katholieke Universiteit Brabant, Tilburg (Neth.), January 1988 (submitted for publication).

BILES, W.E., A response surface method for experimental optimization of multi-response processes. INDUSTRIAL AND ENGINEERING CHEMISTRY, PROCESS DESIGN AND DEVELOPMENT, 14, no. 2, 1975, pp. 152-158.

HOERL, R.W., Ridge analysis 25 years later. The American Statistician, 39, no. 3, August 1985, pp. 186-192.

INTELLICORP, The Simkit system; knowledge-based simulation tools in KEE. User Manual, KEE Version 3.0, Mountain View (California 94040-2216), Sept. 1986.

KHURI, A.I. and M. CONLON, Simultaneous optimization of multiple responses represented by polynomial regression functions, Technometrics, 23, no. 4, Nov. 1981, pp. 363-375.

KIM, C. and H. BLAKE, Optimizing simulation experiments. Dept. of Decision and Information Sciences, University of Santa Clara, Santa Clara (CA. 95053), 1987. (Submitted to: Modelling and Simulation.)

KLEIJNEN, J.P.C., Statistical techniques in simulation, volumes I and II. Marcel Dekker, Inc., New York, 1974/1975. (Russian translation: Publishing House "Statistics", Moscow, 1978.)

KLEIJNEN, J.P.C., Statistical tools for simulation practitioners. Marcel Dekker, Inc., New York, 1987.

MILLER, G.A., The magical number seven plus or minus two: some limits on our capacity for processing information. The Psychological Review, 63, March 1956, pp. 81-97.

MYERS, R.H. and W.H. CARTER, Response surface techniques for dual response systems. TECHNOMETRICS, 15, no. 2, May 1973, pp. 301-317.

ACKNOWLEDGMENT

I benefitted from the discussions with several company employees and with B. Bettonvil (Katholieke Universiteit Brabant/Technische Universiteit Eindhoven) and S. Geldof (TNO). An anonymous referee's comments lead to a drastic improvement in the organization of this paper.

Notes

1) We give the following rough idea of the DSS, developed by the OR team. Each of the 6 machines has a queue of assigned jobs. That queue first has jobs of one type, say type a; next jobs of type b, and so on. Between these subqueues (corresponding to jobs of type a, b, ...) there are open slots (reserve, slack) to accomodate newly arriving jobs of type a, b, ... Moreover, not all jobs are assigned to specific machines, i.e., some jobs are placed in a seventh queue (slack queue). The assignment of a specific job to a queue depends on the DSS parameters x_j ($j = 1,...,15$).

2) Actually we should not only optimize the control variables, but we should also investigate the sensitivity of the optimal solution to variations in the environmental variables such as factors determining the orders (also see Kleijnen, 1987, p. 216). We shall return to this issue in § 9.

3) More accurate estimators of β and γ are possible using Weighted Least Squares, assuming that different control settings yield different variability of y so that $\sigma_i^2 \neq \sigma^2$. But then, we have to estimate these σ_i^2. See Kleijnen (1987, pp. 161-169).

4) Myers and Carter (1973) consider a problem quite analogous to our's, namely maximize the primary response (say y) subject to a restriction on the secondary response (say z); however, their methods (based on Lagrangian multipliers) "are applicable only to quadratic response functions". Biles (1975, p. 155) uses "Rosen's gradient-projection" technique in case one or more constraints are violated.

5) Our approach ignores the random character of the estimated factor effects. A better formulation is: $P(\hat{z} = \hat{\gamma}_0 + \sum_1^{14} \hat{\gamma}_j x_j \leq zmax) \geq P^*$ with prespecified P^*, say $P^* = 0.90$. Then eq. (6.2) - also see Figure 1 - becomes

$$\hat{z} = \hat{\gamma}_0 + \sum_{1}^{14} \hat{\gamma}_j x_j \le zmax - t_v^{\alpha} \hat{\sigma}_z \qquad (6.2')$$

where $v = 16-15$ and $\sigma_z^2 = \mathbf{x}'\Omega_\gamma \mathbf{x}$ where Ω_γ is the covariance matrix of $\hat{\gamma}$ and $\mathbf{x}$ is the vector of control variables including the dummy variable $x_0 = 1$. Also see Khuri and Colon (1981). The practical approach of § 7 does not use eq. (6.2) or its variant (6.2').

6) Our proposal is devised especially for this case study; the general problem of testing regression model adequacy is discussed in Kleijnen (1987, pp. 185-196).

7) If the two fitted equations held over the whole area of interest ($0 \le x_j < \infty$), then the optimum would correspond to the point where $\hat{z} = \hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 x_2 = zmax$ intersects the x_1 axis (assuming $-\hat{\beta}_1/\hat{\beta}_2 > -\hat{\gamma}_1/\hat{\gamma}_2$ as is illustrated), which is well-known from Linear Programming.

8) We could have started the steepest ascent path at a corner of the first local experimental area; see C in Figure 1. Actually, the second local experiment comprises 16 runs (as we shall see), and it does not seem to matter what the exact position is of the second local experimental area, when using RSM.

***

IN 1987 REEDS VERSCHENEN

242 Gerard van den Berg
Nonstationarity in job search theory

243 Annie Cuyt, Brigitte Verdonk
Block-tridiagonal linear systems and branched continued fractions

244 J.C. de Vos, W. Vervaat
Local Times of Bernoulli Walk

245 Arie Kapteyn, Peter Kooreman, Rob Willemse
Some methodological issues in the implementation of subjective poverty definitions

246 J.P.C. Kleijnen, J. Kriens, M.C.H.M. Lafleur, J.H.F. Pardoel
Sampling for Quality Inspection and Correction: AOQL Performance Criteria

247 D.B.J. Schouten
Algemene theorie van de internationale conjuncturele en strukturele afhankelijkheden

248 F.C. Bussemaker, W.H. Haemers, J.J. Seidel, E. Spence
On (v,k,λ) graphs and designs with trivial automorphism group

249 Peter M. Kort
The Influence of a Stochastic Environment on the Firm's Optimal Dynamic Investment Policy

250 R.H.J.M. Gradus
Preliminary version
The reaction of the firm on governmental policy: a game-theoretical approach

251 J.G. de Gooijer, R.M.J. Heuts
Higher order moments of bilinear time series processes with symmetrically distributed errors

252 P.H. Stevers, P.A.M. Versteijne
Evaluatie van marketing-activiteiten

253 H.P.A. Mulders, A.J. van Reeken
DATAAL - een hulpmiddel voor onderhoud van gegevensverzamelingen

254 P. Kooreman, A. Kapteyn
On the identifiability of household production functions with joint products: A comment

255 B. van Riel
Was er een profit-squeeze in de Nederlandse industrie?

256 R.P. Gilles
Economies with coalitional structures and core-like equilibrium concepts

257 P.H.M. Ruys, G. van der Laan
Computation of an industrial equilibrium

258 W.H. Haemers, A.E. Brouwer
Association schemes

259 G.J.M. van den Boom
Some modifications and applications of Rubinstein's perfect equilibrium model of bargaining

260 A.W.A. Boot, A.V. Thakor, G.F. Udell
Competition, Risk Neutrality and Loan Commitments

261 A.W.A. Boot, A.V. Thakor, G.F. Udell
Collateral and Borrower Risk

262 A. Kapteyn, I. Woittiez
Preference Interdependence and Habit Formation in Family Labor Supply

263 B. Bettonvil
A formal description of discrete event dynamic systems including perturbation analysis

264 Sylvester C.W. Eijffinger
A monthly model for the monetary policy in the Netherlands

265 F. van der Ploeg, A.J. de Zeeuw
Conflict over arms accumulation in market and command economies

266 F. van der Ploeg, A.J. de Zeeuw
Perfect equilibrium in a model of competitive arms accumulation

267 Aart de Zeeuw
Inflation and reputation: comment

268 A.J. de Zeeuw, F. van der Ploeg
Difference games and policy evaluation: a conceptual framework

269 Frederick van der Ploeg
Rationing in open economy and dynamic macroeconomics: a survey

270 G. van der Laan and A.J.J. Talman
Computing economic equilibria by variable dimension algorithms: state of the art

271 C.A.J.M. Dirven and A.J.J. Talman
A simplicial algorithm for finding equilibria in economies with linear production technologies

272 Th.E. Nijman and F.C. Palm
Consistent estimation of regression models with incompletely observed exogenous variables

273 Th.E. Nijman and F.C. Palm
Predictive accuracy gain from disaggregate sampling in arima - models

274 Raymond H.J.M. Gradus
The net present value of governmental policy: a possible way to find the Stackelberg solutions

275 Jack P.C. Kleijnen
A DSS for production planning: a case study including simulation and optimization

276 A.M.H. Gerards
A short proof of Tutte's characterization of totally unimodular matrices

277 Th. van de Klundert and F. van der Ploeg
Wage rigidity and capital mobility in an optimizing model of a small open economy

278 Peter M. Kort
The net present value in dynamic models of the firm

279 Th. van de Klundert
A Macroeconomic Two-Country Model with Price-Discriminating Monopolists

280 Arnoud Boot and Anjan V. Thakor
Dynamic equilibrium in a competitive credit market: intertemporal contracting as insurance against rationing

281 Arnoud Boot and Anjan V. Thakor
Appendix: "Dynamic equilibrium in a competitive credit market: intertemporal contracting as insurance against rationing

282 Arnoud Boot, Anjan V. Thakor and Gregory F. Udell
Credible commitments, contract enforcement problems and banks: intermediation as credibility assurance

283 Eduard Ponds
Wage bargaining and business cycles a Goodwin-Nash model

284 Prof.Dr. hab. Stefan Mynarski
The mechanism of restoring equilibrium and stability in polish market

285 P. Meulendijks
An exercise in welfare economics (II)

286 S. Jørgensen, P.M. Kort, G.J.C.Th. van Schijndel
Optimal investment, financing and dividends: a Stackelberg differential game

287 E. Nijssen, W. Reijnders
Privatisering en commercialisering; een oriëntatie ten aanzien van verzelfstandiging

288 C.B. Mulder
Inefficiency of automatically linking unemployment benefits to private sector wage rates

IN 1988 REEDS VERSCHENEN

297 Bert Bettonvil
Factor screening by sequential bifurcation

298 Robert P. Gilles
On perfect competition in an economy with a coalitional structure

299 Willem Selen, Ruud M. Heuts
Capacitated Lot-Size Production Planning in Process Industry

300 J. Kriens, J.Th. van Lieshout
Notes on the Markowitz portfolio selection method

301 Bert Bettonvil, Jack P.C. Kleijnen
Measurement scales and resolution IV designs: a note

302 Theo Nijman, Marno Verbeek
Estimation of time dependent parameters in lineair models
using cross sections, panels or both

303 Raymond H.J.M. Gradus
A differential game between government and firms: a non-cooperative approach

304 Leo W.G. Strijbosch, Ronald J.M.M. Does
Comparison of bias-reducing methods for estimating the parameter in dilution series

305 Drs. W.J. Reijnders, Drs. W.F. Verstappen
Strategische bespiegelingen betreffende het Nederlandse kwaliteitsconcept

306 J.P.C. Kleijnen, J. Kriens, H. Timmermans and H. Van den Wildenberg
Regression sampling in statistical auditing

307 Isolde Woittiez, Arie Kapteyn
A Model of Job Choice, Labour Supply and Wages