Decision Support Systems 9 (1993) 269–280
North-Holland

Simulation and optimization in production planning

A case study

Jack P.C. Kleijnen

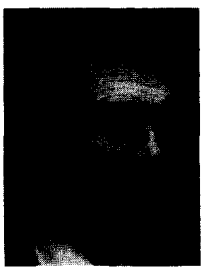
Catholic University Brabant, Tilburg, Netherlands

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 28 response variables were distinguished. These 28 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 products, on order. The major initial problem was the lead times: A drastic reduction (maybe 50%) seemed possible (in Section 7 we shall indeed realize a 37% reduction). First, the company investigated Material Requirements Planning (MRP-I) and Manufacturing Resource Planning (MRP-II), but it found this type of approach not suitable for its 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 would be too risky to



Jack P.C. Kleijnen is Professor of Simulation and Information Systems at the 'Katholieke Universiteit Brabant' (Tilburg University) in Tilburg, Netherlands. He received his doctoral and master's degrees in Management Science, in 1971 and 1964 respectively, at that same university. He spent several years in the U.S.A.: Rutgers University (Summer 1988), Pritsker & Associates (Summer 1984), IBM Research Yorktown (Summer 1981), Indiana University (Summer 1979), IBM Research San Jose (1974), Duke University (Summer 1968, Summer 1969) UCLA (1967/1968). He published four books and more than 100 articles in international journals on simulation, statistics, operations research, computer science, etc. His first book *Statistical Techniques in Simulation* (Dekker, N.Y., 1974/1975) received a Lanchester-Prize Honorable-Mention from the Operations Research Society of America, and was translated into Russian. He received a number of fellowships and awards, both nationally and internationally. He lectured at numerous conferences in many European countries and the USA, consulted several organizations, and is a member of many scientific committees, editorial boards, computer committees, and professional organizations. His research interests are in simulation, mathematical statistics, production management, and management information systems.

Correspondence to: Jack P.C. Kleijnen, Department of Information Systems and Auditing, School of Business and Economics, Catholic University Brabant, 1500 LE Tilburg, Netherlands.

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 production-planning module 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 (SPERRY 1100), 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; see also [5,6].

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 28 responses can be reduced to only two responses (see Section 4). The study also demonstrates the use of mathematical techniques, namely experimental design, regression analysis, and steepest ascent. These techniques are standard for the experts in the various fields; nevertheless, in practice operations researchers are often unfamiliar with techniques such as 2^{k-p} designs. Moreover, we add a novel idea to the steepest ascent technique for situations with multiple responses.

This paper is organized as follows. Section 2 describes the manufacturing process and the production-planning module of the DSS, emphasizing commercial and production goals. Section 3 presents an (inferior) one-factor-at-a-time design to select the simulation inputs, a large set of simulation responses, and the original regression model. Section 4 reduces the original 28 responses to only two responses; the production manager is interested in maximizing response 1 without violating an upper limit for response No. 2, a commercial variable. Section 5 uses a 2^{14-10} design for a local first-order model in the first stage of experimentation, which results in better combinations of DSS parameters and in estimated local first-order effects (which will guide the second stage of experimentation); no DSS parameters are eliminated at this stage! Section 6 applies the steepest ascent technique to the estimated local first-order model for response No. 1 (see Section 4), while considering the linear constraint for response No. 2 to determine the maximum step size along the steepest ascent path (selecting the step size in this way seems novel in Response Surface Methodology or RSM). Section 7 does not determine the maximum step size (since the linear constraint of Section 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 performance. Section 8 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. Section 9 summarizes the paper.

2. A production planning system

The DSS concentrates on the bottleneck process within the total production process (this bottleneck process consists of several consecutive subprocesses). The bottleneck production department has six machines available. To produce a specific product, a single machine suffices. A specific product can be made on more than one machine, but not on all machines. Though a specific machine can make different products (but not all products), the machine cannot make these different products with the same efficiency (also see Section 4). There are about 25 classes of products. 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 two to three hours no production is possible. Switchover costs within a class are minor, namely between five to sixty minutes, usually fifteen minutes. 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 $j = 1, \ldots, 15$. 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

PPS performance. For this paper, *the PPS is a black box.* We can indeed treat the PPS as a black box, since our methodology (2^{k-p} designs and Steepest Ascent) does not depend on specific knowledge about the PPS. (Of course, actual values resulting from the standardized design do depend on the specific system; see tables 2 and 3 later on.) Another reason for treating the PPS as a black box is the proprietary character of the system. Moreover, this paper would become too long, were the details of the PPS heuristics explained. We give the following rough idea of the PPS, developed by the OR team. Each of the six machines has a queue of assigned specific products (there are 700 products; two different customers may order the same product). That queue first has specific products a, next products b, and so on. Between these subqueues (corresponding to specific products a, b,...) there are open slots (reserve, slack) to accomodate newly arriving products a, b,... Moreover, not all products are assigned to specific machines, i.e., some products are placed in a seventh queue (slack queue). The assignment of a specific product to a queue depends on the PPS parameters x_j ($j = 1,\ldots,15$).

So from a *technological* viewpoint there are many different products (at least 700) which can be grouped into 25 'product' classes such that *switchover costs* are minor within a product class and major between classes. From a *commercial* viewpoint, however, 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 Section 8). The PPS was programmed and fed with four months of *historical data* (this period was thought to be representative; also see our comment in Section 8). For that period detailed data are available on orders (several thousands), changes in orders (30% of the orders are revised), machine breakdowns, and so on. By definition, one simulation run implies constant values for the 15 PPS parameters, during those four months.

The *original experimental design* for sensitivity analysis was to use the *one-factor-at-a-time* method:

Run 1. Fix the 15 PPS variables at their base values (say) x_j^b with $j = 1,\ldots,15$. (These base values were suggested by the developers of the PPS using 'common sense'. Common sense implies subjectivity so 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.)

Table 1
2^{14-10} experimental design $\boldsymbol{D} = (d_{ij})$. (+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	−	−	−	−	+	+	+	+	+	+	−	−	−	−

Run 2. Increase variable x_1 by 20% and keep all other 14 variables at their base values. This magnitude of change (20%) was selected rather arbitrarily. It is well-known that RSM does not specify the magnitude of changes. The PPS variables have not much intuitive meaning, so it is difficult to specify a 'high' value. We shall return to this issue in Section 5 (see the discussion of table 2).

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

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

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 [6] and table 1. Moreover, *optimization takes several rounds of experimentation and analysis,* as we shall see in later sections, so *efficient* designs become even more desirable.

The original idea was to *evaluate* each simulation run using the following *28 aspects:*

(a) Average and spread of promised lead times, for orders in classes 1, 2 and 3;
(b) Average and spread in lead time inaccuracy (= absolute value of realized lead time minus promised lead time) for orders in all five classes;
(c) Utilization degree = production hours/(production hours + idle time + switchover time) $\times$ 100%, for each of the six machines;
(d) Switchover degree = switchover time/(production hours + idle time + switchover time) $\times$ 100%, for each machine.

For each aspect (say) y, the OR team wanted to fit a *regression* model. They assumed that the following first-order approximation would be adequate in the first stage of the investigation (where the variables x_j are changed by 20%; also see the last paragraph of Section 6).

$$y_i = \beta_0 + \sum_{j=1}^{15} \beta_j x_{ij} + e_i \quad i = 1, \ldots, 31, \tag{3.1}$$

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

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

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

4. Reconsidering the problem

The preceding section listed 28 aspects thought to be relevant for the evaluation of the PPS. Obviously, managers cannot select a system by considering as many as 28 aspects. (Miller [7] wrote a famous article on this). Therefore we proposed to the client to *reconsider the original problem formulation,* and try to 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 contribution 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. The simulation program can keep track of the number of product units produced during the simulated period. Upon multiplying these numbers by the gross-profit margin per product unit we get the total profit contribution (say) $\tilde{y}$. In this way utilization and switchover degrees for each of the six machines (together $2 \times 6 = 12$ variables) can be combined into a single variable per simulation run, namely *profit contribution.* (As we shall see in Section 5, these accounting data did not become available within the strict time constraint of this project, so we switch to a closely related variable 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; for example, 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 a management's job to specify a maximum for acceptable lead times. (Analogy: Inventory 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. By definition, however, lead times are not critical for class-4 and -5 orders. 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 items account for the 'major' part, namely 80%, of the sales volume.)

In order to further reduce the number of aspects, 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 PPS; 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 *quantiles,* i.e., we use the 90% quantile (say) $z_{.9}$.

$$P(l \leqslant z_{.9}) = 0.90. \tag{4.1}$$

To estimate $z_{.9}$ 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 [6,p.82] for a detailed discussion. Actually we do not need $\hat{\text{var}}(z)$, as we shall see later.

Note that the selection of the 90% instead of the 95% or 99% quantiles, is quite arbitrary.

In summary, we succeeded in reducing the original 28 evaluation aspects of the PPS to only two variables. One variable $\tilde{y}$ is the total profit contribution by the six machines, and should be maximized. The other variable $z_{.9}$ is the 90% quantile of promised (approximately equal to realized) lead times of class-2 orders (the most important order class). So the production manager should try to maximize $\tilde{y}$ 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 PPS.

5. Basic experimental design and results

At the outset the OR team considered 15 PPS parameters or variables which were to be investigated in a first experiment of 31 simulation runs (see Section 3). Note that experimental design theory speaks of "factors" instead of "parameters" or "variables". (Actually we should not only optimize the PPS variables, but we should also investigate the sensitivity of the optimal solution to variations in the environmental variables such as factors determining the orders. We shall return to this issue in Section 9; see also [6,p.216].)

Upon closer examination we find that two of the 15 factors can be combined into a single factor (we do not explain this detail, since we would have to explain the PPS heuristics; see Section 2). To optimize the remaining 14 variables x_j we apply *Response Surface Methodology* (RSM). So we start with a *local first-order* approximation (see also (3.1))

$$\tilde{y} = \beta_0 + \sum_{1}^{14} \beta_j x_j + e. \tag{5.1}$$

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. A one-factor-at-a-time design

with 31 runs was discussed in Section 3. Actually the 15 regression parameters β in (5.1) can be estimated without bias, using a classical 2^{14-10} *design,* which takes only 16 runs (and a single run requires six hours of computer time, so 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, $\text{var}(\hat{\beta}_j)$ would be minimal. We do not use this particular error specification in our analysis. See also [6,pp.334–337].) The design matrix $\boldsymbol{D}$ 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 $5 = 1 \cdot 2$, $6 = 1 \cdot 3, \ldots, 14 = 2 \cdot 3 \cdot 4$ which are also listed in table 1).

To obtain the matrix of independent variables $\boldsymbol{X}$ corresponding to (5.1), we arbitrarily associate the *levels* $+1$ and -1 of $\boldsymbol{D}$ in table 1 with the actual "low" and "high" values of the PPS variables, i.e., in the first local experiment (comprising 16 runs; more experiments will follow) $+1$ in table 1 corresponds to the base values (specified using 'common sense'; see Section 3) and -1 corresponds to 20% higher values. Finally $\boldsymbol{D}$ is augmented with a column of 16 one's corresponding to β_0. The OLS estimator is

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

where the vector $\boldsymbol{y}$ equals $(y_1, \ldots, y_i, \ldots, y_{16})'$ and y_i denotes the total number of *productive hours* of the six machines in run i. In the preceding section we introduced the "profit contribution" $\tilde{y}$. However, it turned out to be impossible to obtain the necessary 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 $\tilde{y}$ also accounts for different technical and financial contributions per machine.

We also measure z_i, the 90% quantile of promised lead times for class-2 orders in run i, and we estimate $\boldsymbol{\gamma}$, the effects of the PPS variables on z:

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

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 combinations of the PPS variables x_j (see fig. 1 later on). As we move in stages through the experimental area, the local first-order effects change. We do not eliminate factors, because a factor that is 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 $\hat{\text{var}}(\hat{\beta}_j)$ and $\hat{\text{var}}(\hat{\gamma}_j)$ and the error specification of (3.2).)

Note that reducing the number of factors from 15 to 14 does not decrease the required number

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

PPS variable x_j	Effect on productive hours y		Effect on 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

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 [6,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 $\text{var}(e) = \sigma^2$ in (3.2), as we explained at the end of the preceding paragraph.

For confidentiality reasons we do not display the values of the simulation responses y_i and z_i ($i = 1, \ldots, 16$). However, we do give some comments on these values, and we do display the changes in y_i and z_i caused by changes in the PPS 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, which was the common sense combination of the PPS variables. Other combinations yield more productive hours, and at the same time result in lower lead times. 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 combinations which *dominate the base combination.*

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

(iii) To evaluate the effect of PPS variable j we should consider, not the unit effects $\hat{\beta}_j$ 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 variable j; see the third column in table 3). The reason is that the variables have different scales and ranges; see also [1].

6. Multi-variate optimization: Theory

Optimization of simulated systems is complicated, since there is no standard mathematical technique to optimize a non-linear, possibly stochastic, system with multiple responses. Kleijnen [6,pp.202–206] surveys different techniques, such as RSM and coordinate search, and complications due to side conditions and multiple responses. Hoerl [4,p.190] states "...multiple responses... is basically an unsolved problem..."; see also [2;3,pp.373–375;8]. In the present case study with its time constraints we needed a fast solution, and we developed the following approach which turns out to work (see the results at the end of Section 7).

Section 4 showed that we wish to optimize $\tilde{y}$, the total profit contribution by the six machines, under the restriction that $z_{.9}$, the 90% quantile of promised lead times for class-2 orders, does not

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

Variable x_j j	Effect $\hat{\beta}_j$	Value of PPS variable x_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

exceed a prespecified limit (say) z_{max}. For practical reasons, we introduced y, the total number of productive hours; see Section 5. Quantifying the commercial limit z_{max} is more difficult. The idea is that a higher z_{max} results in a higher maximum for y, and that—at the end of our investigation—management selects an attractive combination of z_{max} and max (y). (Analogy: In practical inventory control, management selects a combination of service percentage and inventory investment.)

The *mathematical problem* becomes (see also fig. 1 later on)

$$\text{maximize} \quad \hat{y} = \hat{\beta}_0 + \sum_{1}^{14} \hat{\beta}_j x_j \tag{6.1}$$

$$\text{subject to} \quad \hat{z} = \hat{\gamma}_0 + \sum_{1}^{14} \hat{\gamma}_j x_j \leqslant z_{max}. \tag{6.2}$$

Because (6.1) and (6.2) are fitted only locally, we know that these two equations do not hold over the whole area of interest. Therefore it makes no sense to apply Linear Programming to (6.1) and (6.2). Instead, we proceed as follows.

The sign of $\hat{\beta}_j$ shows whether x_j should be increased or decreased in order to maximize y. The relative changes in x_j should follow the path of *steepest ascent.*

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

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

(i) Does y indeed increase?

(ii) Does z indeed remain below zmax?

Fig. 1 illustrates the situation for only two PPS variables, x_1 and x_2. We emphasize that the steepest ascent path is based on local and estimated values $\hat{\beta}_j$. In fig. 1 the local experiment in the first stage is the subdomain represented by the rectangle *ABCD*. An 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

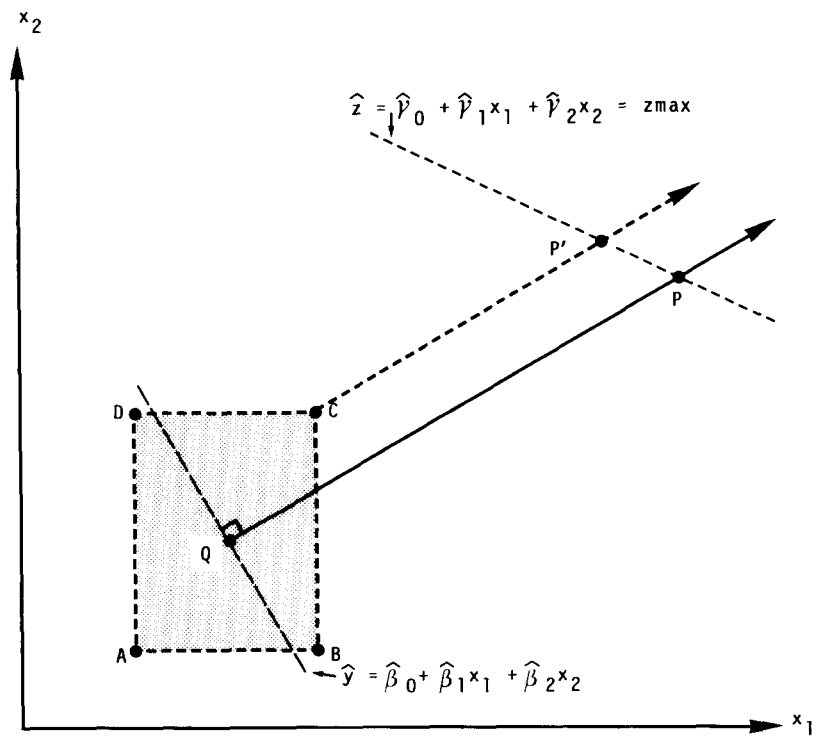


Fig. 1. Steepest ascent path with one restriction.

illustration implies that the condition $z = \hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 a_2 \leqslant z_{\max}$ 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. Actually Q, the starting point of the steepest ascent path, is selected arbitrarily. So several parallel paths could have been drawn in fig. 1; for example, if C shows the highest production, then it seems better to start from C, which leads to P'. The difference between P and P', however, is not really important, because both P and P' are 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', which is not shown in fig. 1.

The second experiment should reveal whether indeed the simulation response y increases (question (i) above) and whether $z \leqslant z_{\max}$ (question (ii)). We can start this second local experiment with a first run in which the PPS variables are fixed to the values corresponding to P. The following situations are possible, where the first experiment comprised n runs ($n = 4$ in fig. 1, but $n = 2^{14-10}$ in table 1) and $n + 1$ corresponds to P.

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

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 $z_{n+1} > z_{\max}$.

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, since we expect to have 'overshot' our goal, i.e., we assume that the steepest ascent path shows the right direction but we have taken too big a step.

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

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 $z_{\max}$, 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 (so S_0 is the distance between P and Q), and defining z_Q to be the z value corresponding to Q.

$$\frac{S_n}{S_0 - S_n} = \frac{z_{\max} - z_Q}{z_{n+1} - z_{\max}}. \tag{6.4}$$

(iv) $y_{n+1} < \max\ y_i$ and $z_{n+1} \leqslant z_{\max}$.

We may proceed as in situation (ii).

Note that as we move into the optimal area, the first-order approximation of (5.1) or (6.1) becomes less adequate so we have to switch to a second-order approximation. This fine-tuning requires the estimation of 91 interactions $\beta_{jj'}$ ($j' = 2, \ldots, 14$ and $j' > j$) and 14 purely quadratic effects β_{jj}. Before this fine-tuning we may eliminate non-significant variables, in order to save computer time. Actually we never got to this stage, as we shall see. See also [6,pp.202–208,312–316].

7. Practical multi-variate optimization

This project was performed under a very strict time schedule: Each simulation run took six hours and results were needed within a few weeks for presentation to top management who had to decide if the project was to be continued. The theoretical approach of the preceding section requires specification of $z_{\max}$, the commercially acceptable maximum value for the 90% quantile of promised lead times for class-2 order. 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 (6.3). We decide to start our search along this path, starting at the *midpoint* of the first experiment; see Q in fig. 1. (We could have started the steepest ascent path at a corner of the first local experimental area; see C in fig. 1. Actually, the second local experiment comprises 16 runs, and it does not seem to matter what the exact position is of the second local experimental area, when using RSM.) The *step size* along this path, is always determined heuristically in RSM, as we saw below (6.3). We may try to make this step

size as big as seems 'possible', which leads to the heuristic developed around (6.4.). However, the latter heuristic requires quantification of z_{max}, 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 PPS variables 'sizably'. We try the following two mutually related *step size heuristics:*

(i) Select a step size such that at least one PPS variable is roughly doubled (or halved), while the other variables are less than doubled (or halved): See x_{12} in table 3 (columns 3 and 4).

(ii) Further increase the step size such that at least one other variable becomes roughly halved: See x_4 and x_5 in table 3, columns 3 and 5. The variable x_{12} is kept at roughly the same value as in heuristic (i), because of the specific interpretation that variable has; so we deviate from the steepest ascent path. The heuristic (i) and (ii) combined with the steepest ascent equation (6.3) imply

$$x_j^{(i)} = x_j^{(1)} + 0.0005\hat{\beta}_j \quad j = 1, \ldots, 14, \qquad (7.1.a)$$

$$x_j^{(ii)} = x_j^{(1)} + 0.002\hat{\beta}_j \quad j = 1, \ldots, 11, 13, 14, \qquad (7.1.b)$$

where $x_j^{(i)}$ and $x_j^{(ii)}$ denote the x_j-value according to heuristic (i) and (ii), respectively, and $x_j^{(1)}$ denotes the x_j-value in base run 1. Table 3 shows that the other 11 variables do not change substantially, when we apply 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 y indeed exceed the values in the first experiment except for two combinations (namely y_5 and y_{12}; also z_5, the lead time quantile for class-2 orders, is smaller). Heuristic (ii) gives even better results: Its productive hours y_{17} exceed the hours in the first experiment except for one combination (namely 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)). When we further explore the neighborhood of the estimated steepest-ascent path (using a third simulation run), the results become bad: y decreases and z increases. 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 (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 PPS variable increases by 20%; for example, x_1 becomes $132 \times 1.2 = 158.4$.

Table 4
Responses of second 2^{14-10} experiment. Relative to base run

Run i	$(y_i - y_1)/y_1 \times 100$	$(z_i - z_1)/z_1 \times 100$
17	1.68	−8.24
18	1.47	−36.19
19	1.45	−10.58
20	0.57	−10.43
21	1.25	−37.22
22	0.79	−6.64
23	0.50	−18.42
24	1.53	5.80
25	0.09	−17.15
26	2.58	0.36
27	−0.80	−22.77
28	−0.37	−1.75
29	1.48	−4.70
30	1.92	−7.09
31	1.93	−12.53
32	−0.78	−20.36

The second experiment yields the results of table 4, where for confidentiality reasons we do not display the y_i and z_i themselves but only the response increases relative to the base-run responses y_1 and z_1.

(i) The second 2^{14-10} experiment is performed in the neighborhood of the new base run (see point P in fig. 1); so some y-values are higher than y_{17} (namely runs 26, 30, 31) and some are not; also notice that 9 z-values are smaller than z_{17}.

(ii) Compared to the base run of experiment 1 (the initial common sense combination) only three out of 16 y-values are *not* 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 the corresponding z (z_{27}, z_{28} and z_{32} are smaller than z_1). So RSM does yield better combinations of the PPS variables; also see the results (iii)–(v).

(iii) The maximum y-value is y_{26}, which is 2.6% higher than y_1. And z_{26} happens to be equal to z_1.

(iv) Other combinations improve y only a little, but they improve z drastically. For example, run 31 improves the base run's y_1 by 1.9% while z decreases with 12.5%.

(v) Run 21 gives the minimum lead time quantile: z_1 is reduced by 37.2%. 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?). Maybe the intuitively selected combination for the PPS variables x_j is close to the optimum? The intuitive combination, however, does not give good delivery times; for example, run 21 decreases z_1 by 37% (while its y is still 1.3% higher than y_1; see result (v) above). And it was the delivery times that initiated this PPS (see Section 1). We cannot explain why z decreased so much, while y is the variable to be maximized in our RSM procedure. 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, including the PPS.

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 class-2 orders. 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), which we examine next.

The OR team's simulation model represents the 'factory' (six machines) and the PPS, 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 practical complications that 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 still represents the PPS, 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.) 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 simulation model of the factory and the human planner is realistic. After validation of that model and optimization of the PPS, the sensitivity analysis should concentrate on changes in the order stream and in the factory, in order to check the robustness of the PPS versus the human planner; for example, can be PPS cope with a labor strike (the simulation model already includes historical machine breakdowns)?

This paper concentrates on optimizing the heuristic production planning system (PPS). The original idea, however, was to use this system as part of 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: The PPS;
(ii) Evaluate and optimize this PPS (this is the topic of our paper);
(iii) Let the human planner be 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 that are clearly inferior. (Remember that our optimization considered only two 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 Production Planning System (PPS) with 15 control variables and as many as 28 response variables. We *reformulated the problem such that only two response variables remained:* y, the number of productive hours (which excludes idle times and switchover times), and z, the 90% quantile for promised lead times of class-2 orders. We wished to maximize y since it directly affects profits, and originally we wished to keep z below some commercially acceptable limit, $z_{\max}$. Unfortunately, in the few weeks of this project we could not obtain a 'hard' value for $z_{\max}$. 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 ($j = 1, \ldots, 14$). Our first experiment investigated the 14 PPS variables in only 16 runs (a 2^{14-10} design), increasing each variable by 20%. This experiment showed that other combinations of the 14 PPS 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 PPS variable by 20%. Several combinations in the second experiment were better than the initial base combination, that is, y_1 was smaller *and* z_1 was higher. The maximum increase in y was 2.6% while the corresponding z remained equal to z_1. For practical reasons we could not continue our (steepest ascent) search for better combinations of the 14 PPS variables; 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 combinations of the PPS variables that are better than the intuitively selected base combination. We also indicated an extension of RSM methodology that seems novel (see Section 6). Our case study illustrates *practical problems* such as lack of data (see the variables $z_{\max}$ and profit contribution $\tilde{y}$), time pressure, and organizational politics.

Acknowledgment

I benefitted from the discussions with several company employees and with B. Bettonvil (KUB/TUE) and S. Geldof (ITP-TUE/TNO).

References

[1] B. Bettonvil and J.P.C. Kleijnen, Measurement scales and resolution IV designs, American Journal of Mathematical and Management Sciences 10, nos. 3, 4 (1990) pp. 309–322.
[3] G.E.P. Box and N.R. Draper, Empirical model-building with response surfaces (Wiley, New York, 1987).
[4] R.W. Hoerl, Ridge analysis 25 years later, The American Statistician 39, no. 3 (August 1985) pp. 186–192.
[5] J.P.C. Kleijnen, Statistical techniques in simulation, volumes I and II. (Dekker, New York, 1974/1975). (Russian translation: Publishing House "Statistics", Moscow, 1978.)
[6] J.P.C. Kleijnen, Statistical tools for simulation practitioners (Dekker, New York, 1987).
[7] G.A. Miller, 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.
[8] R.H. Myers and W.H. Carter, Response surface techniques for dual response systems, Technometrics 15, no. 2 (May 1973) pp. 301–317.