European Journal of Operational Research 36 (1988) 167–173
North-Holland

Theory and Methodology

Optimizing the performance of a blood analyser: Application of the set partitioning problem

W.M. NAWIJN
Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands

Abstract: The paper presents some models for optimizing the production rate of an automatic multitest blood analyser. A blood analysis may require one or more tests. The performance of the analyser depends on the test compositions of the analyses and the design of the analyser itself. The design can be changed. The optimal design can be determined by solving a set partitioning problem, given a representative sample of blood analyses.

Keywords: Production, zero–one programming, set partitioning

Problem description

We consider the optimization of the production rate of an automated blood analyser for chemical blood testing. Usually several tests have to be performed on a single blood specimen. The blood analyser can perform these tests simultaneously. To execute the tests on a particular blood specimen, several cuvettes are automatically filled with blood from this specimen. Thereafter testing agents are added to the cuvettes and the tests start. Once a test is completed the cuvette is washed and ready for subsequent use. There are a fixed number of cuvettes equidistantly spaced on the circumference of a turntable. The two filling stations, for blood and agents respectively, the washing and testing stations are situated along the circumference. All operations take place simultaneously during a fixed time period in which the table is at rest. Subsequently, the table turns to the next position and all operations are repeated. The time between two subsequent stops is constant too, so all operations of the analyser proceed in fixed time steps. The occupancy time of a cuvette per blood specimen is fixed and the same for all tests.

However, the number of cuvettes needed for a particular specimen varies. Not only does it depends on the number of tests but it also depends on the particular test combination. That is, two samples with the same number of tests may need a different number of cuvettes, due to the design of the filling head for the testing agents. The openings of the supplying tubes in the filling head are arranged in a rectangular grid. To be specific, there are five clusters of four openings and the agent for each test is allocated to one of these twenty positions, see Figure 1. So, in our case the analyser can perform 20 tests.

Suppose we have a blood specimen that requires the tests 2, 3, 6 and 13. Now remember that the analyser proceeds in discrete time steps. First, it fills 4 cuvettes with blood from the patient's blood specimen. At the next stop two of these cuvettes are filled with agents for the tests 2 and 3, allocated in the first cluster of tests. At the same time 4 more cuvettes are filled with blood (from the same specimen). After one further time step

Received June 1987; revised October 1987

1	5	9	13	17
2	6	10	14	18
3	7	11	15	19
4	8	12	16	20

Figure 1. Layout of the filling head

one of the latter cuvettes is filled with the agent for test 6, belonging to the second cluster of 4 tests, and, simultaneously, a third group of 4 cuvettes is filled with blood. At the next stop one of these cuvettes is filled with the agent for test 13, pertaining to the fourth cluster and at the same time again 4 cuvettes are filled with blood from the next patient's specimen to be analysed, and so on. Consequently, our blood specimen requires three groups of 4 cuvettes, due to the allocation of the tests to the five clusters.

There are 192 cuvettes, i.e. 48 groups of 4 cuvettes. Since the time between two subsequent stops is 12 seconds (including the stop itself), continuous operation in the way described above is garanteed if the fixed occupancy time per group of cuvettes is less than 12×48 sec., which is indeed the case, since every test takes 7.5 min. When C denotes the average number of groups of cuvettes required per blood specimen the number of blood specimens processed per minute equals $5/C$.

The connections of the supplying tubes to the filling head can be changed, so the allocation of tests to the five clusters can be chosen. Since this takes quite some time our objective is to find an allocation of the twenty tests to the five clusters such that C is minimized in the long run.

Modelling the problem

Let n be the number of tests the analyser can perform. The n tests should be partitioned into k clusters of m tests, where $k \times m = n$. Let the stochastic n-vector $(Z_1, Z_2, \ldots, Z_n)$ denotes the test composition of an arbitrary blood specimen, in which $Z_j = 1$ if test j is required and $Z_j = 0$ if it is not.

Remark 1. The sample space of the n-vector $(Z_1, Z_2, \ldots, Z_n)$ should be defined over all demand sources, such as general physicians and the available medical specialists in the hospital. These sources may have different demand profiles with regard to test compositions. Therefore, the sample data of a day may exhibit nonhomogeneity due to administrative collection procedures.

Since a blood specimen requires at least one test, one obviously has

$$1 \leqslant \sum_{j=1}^{n} Z_j \leqslant n \tag{1}$$

so the variables Z_j $(j = 1, 2, \ldots, n)$ are dependent.

Let

$$x_{jr} = \begin{cases} 1 & \text{if test } j \text{ is assigned to cluster r,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Then the expected number of clusters (in fact m-tuples of cuvettes) used per specimen, is given by

$$E\{C\} = \sum_{r=1}^{k} E\left\{\min\left(1, \sum_{j=1}^{n} Z_j x_{jr}\right)\right\} \tag{3}$$

or equivalently,

$$E\{C\} = \sum_{r=1}^{k} \Pr\left\{\sum_{j=1}^{n} Z_j x_{jr} \geqslant 1\right\}. \tag{4}$$

Hence, the optimization problem to be solved reads

$$\text{minimize} \quad \sum_{r=1}^{k} \Pr\left\{\sum_{j=1}^{n} Z_j x_{jr} \geqslant 1\right\}, \tag{5.1}$$

subject to

$$\sum_{j=1}^{n} x_{jr} = m, \quad r = 1, 2, \ldots, k, \tag{5.2}$$

$$\sum_{r=1}^{k} x_{jr} = 1, \quad j = 1, 2, \ldots, n, \tag{5.3}$$

$$x_{jr} \in \{0, 1\}, \quad j = 1, 2, \ldots, n; \; r = 1, 2, \ldots, k. \tag{5.4}$$

Constraint (5.2) expresses the fact that every cluster of tests must contain exactly m tests, while constrain (5.3) specifies that each test should be allocated to exactly one cluster.

Remark 2. It is sufficient to define the stochastic vector $(Z_1, Z_2, \ldots, Z_n)$ over the sample space of blood specimens requiring two or more tests, since

the contribution of single test blood specimens to the objective function in (4.1) is a constant, as is readily verified. For the same reason one could further restrict the sample space to blood specimens requiring at most $m(k-1)$ tests.

Since

$$\Pr\left\{\sum_{j=1}^{n} Z_j x_{jr} \geqslant 1\right\} = 1 - \Pr\left\{\sum_{j=1}^{n} Z_j x_{jr} = 0\right\} \qquad (6)$$

it follows that

$$\sum_{r=1}^{k} \Pr\left\{\sum_{j=1}^{n} Z_j x_{jr} \geqslant 1\right\} = k - \sum_{r=1}^{k} \Pr\left\{\sum_{j=1}^{n} Z_j x_{jr} = 0\right\}. \qquad (7)$$

Hence minimizing the left hand side is equivalent to maximizing the sum on the right hand side.

Now suppose that the dependence between Z_j's is so weak that we approximately have

$$\Pr\left\{\sum_{j=1}^{n} Z_j x_{jr} = 0\right\} = \prod_{j=1}^{n} \Pr\{Z_j x_{jr} = 0\}. \qquad (8)$$

In this case we are faced with the problem of maximizing the expression in (8) subject to (5.2)–(5.4). The optimal solution then follows from the following lemma.

Lemma 1. *Let $a_1, a_2, \ldots, a_n$ be a sequence of (nonnegative) numbers and $(J_1, J_2, \ldots, J_k)$ be a k-partition of the index set $G = \{1, 2, \ldots, n\}$, such that $|J_r| = m$ $(k \times m = n)$. Moreover, let*

$$P(J_r) = \prod_{i \in J_r} a_i \quad \text{and} \quad S = \sum_{r=1}^{k} P(J_r).$$

Then S is maximized over all k-partitions of G if

$$\min\{a_i \mid a_i \in J_r\} \geqslant \max\{a_i \mid a_i \in J_{r+1}\} \qquad (C)$$

for $r = 1, 2, \ldots, k-1$.

Proof. Suppose that under condition (C) $a_i \in J_r$ and $a_j \in J_t$ with $1 \leqslant r \leqslant t \leqslant k$, such that $a_i > a_j > 0$. By interchanging a_i and a_j only the products $P(J_r)$ and $P(J_t)$ change and become $a_j P(J_r)/a_i$ and $a_i P(J_t)/a_j$, respectively. The new value of S, denoted by S', becomes

$$S' = S - (a_i - a_j)\left[P(J_t)/a_j + P(J_r)/a_i\right].$$

Hence $S' < S$. Since any partition other than (C) can be obtained from C by a sequence of such interchanges, condition (C) guarantees optimality. □

Hence, in case of 'nearly independent' Z_j's one should order the tests according to decreasing values of $\Pr\{Z_j = 0\}$ and allocate the first m tests from the ordering to the first cluster, the next m tests to the second cluster and so on.

It turns out, however, that in our case some tests are highly correlated, as will be shown later.

Remark 3. If the random variables $Z_1, \ldots, Z_n$ are symmetrically dependent, i.e. if all permutations of these variables have the same distribution, see Feller [1, p. 225], then (5.1) is independent of the x_{jr}'s and, consequently, the analyser's production rate is independent of the clustering.

Apart from this symmetric case almost nothing can be said about the optimal allocation for correlated pairs.

Consider the following special case. Suppose certain disjoint pairs of tests are correlated while test pairs are independent of each other and of the other tests. Moreover, let us assume that $m = 2$ and n is even. Although a number of necessary ordering conditions can be given for the optimal partition, no simple partition rule seems to exist. In fact this problem can be formulated as a weighted matching problem in a complete graph. The nodes of the graph correspond to the tests and the weight of the arc between two tests i and j either equals the product of $\Pr\{Z_i = 0\}$ and $\Pr\{Z_j = 0\}$ if they are independent, or equals $\Pr\{Z_i = 0, Z_j = 0\}$ if they are correlated. The algorithms to solve this problem are already quite complicated, see [3].

Obviously, in practice the simultaneous probability distribution of the Z_j's is unknown. So, the optimization problem must be reformulated in terms of estimates of the underlying probabilities, which are based on the test requirements of a sufficiently large number of blood specimens.

Let p be the sample size, i.e. the number of blood specimens, and let z_{ij} $(1 \leqslant i \leqslant p,\ i \leqslant j \leqslant n)$ be defined by

$$z_{ij} = \begin{cases} 1 & \text{if specimen } i \text{ requires test } j, \\ 0, & \text{otherwise.} \end{cases} \qquad (9)$$

Then in problem (5) the objective (5.1) should be replaced by

$$\text{minimize} \quad \frac{1}{p}\sum_{i=1}^{p}\sum_{r=1}^{k}\min\left(1, \sum_{j=1}^{n} z_{ij}x_{jr}\right) \tag{10}$$

which denotes the average number of clusters used per blood specimen, for the given sample. Notice that the objective function is nonlinear. It is possible however to formulate this problem as a linear 0–1 programming problem in the following way.

Let

$$y_{ir} = \max_{i \leqslant j \leqslant n} z_{ij}x_{jr}, \quad 1 \leqslant i \leqslant p,\ i \leqslant r \leqslant k. \tag{11}$$

Observe that

$$y_{ir} = \min\left(1, \sum_{j=1}^{n} z_{ij}x_{jr}\right) \tag{12}$$

since z_{ij} and x_{jr} are 0–1 variables.

Hence, in view of (11) and (12), problem (10) is equivalent to

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{p}\sum_{i=1}^{p}\sum_{r=1}^{k} y_{ir},\\
&\qquad y_{ir} \geqslant z_{ij}x_{jr}, \quad 1 \leqslant i \leqslant p;\ i \leqslant r \leqslant k;\ 1 \leqslant j \leqslant n,\\
&\sum_{j=1}^{n} x_{jr} = m, \qquad 1 \leqslant r \leqslant k,\\
&\sum_{r=1}^{k} x_{jr} = 1, \qquad 1 \leqslant j \leqslant n,\\
&x_{jr} \in \{0, 1\}.
\end{aligned} \tag{13}$$

The disadvantage of this formulation is that for a large sample size p, the number of inequality constraints become quite large. For our problem $k = 5$, $n = 20$ and $p = 2208$, which already gives more than 200 000 inequality constraints. Instead of solving the above problem, we choose to solve problem (10) by reformulating it as a set partitioning problem.

Let H_r be an m-subset of $\{1, 2, \ldots, n\}$, i.e. a cluster of m tests. Notice that there are $s = \binom{n}{m}$ such subsets. Associated with H_r we define the following 'cost' coefficient

$$C_r = \frac{1}{p}\sum_{i=1}^{p}\min\left(1, \sum_{j=1}^{n} z_{ij}I_j(H_r)\right) \tag{14}$$

in which the indicator function $I_j(H)$ is defined by $I_j(H) = 1$ if $j \in H$ and $I_j(H) = 0$ if $j \notin H$.

Observe that C_r equals the fraction of samples using cluster H_r. From the definition of $I_j(H_r)$ it follows that problem (10) is equivalent to the following well known set partitioning problem, see e.g. Syslo et al. [4],

$$\text{minimize} \quad \sum_{r=1}^{s} C_r x_r, \tag{15}$$

subject to

$$\begin{aligned}
&\sum_{r=1}^{s} I_j(H_r)x_r = 1, \quad 1 \leqslant j \leqslant n,\\
&x_r \in \{0, 1\}, \qquad 1 \leqslant r \leqslant s,
\end{aligned} \tag{16}$$

in which (16) expresses the requirement that each test j is contained in exactly one subset H_r.

The advantage of the above formulation over formulation (13) is that it is independent of the sample size, although at the cost of determining the C_r's for all H_r.

The solution

Since the maximum number of tests per blood specimen $n = 20$ and the tests should be partitioned into 5 clusters of 4 tests, the total number of 4-element subsets H_r in our case equals $s = \binom{20}{4} = 4845$. The solution to the set partitioning problem has been calculated using an available computer code, developed by Fleuren [2]. The underlying algorithm is based on a combination of techniques. Lagrangean relaxation and dual heuristics are used to obtain bounds for the objective function, which in turn are used to obtain the optimum solution by a Branch and Bound procedure.

The statistical data consists of 2208 analyses, containing at least two tests, which were collected over a period of approximately two weeks by tapping the data line from the central laboratory computer to the analyser.

As space limitations makes it impossible to present the statistical data, we confine ourselves to give a cross frequency table in which for all specimens requiring more than one test the frequency of simultaneous occurrence of a pair of tests is given. The $n \times n$ (symmetrical) matrix (f_{ij}), in which f_{ii} gives the number of specimens requiring test i and f_{ij} equals the number of specimens requiring test i and test j simultaneously, is pre-

Table 1
Cross frequencies and cross correlations for test pairs ($p = 2208$)

	1	2	3	4	5	6	7	8	9	11	12	13	14	15	16	18	20
1	869	0.47	0.47	0.38	0.40	0.37	0.11	0.11	0.7	−0.07	0.05	0.21	0.08	0.32	0.32	0.07	0.11
2	666	1054	0.94	0.31	0.27	0.36	0.09	0.09	0.14	0.01	0.05	0.16	0.01	0.35	0.35	0.05	−0.06
3	672	1021	1059	0.33	0.29	0.35	0.09	0.09	0.14	0.02	0.05	0.16	0.01	0.35	0.35	0.05	−0.06
4	208	207	213	220	0.01	0.09	0.29	0.32	0.11	−0.03	0.19	0.06	0.08	0.17	0.17	0.15	0.03
5	713	759	769	132	1277	0.51	−0.02	0.02	−0.03	0.11	0.04	0.17	0.06	0.25	0.25	0.05	0.11
6	693	796	796	154	1002	1260	0.08	0.04	0.03	0.06	0.01	0.13	−0.01	0.23	0.23	0.03	0.03
7	192	218	216	110	210	245	375	0.86	0.17	0.05	0.18	0.08	0.07	0.15	0.15	0.11	0.02
8	159	179	177	102	181	188	297	301	0.13	0.02	0.17	0.09	0.09	0.16	0.16	0.14	0.04
9	340	440	442	113	433	457	199	152	772	0.51	0.19	0.15	0.14	0.14	0.14	0.08	0.07
11	520	682	690	131	875	840	262	199	751	1418	0.16	0.14	0.09	−0.02	0.02	0.03	0.07
12	113	133	133	64	155	140	89	73	147	209	242	0.19	0.26	0.15	0.15	0.17	0.20
13	289	308	307	66	360	341	112	96	239	378	108	491	0.48	0.25	0.25	0.24	0.48
14	190	194	192	60	254	220	90	80	193	289	113	258	394	0.20	0.20	0.24	0.76
15	363	426	430	103	434	418	147	125	256	342	103	221	171	545	1.00	0.20	0.13
16	363	426	430	103	434	418	147	125	256	342	103	221	171	545	545	0.20	0.13
18	48	51	51	28	60	55	33	32	47	61	32	62	54	58	158	86	0.23
20	281	248	246	66	389	342	105	91	234	404	123	321	386	199	199	66	578

sented in Table 1 (restricted to its lower triangular part).

Notice from this table that the tests 10, 17 and 19 were never required and omitted.

From the matrix (f_{ij}) we can calculate the correlation coefficients ρ_{ij} between the tests i and j. Since Z_i and Z_j are binary variables, it follows that

$$\operatorname{cov}(Z_i, Z_j) = \Pr\{Z_i = 1,\ Z_j = 1\} - \Pr\{Z_i = 1\}\ \Pr\{Z_j = 1\}, \qquad (17)$$

$$\operatorname{var}(Z_i) = \Pr\{Z_i = 1\}\left[1 - \Pr\{Z_i = 1\}\right].$$

In view of these relations we estimate the correlation coefficients by

$$\rho_{ij} = \frac{pf_{ij} - f_{ii}f_{jj}}{\{f_{ii}f_{jj}(p - f_{ii})(p - f_{jj})\}^{1/2}}. \qquad (18)$$

The matrix (ρ_{ij}) is presented in Table 1 by its upper triangular part. From these data we see that for instance the test pairs (1, 2), (1, 3), (2, 3), (5, 6), (7, 8), (9, 11), (13, 14), (13, 20) and (14, 20) are, relatively speaking, highly correlated. The tests 15 and 16 always occurred simultaneously. These facts are reflected in the optimal solution found, which is presented in Table 2.

Observe that the tests 10, 17 and 19 which were never required and consequently have (almost) no relation with the other tests are allocated to one cluster, combined with test 18 which has the lowest frequency among all other tests. The optimal number of clusters used per (multi-test) blood specimen equals 2.22, see Table 2, taking into account $p = 2208$. It should be noted that the average number of tests per (multi-test) specimen equals 5.1, so the average number of clusters needed per specimen is at least 1 + entier (5.1/4) = 2.

The correlated pairs mentioned earlier can be used to generate a set of clusterings, the best of which provides an upperbound of the objective function in the set partitioning problem, which in turn can be used when determining the optimal solution. It should be noted however that the quality of the generated clusterings is rather variable. For instance for the clustering obtained by interchanging the pairs (9, 11) and (15, 16) in the optimal solution, viz. Table 2, the value of the objective function is 9.6% higher. The original clustering used by the hospital, depicted in fig. 1, in which the tests are grouped, as far as possible,

Table 2
The optimal clustering

	Tests	pC_r
Clusters	(1, 2, 3, 4)	1287
	(5, 6, 9, 11)	1915
	(7, 8, 15, 16)	775
	(12, 13, 14, 20)	843
	(10, 17, 18, 19)	86

according to physiological functions related to e.g. blood, liver, reins, etc., has an objective function value of 2.68, which is 20.4% higher than the optimal solution. The clustering obtained by ordering the tests according to Lemma 1, which can be obtained from the entries on the main diagonal in Table 1, gives an objective value that is 10.7% higher. Finally it should be mentioned that although the optimal solution was at first obtained for $p = 2208$, it turned out that for $p \geqslant 400$ we always found one and the same optimal solution. For different subsamples with $p < 400$ the optimal solutions differ and moreover, for a given sample several optimal solutions may exist. These observations relate to a statistical as well as an optimizational aspect of the sample size. The latter aspect plays a role in the sensitivity of the solution to changes in the coefficients C_r (c.q. the sample size) of the objective function (15). Unfortunately it is extremely difficult to say something about the (in)sensitivity on p. From a statistical point of view one may ask for the minimal sample size such that the true value of the objective function will be estimated with a relative error of 5%, say, with a confidence level of 99%. To give a crude answer to this question observe that in our case (15) is the sum of five estimated probabilities. Let $Q = \Pr\{Z_{i_1} + Z_{i_2} + Z_{i_3} + Z_{i_4} \geqslant 1\}$ for some set $H_r = \{i_1, i_2, i_3, i_4\}$, then it is known that the estimator $\hat{C}_r$ for this probability is normally distributed with mean Q and variance $Q(1 - Q)/p$ for p large enough, provided all the analyses are independent. Now the values of Q for the five contributing sets range somewhere between 0.9 and 0.05, as can be verified from some trial solutions, and their sum lies between 2.0 and 2.5. Assuming that the estimators are independent random variables their sum has a total variance of approximately $0.8/p$. For a confidence level of 99% and a relative error of 5% we should have

$$2.33 * \sqrt{\frac{0.8}{p}} < 0.05 * 2.00$$

which gives $p > 434$. For a relative error of 2% and the same confidence level we obtain $p > 2714$.

Acknowledgement

The author wishes to express his thanks to H. Fleuren for using his set partitioning code.

References

[1] Feller, W., *An Introduction to Probability Theory and Its Applications*, Volume II, Wiley, 1966.

[2] Fleuren, H., "Computational and applicability study of the set partitioning approach for vehicle routing and scheduling", unpublished Ph.D. thesis, University of Twente, April 1988.

[3] Papadimitriou, C.H., Steiglitz, K., *Combinatorial Optimization*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[4] Syslo, M.M., Deo, N., and Kowalik, J.S., *Discrete Optimization Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1983.