The best variety or an almost best one? : a comparison of subset selection procedures

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 92-22

The best variety or an almost best one? A comparison of subset selection procedures

P. van der Laan

Eindhoven, June 1992
The Netherlands

Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, statistics, operations research and systems theory
P.O. Box 513
5600 MB Eindhoven - The Netherlands

Secretariate: Dommelbuilding 0.03
Telephone: 040-47 3130

ISSN 0926 4493

PAUL VAN DER LAAN

Department of Mathematics and Computing Science
Eindhoven University of Technology
Eindhoven, The Netherlands

THE BEST VARIETY OR AN ALMOST BEST ONE? A COMPARISON OF SUBSET SELECTION PROCEDURES

Summary: Given are k varieties. The best variety is defined as the variety with the largest average yield per plot of common unit size. An almost best or an ε-best variety is a variety with an average yield on a distance not larger than $\varepsilon(\geq 0)$ from the best variety. Subset selection is considered for selection of the best variety, but also for selection of an ε-best variety. A comparison between these two selection goals is made by investigating the relative efficiency of subset selection of an ε-best variety. An application is the field of variety testing is presented.

1. INTRODUCTION

Given are k (integer $k \geq 2$) varieties $V_1, V_2, ..., V_k$, and k independent random variables $X_1, X_2, ..., X_k$ corresponding with these varieties based on a common number of n independent observations. The random variable X_i $(i = 1, 2, ..., k)$ may be the average yield per plot of common unit size. We assume that the random variable X_i has a continuous distribution function $F(x-\mu_i)$ and density function $f(x-\mu_i)$ with unknown expectation μ_i $(i = 1, 2, ..., k)$.

The problem considered is to select a non-empty subset of the collection of k varieties, as small as possible, such that the probability to include the best variety into the subset is at least equal to P^*, with $k^{-1} < P^* < 1$. The best variety is defined as the variety with the largest value of the location parameter μ. We suppose that the best variety is unique, otherwise we assume that tagging has been used.

For this problem of selecting the best variety the subset selection procedure of Gupta can be used (cf. Gupta (1965)). The formulation of the sketched problem as a selection problem enables the experimenter to answer his question regarding the best variety in an adequate way. In this paper we shall present in section 2 Gupta's statistical subset selection procedure. In section 3 selection of an almost or an ε-best variety is considered. The efficiency of selecting an ε-best variety is investigated in section 4. Section 5 presents an application in the field of variety testing. Some concluding remarks are given in section 6.

2. SUBSET SELECTION OF THE BEST VARIETY

In this section we shall indicate the subset selection procedure of Gupta. The subset selection approach of Gupta provides a subset of varieties from the total collection of k varieties in such a way that the probability of a correct selection is at least P^*, with $k^{-1} < P^* < 1$. A correct selection is defined as a selection for which the best variety is an element of the selected subset. The confidence level P^* is predetermined by the experimenter. The size S of the subset is a random variable and must be as small as possible. The statistical subset selection rule is defined as follows:

Variety V_i is an element of the subset if and only if

$$X_i \geq \max_{1 \leq j \leq k} X_j - d \,,$$

where the selection constant d $(d > 0)$ has to be determined such that the probability of selecting the best variety in the subset is at least equal to P^*. The selection constant d can be found in Gibbons, Olkin and Sobel (1977) for standard Normal distributions for the X_i $(i = 1, 2, ..., k)$. Extensive tables of d for standard Normal distributions can be found in Butler and Butler (1987). For common unknown σ^2 of the Normal distributions one has to use the pooled estimator of σ^2, namely

$$s^2 = \{k(n-1)\}^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2 = k^{-1} \sum_{i=1}^{k} s_i^2 \,,$$

where x_{ij} $(j = 1, 2, ..., n)$ are the n independent observations on X_i and $\bar{x}_i$ is the corresponding sample mean. The selection constant d has to be replaced by a different selection constant

d_2 which can be derived from tables in Gibbons, Olkin and Sobel (1977).

Subset selection is a flexible form of selection, in the sense that the number of replications has not to be determined in advance. Also after the experiment has been executed, the selection can be carried out. Of course, the number of replications of the experiment has influence on the (expected) size of the subset. A relatively large subset indicates, apart from random fluctuations, that the number of replications is small or that the expected yields of the varieties are close together.

3. SUBSET SELECTION OF AN ε-BEST VARIETY

The requirement to select precisely the best variety may be a strong qualification for selection in the case that the best variety is in the neighbourhood of other varieties. The result will be that in general large subsets will be obtained. A possibility to overcome this difficulty is to increase the number of replications. But if this possibility is not realistic in practice, then there exists an alternative way out. This alternative is to transform the goal 'to select the best variety' into a different goal, namely to select 'an almost best variety'.

Let us assume that the best variety and the second best are near each other. Formulated more precisely this means that the average yields are close together, say on a distance less than ε, with $\varepsilon \geq 0$. If ε is relatively small, then it is in practice usually not of interest whether one selects the best variety or the next best. Not every difference in average yield is practically important. In other words there will be practical situations for which a more or less imprecise selection is adequate. Selection of either the best variety or an almost best variety will both be allowed. In this context an almost best variety is a variety with expected yield on a distance less than ε from the best one. This last variety will be called an ε-best variety. A variety V_i is an ε-best variety if $\mu_i \geq \max_{1 \leq j \leq k} \mu_j - \varepsilon$. If is clear that the best variety is also an ε-best variety. Thus for each $\varepsilon \geq 0$ there always exists at least one ε-best variety. If $\varepsilon = 0$, then there exists one and only one ε-best variety, namely the best variety, assuming there are no ties. A correct selection in this context is the selection of a subset which contains at least one ε-best variety. The selection rule considered is:

Select variety V_i $(i = 1, 2, ..., k)$ in the subset if and only if

$$X_i \geq \max_{1 \leq j \leq k} X_j - c \ ,$$

where the selection constant $c \geq 0$ has to be determined such that $P(CS) \geq P^*$, with $k^{-1} < P^* < 1$.

For Normal distributed variables X_i $(i = 1, 2, ..., k)$ with common scale parameter σ one can prove that for the selection constant c the following holds:

$$c = (d - \varepsilon)\sigma \ ,$$

where d en ε are measured in units of σ. Details can be found in Van der Laan (1992).

4. SELECTION OF AN ε-BEST VARIETY IN COMPARISON WITH SELECTION OF THE BEST

In this section a comparison is made between selection of the best variety and selection of an ε-best variety. This comparison will be based on the P^* value for selecting an ε-best variety and the probability of selecting the best variety, both for the statistical selection rule $X_i \geq \max_{1 \leq j \leq k} X_j - d\sigma + \varepsilon\sigma$. Without loss of generality we can assume that $\sigma = 1$. The efficiency of the procedure to select an ε-best variety, relative to selecting the best variety, is defined as the relative gain G_r: the gain in minimal probability of correct selection of an ε-best variety relative to that of the best variety. The computations are based on interpolations so the values of G_r are of limited accuracy. In table 1 some results are presented.

Table 1

The relative gain G_r in percentages for selecting an ε-best variety in comparison with selecting the best variety, for $P^* = 0.90$ and some values of ε and k.

ε	$k = 5$	$k = 10$	$k = 25$	$k = 100$
0.2	4.3%	4.4%	4.7%	4.8%
0.5	11.2%	11.9%	12.4%	12.9%

It is to be expected that for larger values of ε the gain is much larger. For instance, for $\varepsilon = 1$ and $k = 10$ one find $G_r = 37\%$.

5. AN APPLICATION IN THE FIELD OF VARIETY TESTING

In a 1991 Sugar Beet trial twenty five varieties were investigated in a complete randomized block design with $b = 4$ blocks. The results, the average root yield of sugar (arys), are given in next table. The original situation of this application is changed in view of the confidentiality of the results of the experiment. No practical conclusion can be deduced from the experiment and the data.

variety nr.	1	2	3	4	5	6	7
arys	645.8	583.1	652.1	633.3	614.5	645.8	608.2
variety nr.	9	10	11	12	13	14	15
arys	620.7	652.1	664.6	645.8	670.9	627.0	608.2
variety nr.	16	17	18	19	20	21	22
arys	683.4	620.7	620.7	670.9	645.8	633.3	595.6
variety nr.	23	24	25				
arys	620.7	645.8	633.3				

The standard error obtained from the analysis of variance is equal to 28.5975. The standard deviation has been taken as $28.5975/\sqrt{4} = 14.299$. In the next illustration we shall use this value as if it is the known value of the standard deviation.
Using subset selecting with confidence level $P^* = 0.90$, we find that a variety has to be taken in the subset if and only if the average is larger than or equal to

$$683.4 - 3.391133 * 14.299 = 634.91 .$$

The subset consists of the following varieties, indicated by their variety number:

$$\{1, 3, 4, 7, 10, 11, 12, 13, 16, 19, 20, 24\} .$$

For $\varepsilon = 1$ a variety is taken in the subset if and only if its average is larger than or equal to

$$683.4 - (3.391133 - 1) * 14.299 = 649.21 .$$

The corresponding subset is

$$\{3, 10, 11, 13, 16, 19\} .$$

The number of selected varieties is fifty percent of the number of varieties in the first subset.

6. SOME CONCLUDING REMARKS

In this paper we have discussed some aspects of subset selection which are of interest from a practical point of view. A conclusion may be that instead of trying to select the best variety, it may be worthwile to consider as selection goal: selection of an almost best variety. To consider this selection goal it is possible to get on the average smaller subsets. This means in practice in most cases an important profit.

Acknowledgement: I wish to express my gratitude to Dr. Marta Morales (VanderHave Research, Rilland) for providing me the practical application in which the concept of subset selection is of importance.

REFERENCES

Butler K.L., Butler D.G. (1987). Tables for selecting the best population. Queensland Biometrical Bulletin 2, Queensland Department of Primary Industries, Brisbane, QLD 4001, Australia.

Gibbons J.D., Olkin I., Sobel M. (1977). Selecting and Ordering Populations: A New Statistical Methodology. John Wiley & Sons, Inc., New York.

Gupta S.S. (1965). On some multiple decision (selection and ranking) rules. Technometrics 7, 225 - 245.

Laan P. van der (1992). Subset Selection of an almost best treatment. Biometrical Journal 34, no. 5, 1-10.

List of COSOR-memoranda - 1992

Number	Month	Author	Title
92-01	January	F.W. Steutel	On the addition of log-convex functions and sequences
92-02	January	P. v.d. Laan	Selection constants for Uniform populations
92-03	February	E.E.M. v. Berkum H.N. Linssen D.A. Overdijk	Data reduction in statistical inference
92-04	February	H.J.C. Huijberts H. Nijmeijer	Strong dynamic input-output decoupling: from linearity to nonlinearity
92-05	March	S.J.L. v. Eijndhoven J.M. Soethoudt	Introduction to a behavioral approach of continuous-time systems
92-06	April	P.J. Zwietering E.H.L. Aarts J. Wessels	The minimal number of layers of a perceptron that sorts
92-07	April	F.P.A. Coolen	Maximum Imprecision Related to Intervals of Measures and Bayesian Inference with Conjugate Imprecise Prior Densities
92-08	May	I.J.B.F. Adan J. Wessels W.H.M. Zijm	A Note on "The effect of varying routing probability in two parallel queues with dynamic routing under a threshold-type scheduling"
92-09	May	I.J.B.F. Adan G.J.J.A.N. v. Houtum J. v.d. Wal	Upper and lower bounds for the waiting time in the symmetric shortest queue system
92-10	May	P. v.d. Laan	Subset Selection: Robustness and Imprecise Selection
92-11	May	R.J.M. Vaessens E.H.L. Aarts J.K. Lenstra	A Local Search Template (Extended Abstract)
92-12	May	F.P.A. Coolen	Elicitation of Expert Knowledge and Assessment of Imprecise Prior Densities for Lifetime Distributions
92-13	May	M.A. Peters A.A. Stoorvogel	Mixed H_2/H_∞ Control in a Stochastic Framework

Number	Month	Author	Title
92-14	June	P.J. Zwietering E.H.L. Aarts J. Wessels	The construction of minimal multi-layered perceptrons: a case study for sorting
92-15	June	P. van der Laan	Experiments: Design, Parametric and Nonparametric Analysis, and Selection
92-16	June	J.J.A.M. Brands F.W. Steutel R.J.G. Wilms	On the number of maxima in a discrete sample
92-17	June	S.J.L. v. Eijndhoven J.M. Soethoudt	Introduction to a behavioral approach of continuous-time systems part II
92-18	June	J.A. Hoogeveen H. Oosterhout S.L. van der Velde	New lower and upper bounds for scheduling around a small common due date
92-19	June	F.P.A. Coolen	On Bernoulli Experiments with Imprecise Prior Probabilities
92-20	June	J.A. Hoogeveen S.L. van de Velde	Minimizing Total Inventory Cost on a Single Machine in Just-in-Time Manufacturing
92-21	June	J.A. Hoogeveen S.L. van de Velde	Polynomial-time algorithms for single-machine bicriteria scheduling
92-22	June	P. van der Laan	The best variety or an almost best one? A comparison of subset selection procedures