ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Pumplün, Constanze; Weihs, Claus; Preusser, Andrea

Working Paper

Experimental Design for Variable Selection in data bases

Technical Report, No. 2004,72

Provided in Cooperation with:
Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475), University of Dortmund

Suggested Citation: Pumplün, Constanze; Weihs, Claus; Preusser, Andrea (2004) : Experimental Design for Variable Selection in data bases, Technical Report, No. 2004,72, Universität Dortmund, Sonderforschungsbereich 475 - Komplexitätsreduktion in Multivariaten Datenstrukturen, Dortmund

This Version is available at:
https://hdl.handle.net/10419/22585

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Experimental Design for Variable Selection in data bases

Constanze Pumplün, Claus Weihs, and Andrea Preusser

University of Dortmund*
Department of Statistics
44221 Dortmund, Germany

Abstract. This paper analyses the influence of 13 stylized facts of the German economy on the West German business cycles from 1955 to 1994. The method used in this investigation is Statistical Experimental Design with orthogonal factors. We are looking for all existing Plackett-Burman designs realizable by coded observations of these data. The plans are then analysed by regression with forward selection and various classification methods to extract the relevant variables for separating upswing and downswing of the cycles. The results are compared with already existing studies on this topic.

1 Introduction

In the following, existing data are analysed using the method of statistical experimental design. The aim of experimental design is to estimate factor effects with the highest accuracy possible. Usually, an experimental design with fixed factor levels is taken and the response of the experiment is used to find factors of high influence with as few experiments as possible. Thus the optimal factors determining the response are found faster and with less expense than by carrying out all experiments with all possible factor level combinations. In order to detect the variables which do influence the up- and downswing phases of the economy, we use a special type of screening plans, namely Plackett-Burman plans. Contrary to the method of full factorial designs, which investigate main effects and all possible interactions, these plans are employed to find only the main effects in the model.

The original data used here are highly correlated. In order to eliminate these correlations, the data are coded by -1 and +1 only and then special observations are selected building Plackett-Burman plans. The main advantage of this method is that it selects the most important factors not disturbed by correlations in the data. By this procedure, on the one hand, the data are reduced by the discrete coding by -1 and +1 and on the other hand by choosing special observations only. In order to at least partially compensate this, we are analysing all existing Plackett-Burman plans with respect to the data

* This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475. We also thank Uwe Ligges and Karsten Luebke for their support.

and finally choose those variables which are, what we call, uniquely correlated to the up- and down phases of the economy. The following investigations are based on 13 stylized facts of the West German economy (cf. Heilemann and Münch (1999)) which have been selected by Heilemann and Münch to explain the German business cycle. There exists already a number of papers which analyse and interpret these data based on, e.g. classification methods like linear discriminant analysis and time series analysis (cp. Heilemann and Münch (1999), Weihs, Röhl and Theis (1999), Weihs and Garczarek (2002)).

In this paper in the first step, we code the data to -1 and +1 and in the second step we look for all Plackett-Burman plans in the coded data. All these plans are analysed by stepwise regression with forward selection, by unpruned classification trees, by trees consisting only of the tree stump and by stepwise linear discriminant analysis (cp. Röver (2003)). All this is based on an a priori classification of the response in the phases 'up', and 'down' in the years under investigation, based on Heilemann and Münch (1999). Finally, the variables which have turned out to be important are compared with the results of existing studies.

2 Data

The predictor data set consists of 13 variables which have been measured quarterly (157 quarters) in the years 1955/4 to 1994/4 (price index base is 1991) (cf. Heilemann and Münch (1999)). The variables (and their abbreviations) are real-gross-national product-gr (BSP91JW), real-private-consumption-gr (CP91JW), government-deficit-rate (DEFRATE), wage-and salary-earners-gr (EWAJW), net-export-rate (EXIMRATE), money-supply-M1-gr (GM1JW), real-investment-in-equipment-gr (IAU91JW), real-investment-in-construction-gr (IB91JW), unit-labour-cost-gr (LSTKJW), GNP-price-deflator-gr (PBSPJW), consumer-price-index-gr (PCPJW), nominal-short-term-interest-rate (ZINSK), real-long-term-interest-rate (ZINSLR). The letters 'gr' are an abbreviation of 'growth rates relative to last years corresponding quarter'.

3 Plackett-Burman designs

Heilemann and Münch (1999) distinguish 4 phases of the business cycle: 'upswing', 'upper turning point', 'downswing' and 'lower turning point'. Each quarter has been assigned one of these phases which we assume to be the correct one. Here only the phases 'up-', and 'downswing' are considered. Therefore, the phases 'upper turning point' and 'lower turning point' are split in the middle, i.e. if, e.g., the 'upper turning point' phase lasts for k quarters, $k \in \mathbb{N}, [k/2]$ quarters will be added to the 'upswing' phase and $k - [k/2]$ quarters will be added to the succeeding 'downswing' phase, where $[x]$ denotes the so called Gauß brackets, i.e. the largest integer less or equal

to x, $x \in \mathbb{N}$. An analogous convention holds for the 'lower turning point' phase. These two phases 'upswing' and 'downswing' are coded by 0 and 1, respectively. Note that two phase consideration is standard in business cycle analysis. Thus, it is the natural starting point for our studies. Extensions to 4 classes are planned.

Plackett-Burman plans only exist if the number of experiments n is a multiple of four and the number of variables is n-1 (cf. Plackett and Burman (1946), Weihs and Jessenberger (1999)). The Plackett-Burman plan for $n = 8$ is shown in Table 1.

	x1	x2	x3	x4	x5	x6	x7
1	-	-	-	-	-	-	-
2	-	-	+	-	+	+	+
3	+	-	-	+	-	+	+
4	+	+	-	-	+	-	+
5	+	+	+	-	-	+	-
6	-	+	+	+	-	-	+
7	+	-	+	+	+	-	-
8	-	+	-	+	+	+	-

Table 1. Plackett-Burman plan with 8 experiments.

The second row is called generating row, as it generates the rows 3–8 of the matrix by being shifted one position to the right at each step. Plackett-Burman plans are orthogonal arrays in the sense of (Hedayat et al. (1999)), they are of the form $OA(4\lambda, 4\lambda - 1, 2, 2)$, $\lambda \in \mathbb{N}$, ($\lambda = 2$ in Table 1), i.e. each factor has only two levels -1 and +1, the sum of each column is 0 and columns are pairwise orthogonal. If an 8th column consisting only of +1's is added to the matrix, one gets a unique Hadamard matrix of order 8 (cp. Hedayat et al. (1999)). Therefore it is necessary to code the existing data in +1 and -1, in order to look for Plackett-Burman designs. For each variable, all values less than its median are taken as -1 and all values greater than or equal to its median are taken as +1. As there are 13 variables, one looks for Plackett-Burman plans with $n = 8$ or $n = 12$ in the coded data. 113 different plans were found for $n = 8$ and none for $n = 12$.

The algorithm for finding these plans is first to look for all rows which contain at least seven times the number -1. The corresponding columns are then searched for the generating row. After this has been found, the search continues for the generating row shifted one position to the right, etc. This process has to be carried out for all possible permutations of the original seven columns. A much faster algorithm has been suggested by S.Haustein (private communication), where one looks for the base row $u0 = (- - - - - - -)$ and then searches for a row v in the corresponding columns with Hamming distance 4 to $u0$. After this has been found, one looks for a row $v1$ with Hamming distance 4 to $u0$ and v. This process is continued until eight rows have been found which are equidistant with Hamming distance 4. These eight

rows form a Plackett-Burman plan for $n = 8$, because the Plackett-Burman plan for $n = 8$ is an orthogonal array of the form $OA(8, 8-1, 2, 2)$ and this class has only one isomorphism class. Here two arrays are said to be isomorphic (cf. Hedayat et al. (1999)), if one can be obtained from the other by permutations of rows, columns or factor levels.

In the following investigations, a linear screening model is used, $y = X\beta + \epsilon$, where $X = (\mathbf{1}, A)$ is an $(n \times n)$ matrix with $\mathbf{1} = (1, 1, 1, ...)^t$ and A the Plackett-Burman matrix. β is the vector of unknown coefficients, y the result vector with the coded business cycle phases and ε the error vector.

4 Results

4.1 Stepwise regression by forward selection

113 different Plackett-Burman plans were found by the method described in 3. When evaluating these plans by stepwise regression with forward selection with respect to y (cp. Weihs, Jessenberger (1999)), we used the F-test at level 0.2. Figures 1, and 2 show the absolute and the relative frequency of the selected variables (dark bars). The light bars show how often each variable appears in all 113 Plackett-Burman plans. Figure 1 thus shows that each variable is at least once in a plan (light bars). The variables which turn out to be most important by this method are 'DEFRATE', 'EXIMRATE', 'LSTKJW', 'IAU91JW' and 'ZINSK'(cp. Figure 2). If one uses the F-test with level 0.05 one gets the same variables except 'EXIMRATE'. It is also interesting that in almost half of all cases none of the variables turns out to be important. Furthermore it strikes that for all variables the dark bars are rather small, compared to the light ones. That means that although a variable appears often in the plans it is chosen only a few times as important concerning the up- and down of the economy.

4.2 Classification methods

In the 113 plans, variables are selected also by different classification methods, i.e.unpruned classification trees (TreeAllNodes), classification trees with only the tree stump (TreeStump) and stepwise linear discriminant analysis (cp. Röver (2003)). Figures 3, and 4 again show the absolute and the relative frequency of selected variables by the different methods. The number in brackets following the variable name indicates how often the variable appears in a Plackett-Burman plan. Classification by unpruned trees yields as important variables 'BSP91JW', 'CP91JW', 'DEFRATE' and 'EXIMRATE'. Using only the tree stump yields the same variables without 'CP91JW' as important. This is the same result one gets by stepwise linear discriminant analysis. On the whole, these three classification methods yield similar results but on different levels.

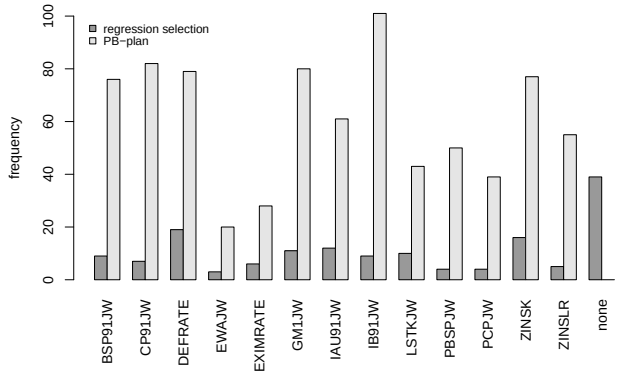

Fig. 1. Absolute frequency of variable selected by stepwise regression with forward selection.

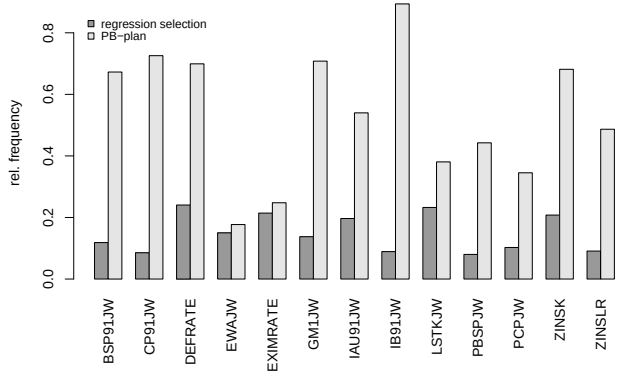

Fig. 2. Relative frequency of variable selected by stepwise regression with forward selection.

For all used classification methods as well as for stepwise regression with forward selection it is important to know how the rows which build the Plackett-Burman plans are distributed. This is illustrated in Figure 5 which shows how often each row is contained in a plan. Note that the outstanding row number 72 refers to the 4th quarter of 1972 and row number 145 to the 1st quarter of 1991. These years are special years from an economic point of view, as in 1972 the German economy suffered from the oil price shock. The German unification influences the post 1990 data, an effect shown in the first quarter of 1991.

4.3 Variable assessment

If one wants to decide which of the above variables plays a dominant role with respect to the business cycle, it is important to assess their correlation

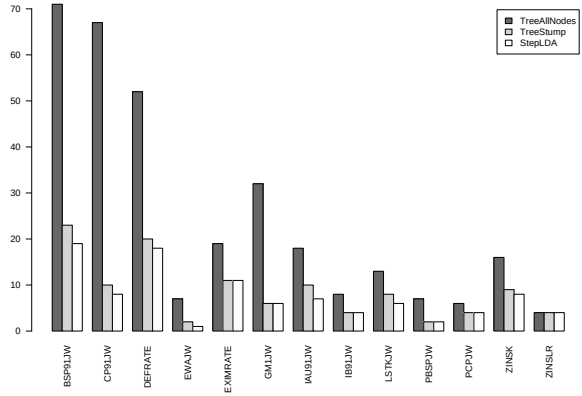

Fig. 3. Absolute frequency of variables selected in Plackett-Burman design.

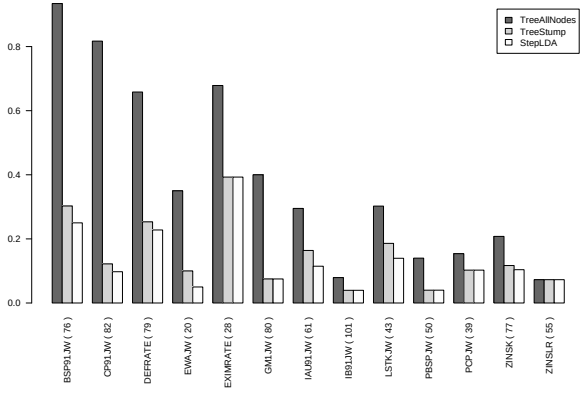

Fig. 4. Relative frequency of variables selected in Plackett-Burman design.

in all those Plackett-Burman plans where the corresponding variable was included. It turns out (see Table 2) that unit labour costs('LSTKJW') is clearly positively correlated to y (84% of all cases) and the government deficit ('DEFRATE') can still be considered as positive correlated (78% of all cases), taking into account a possible error margin. No variable is clearly negatively correlated to y. Hence, one may finally consider those variables as important which on the one hand are chosen most often, both by regression and by classification, and which on the other hand possess a distinct positive or negative correlation to y. Using this decision criterion, one gets 'unit labour costs' ('LSTKJW') and 'government deficit' ('DEFRATE') as variables which clearly determine the West German business cycles.

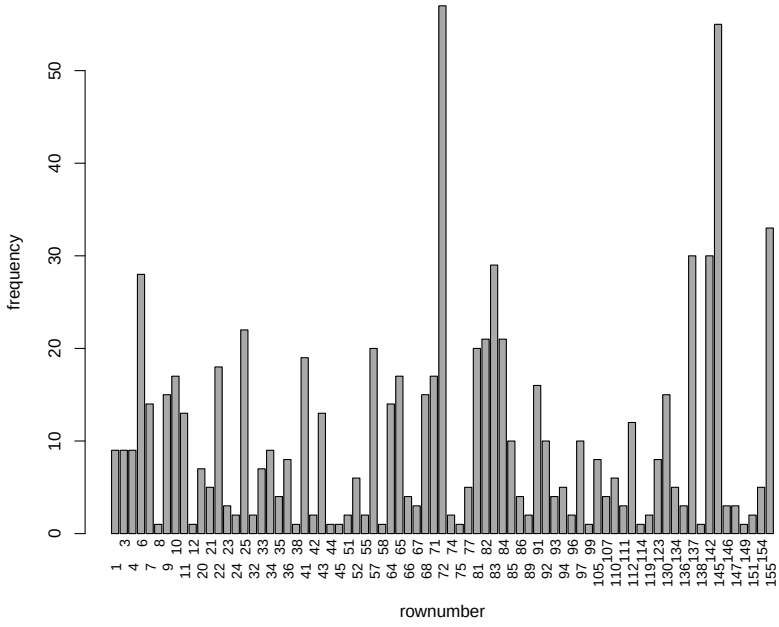

Fig. 5. Absolute frequency of rows in all Plackett-Burman plans.

In previous studies of this topic (cp. e.g. Weihs and Garczarek (2002), Weihs et al. (1999)) the variables most influential for the West German business cycle in the 4 phase case were 'wage and salary earners' ('EWAJW') and 'unit labour costs' ('LSTKJW'). Moreover, if one compares the above method to stepwise regression by forward selection on the whole data set,

	Positive	**Negative**	**No Cor.**	**% positive**
BSP91JW	18	41	17	24
CP91JW	32	24	26	39
DEFRATE	62	3	14	78
EWAJW	13	2	5	65
EXIMRATE	19	5	4	68
GM1JW	8	53	19	10
IAU91JW	5	40	16	8
IB91JW	21	51	29	21
LSTKJW	36	1	6	84
PBSPJW	29	6	15	58
PCPJW	24	6	9	49
ZINSK	50	14	13	65
ZINSLR	25	18	12	45

Table 2. Correlation with respect to y.

again taking level 0.2 in the F-test, the model 'LSTKJW' + 'IAU91JW' + 'DEFRATE' + 'ZINSK' + 'CP91JW' + 'BSP91JW' is chosen. This strongly indicates the importance of 'LSTKJW' and 'DEFRATE'. Also stepwise linear discriminant analysis, classification by unpruned trees and classification trees using only the tree stump were applied on the whole data set. The application of unpruned classification trees shows 'IAU91JW' to be the most important variable, as does classification trees using only the tree stump. Stepwise linear discriminant analysis shows that besides 'IAU91JW', also two other variables are important, 'LSTKJW' and 'PCPJW'.

5 Conclusion

'Unit labour costs' ('LSTKJW') has been detected as an important variable by this method as well as by previous methods (cp. 4.3). This strongly indicates that this variable has a great influence on the West German business cycle. The question why the 'government deficit' ('DEFRATE') turns out to be important here, but does not so in previous studies, requires a thorough analysis of the influence of the methods applied here on the results. The advantage of using Plackett-Burman plans lies in the clean and easy selection of variables in determining the important variables. This is only a first step in this direction. Right now, we are investigating only the correlations of those variables with the business cycle, which have turned out to be important in the above described investigations. A next step could be to investigate a similar procedure with full factorial designs or fractional factorial designs. These plans also respect orthogonality, but in addition permit interactions between the factors.

References

HEDAYAT, A., SLOANE, N. and STUFKEN, J. (1999): *Orthogonal Arrays.* Springer Verlag, New York, Berlin, Heidelberg.

HEILEMANN, U. and MÜNCH, J. (1999): Classification of West German Business Cycles. *Technical Report 11, SFB 475 Universität Dortmund.*

PLACKETT, R. L. and BURMAN, J. P. (1946): The design of optimum multifactorial experiments. *Biometrica, 33, 305–325.*

RÖVER, C. (2003): Musikinstrumentenerkennung mit Hilfe der Hough - Transformation. *Diplomarbeit, Fachbereich Statistik, Universität Dortmund.*

WEIHS, C. and JESSENBERGER, J. (1999): *Statistische Methoden zur Qualitätssicherung und -optimierung in der Industrie.* Wiley-Vch, Weinheim.

WEIHS, C., RÖHL, M.C. and THEIS, W. (1999): Multivariate Classification of Business Phases. *Technical Report 26, SFB 475, Universität Dortmund.*

WEIHS, C. and GARCZAREK, U. (2002): Stability of multivariate representation of business cycles over time. *Technical Report 20, SFB 475, Universität Dortmund.*