ETM

ENGINEERING & TECHNOLOGY MANAGEMENT

Title: Large Scale Analysis of Messy Data: Analysis of Programming Practices

Course: EMGT 555/655
Term: Fall
Year: 1997
Author(s): T. Mackenzie

Report No: P97076

ETM OFFICE USE ONLY	
Report No.:	See Above
Type:	Student Project
Note:	This project is in the filing cabinet in the ETM department office.

Abstract: Businesses today are trying to economize their practices to increase efficiencies and profits in a highly competitive society. One way a business could do this would be to make a graph of productivity of workers vs. length of work shift. The graph may show that a certain length of work shift may show a greater productivity rate than other shifts. By doing this, the business can decide to reduce the workload or increase the workload to increase productivity, and possibly increase company profits. However, with so many factors and variables, its simple plot may not provide enough information to make such a decision. There are several methods of analysis that can assist the decision-maker in determining a course of action. The following paper deals with software cost estimation, which contains 26 variables that deal with computer programming. Cluster analysis and Factor analysis are introduced to help make interpretations and correlation within the variables. Large Scale Analysis of Messy Data: Analysis of Programming Practices

Large Scale Analysis of Messy Data
Analysis of Programming Practices

Troy Mackenzie

EMP-P9776

LARGE SCALE ANALYSIS
OF MESSY DATA

ANALYSIS OF PROGRAMING PRACTICES

BY TROY MACKENZIE
EMGT 565 FINAL PROJECT

INTRODUCTION

Businesses today are trying to economize their practices to increase efficiencies and profits in a highly competitive society. One way a business could do this would be to make a graph of productivity of workers vs. length of work shift. The graph may show that a certain length of work shift may show a greater productivity rate than other shifts. By doing this, the business can decide to reduce the workload or increase the workload to increase productivity, and possibly increase company profits. However, with so many factors and variables, it a simple plot may not provide enough information to make such a decision. There are several methods of analysis that can assist the decision-maker in determining a course of action. The following paper deals with software cost estimation, which contains 26 variables that deal with computer programming. Cluster analysis and Factor analysis are introduced to help make interpretations and correlation within the variables.

DESCRIPTION OF VARIABLES

The following descriptions explain the variables used in the projects that were analyzed:

TYPE - Type of program:

- Business (BUS) - Data processing programs for business purposes such as payroll, scheduling, and accounting.
- Control (CTL) -Operating system programs responsible for the overall management of the computer and its resources.
- Human/Machine Interface (HMI) – Programs to handle any boundary at which people interact with machines.
- Scientific (SCI) – Programs designed to handle mathematical formulas and matrices.
- Support (SUP) – Programs to get data and programs into the CPU, get processed information out, and store data and programs for ready access to the CPU.
- System (SYS) – Programs that control the internal operations of the computer system, such as operating systems, computers, interpreters, assemblers, graphics support programs, & mathematical routines.

YEAR - Year program was created

PRODUCT ATTRIBUTES

- LANG - Programming language used
- RELY - Required software reliability
- DATA - Data base size
- CPLX - Product complexity
- AAF -

COMPUTER ATTRIBUTES

- TIME - Execution time constraint

STOR - Main storage constraint
VIRT - Virtual machine volatility
TURN - Computer turnaround time
TYPE -

PERSONAL ATTRIBUTES

ACAP - Analyst capability
AEXP - Applications Experience
PCAP - Programmer capability
VEXP - Virtual machine experience
LEXP - Programming language experience
CONT -

PROJECT ATTRIBUTES

MDOP - Modern programming practices
TOOL - Use of software tools
SCED - Required development schedule
RVOL -

MAN MONTHS

NOM – Nominal man months required
EST - Estimated man months required
ACT - Actual man months required

Numbers using 1.0 as a standard starting point represented most of the variables above. Numbers greater than 1.0 indicated better performance scores for that variable, where numbers greater than 1.0 indicated lower performance scores for those variables. The year variable indicates the last two digits of the year the program was created. For example: The first specimen on Table 1 indicates the year 72, with means the program was created in the year 1972. The language variable is marked by the type of language it used to create the program. There are a total of 7 different programming languages included in this study and they are:

- COBOL (COB) – Acronym for Common Business Oriented Language, which is a high level language developed for business data processing applications.
- Fortran (FTN) – Acronym for FORmula TRANslator, a widely used high level programming language for performing mathematical, scientific and engineering computations.
- Hollerith Code (HOL) – Particular code used to represent alphanumeric data on punched cards, named after Herman Hollerith, who was the originator of punched card tabulating.
- Jovial (JOV) – Jule's Own Version of the International Language, a high level programming language used in the 1960-1970's.
- Modula-2 (MOL) – High level programming language similar to Pascal. It supports a separate compilation of modules, where Pascal does not.
- Pascal (PSC) – High level structured programming language that has gained wide acceptance as a tool for both applications programming and system development.

PL/I (PLI) – High level programming language designed to process both scientific and business applications. Contains many of the best features of Fortran, COBOL, ALGOL, and others.

The other variables not mentioned above are self explanatory.

ANALYSIS METHODS

Several methods of analysis were considered in this project. When one looks at the data as shown in Table 1 (in the appendix), it is extremely difficult to make any comparisons yet find one that is correlated to another. A data reduction technique is needed to be able to reduce the variables into groups to allow us to make interpretations within the data set. Two methods were found to be the best to analyze this 'messy' set of data. The methods tried were cluster analysis and factor analysis. There may be several other methods available that may be better suited for this particular problem, but only the methods taught in Engineering Management, course 565 were considered. In the following analysis, the goal was to try to make the 'messy' data meaningful and useful by finding correlation between variables or finding groups that correlated with each other.

FACTOR ANALYSIS

Factor analysis was used to determine the correlational relationships between the characteristics of different programming projects. By using factor analysis, we could represent the variables with a smaller set of "derived" variables, or *factors*. There are twenty-six variables in this study, which makes it very difficult to find any correlation between the variables. Factor analysis will detect structure in between the variables above and classify them into smaller, more manageable, and interpretable number of *factors*. A plot of the eigenvalues, which vary from near 5.5 to close to zero, is shown in Figure 1. The eigenvalues corresponds to the equivalent number of variables, which the factor represents. For example, a factor associated with an eigenvalue of 5.5 indicates that the factor accounts for as much variance in the data collection as would 5.5 variables, on average. We are only interested in the factors that make up most of the variance, so a scree test was performed to determine how many factors should be kept for the analysis. The cutoff point was assumed to be where the continuous drop in eigenvalues leveled off. After this cutoff point, random "noise" is being extracted by the additional factors. Looking at Figure 1, the graph levels of at two locations, one of the two is questionable. The first location is after factor 4; it levels of slightly and then decreases rapidly again. The second critical location is after factor 6, where it levels off dramatically, and there is not appreciable drop in values after factor 6. Retaining 4 factors would account for 55.2% of the total variance of the data. Retaining 6 factors would account for 67.5% of the total variance. Using 4 or 6 factors results in accounting for most of the variance in the data with only a few factors, so factor analysis using four and six factors were investigated.

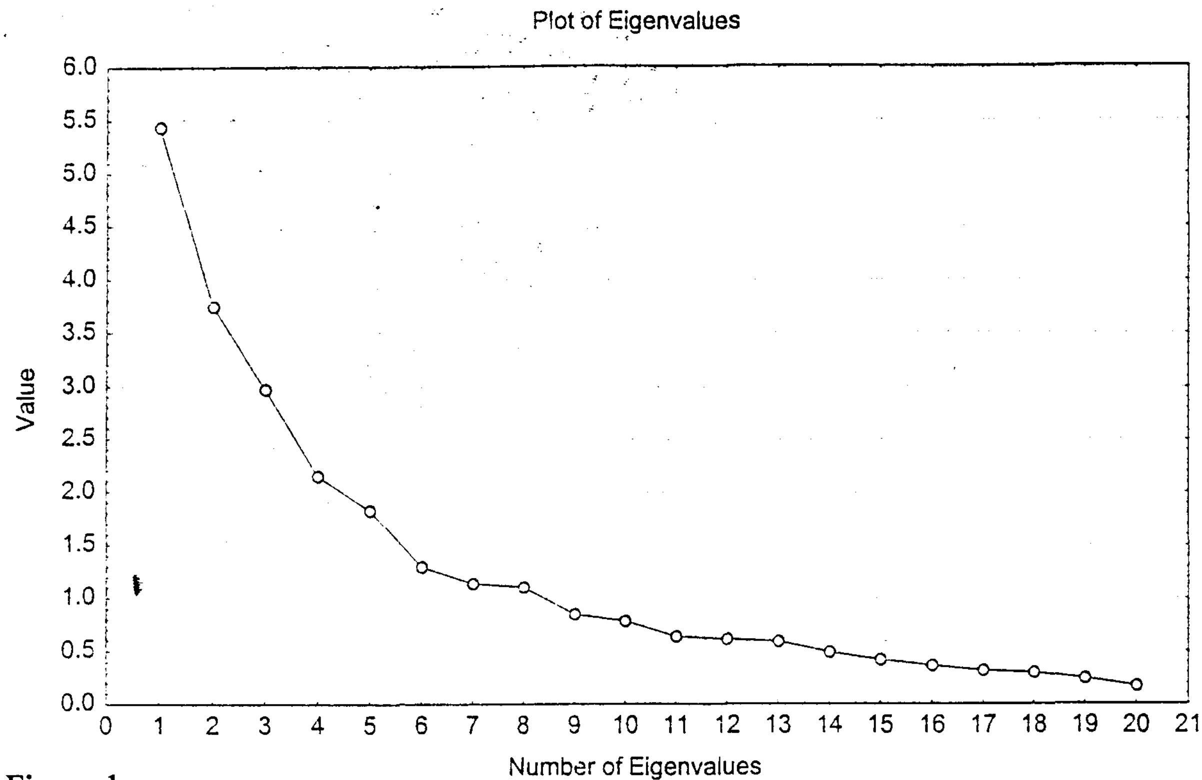


Figure 1

Tables 2 and 3 show the factor loadings for 4 and 6 factors respectively. The factor loadings can be interpreted as the correlation between the factors and the variables, which represents the most important information on which the interpretation of factors is based. A 'perfect' example of factor loadings would show that the first factor has most of the highest loadings, where each successive factor has lower numbers of high loadings. This is due to the fact that factors are extracted so that successive factors account for less and less variance. This pattern cannot be seen the either of the 4 or 6 factor loadings. The number of high loading seem to be randomly dispersed throughout the different factor groups with no significant pattern of showing more high loadings in the first factors than the successive factors. The high loadings are those that are 0.70 or greater and are the ones that provide the meaning and interpretation of the factor. Those with low or zero loadings on a particular factor will not contribute to the meaning of that factor, but rather will tend to contribute to the meaning of one of the other factors by virtue of their high loadings on those factors. Several methods of rotation were tried and the varimax raw rotation was retained since it showed the highest number of high factor loadings. At this point a decision was made to drop the analysis using four factors and use the six factors in the analysis. Looking at the factor loadings for factor one, high loadings are shown for required software liability, execution time constraint, and main storage constraint. The second factor has high loadings for the estimated man months required and the actual man

Table 2

STAT. FACTOR ANALYSIS	Factor Loadings (Varimax raw) (egmt.sta) Extraction: Principal components (Marked loadings are > .700000)			
Variable	Factor 1	Factor 2	Factor 3	Factor 4
TYPE	.181507	-.201512	.194318	.311037
YEAR	.055019	-.061613	.596012	-.229451
LANG	-.423549	-.262106	.159643	.602805
RELY	-.824053*	.175299	.202212	.206479
DATA	.092460	.583622	-.122889	-.378173
CPLX	-.484133	-.050516	.238269	.502554
AAF	.460524	.032522	.117918	.469761
TIME	-.685713	.082993	.075072	.368049
STOR	-.830942*	.037633	-.121746	.171007
VIRT	-.273439	-.063876	-.199027	.730909*
TURN	-.054037	.177544	-.554013	.081114
TYPE	.630143	.260416	.026747	.034212
ACAP	.359444	-.235888	-.549940	-.367318
AEXP	.408253	-.046498	.050148	.155355
PCAP	.219817	.093196	-.651996	-.377859
VEXP	-.051781	.044378	-.052911	.808130*
LEXP	-.264838	.049192	.007032	.825673*
CONT	.184039	-.200861	.443901	-.288439
MODP	.041877	.139975	-.882749*	.071631
TOOL	-.020649	-.133969	-.540979	.601633
SCED	.184773	.012521	.038782	.417484
RVOL	-.254662	-.010444	-.429904	.268425
NOM	.001004	.842294*	.093007	-.120546
EXT	-.045946	.935183*	-.151027	.069585
ACT	-.093972	.852375*	-.015893	.005645
_E_A__A	.029992	.505555	-.215633	.145653
Expl.Var	3.548776	3.314968	3.179238	4.252233
Prp.Totl	.136491	.127499	.122278	.163547

months required. Factor three shows high loadings for computer turnaround time, modern programming practices, and use of software tools. The fourth factor has high loadings for
virtual machine experience, programming language experience and virtual machine volatility. Factor 5 shows high loadings for applications experience and required development schedule. The last factor shows high loadings with analyst capability and programmer capability. At this point, names could be added to some of the factors as there is a general area as to what variables have the higher loadings. Discussion and naming of the factors are mentioned in the results.

Table 3

STAT. FACTOR ANALYSIS	Factor Loadings (Varimax raw) (egmt.sta) Extraction: Principal components (Marked loadings are > .700000)					
Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
TYPE	-.370514	-.177356	-.005390	.282171	-.423800	.340916
YEAR	.023398	-.034077	-.673030	-.166846	.270046	.098585
LANG	.347780	-.271331	.035537	.524646	-.154237	.398260
RELY	.817680*	.168157	-.102070	.164803	-.057666	.277747
DATA	.020451	.577486	-.011835	-.333699	.270224	-.279750
CPLX	.464625	-.107327	.111005	.283371	.056268	.560278
AAF	-.402271	.007470	-.002487	.400098	.337929	.166594
TIME	.719726*	.060135	.021494	.306236	.077424	.225212
STOR	.849427*	.011137	.179893	.110807	-.048272	.102982
VIRT	.320095	-.077294	.201758	.735221*	.142762	.033201
TURN	.130627	.093588	.688157	-.102696	.199272	-.083529
TYPE	-.624428	.201637	.247226	-.166134	.207382	.231183
ACAP	-.193855	-.226332	.138756	-.203122	.301684	-.745413*
AEXP	-.177677	-.077146	-.162962	.153556	.727457*	-.168198
PCAP	-.161995	.148915	.165172	-.096963	-.064406	-.830374*
VEXP	.060853	.062957	.020852	.885294*	.084146	.051176
LEXP	.251851	.048283	.080673	.822301*	.033558	.222819
CONT	-.133154	-.219425	-.346160	-.378134	.230511	.208569
MODP	-.025693	.118529	.743730*	.103523	-.092751	-.482446
TOOL	-.018949	-.183128	.696720	.480275	-.088302	.041178
SCED	-.005946	-.041666	.025168	.316958	.621471	.072363
RVOL	.298856	.005006	.221790	.390632	-.000126	-.331248
NOM	-.046330	.873147*	-.106017	-.047195	-.111821	.002184
EXT	.052748	.909151*	.271486	.005038	.065098	.058043
ACT	.071095	.887657*	-.060174	.117519	-.080074	-.085162
_E_A__A	-.002107	.418109	.533157	-.108732	.184121	.239877
Expl.Var	3.335072	3.273544	2.746419	3.645834	1.726829	2.691272
Prp.Totl	.128272	.125906	.105631	.140224	.066416	.103510

A plot of the factor loadings was created to help discriminate the different factors. In Figure 2, it shows a plot of factor loadings with Factor 1 vs. Factor 2. The independent factors are shown at the right and top, respectively. The variables classified within the factor are shown on the graph with high loadings while all the other variables are scattered all around the graph between factor loadings of about -.75 to .75. Figure 3 shows Factor 1 vs. Factor 3, which shows a similar pattern with groupings of variables in Factor 1 on the right, and Factor 3 on the top of the graph. Other low ranking variables are also scattered throughout the plot. Figures 4, 5, and 6 show the plots of loadings for Factor 1 vs. Factor 4, Factor 1 vs. Factor 5, and Factor 1 vs. Factor 6, respectively. The last three plots produce similar results as mentioned above. Figures 2-6 can be found in the appendix. The above plots reinforce the groupings of variables into factors that are mentioned by looking at the factor loadings.

CLUSTER ANALYSIS

Cluster analysis was performed to see if we could cluster groups into interpretable patterns, and to see if it agreed with the factor analysis groupings. In using cluster analysis, we are trying to classify a "mountain" of information into manageable meaningful groups, so that we may use the information to our advantage. When using cluster analysis, a tree diagram is drawn showing the linkage distances between the variables. First we have to determine what distance measure to use in order to produce meaningful results. The most commonly chosen type of distance measure, or Euclidean distance, was used since it measured the similarity between the variables.. Next, the amalgamation or linkage rule had to be determined to determine when two clusters are sufficiently similar to be linked together. Out of the several possible methods of amalgamation or linkage, the single linkage, or nearest neighbor, was chosen. This method links two clusters together when any two objects in the two clusters are closer together than the respective linkage distance. By using this method, we can produce a hierarchical tree diagram that form clusters sequentially in a hierarchical or "nested" fashion in which smaller clusters occur within larger ones, which would produce the most interpretable results.

Looking at the tree diagram for the 26 variables using the single linkage as the amalgamation or linkage method using Euclidean distances, we can see that most of the variables are clustered at a short linkage distance with a few variables being clustered near the 10,000 linkage distance. By relying on this graph only, we cannot make any conclusions since we cannot see how the variables are clustered together at the beginning. By looking at the amalgamation schedule, we can see where variables are clustered and what their Euclidean distances are. By looking at Table 4 in the appendix, 6 distinctive clusters were found. The first cluster consists of virtual machine experience and programming language experience at a linkage distance of .479. Use of software tools joined up with the first cluster at a distance of .63. The second cluster was formed at a linkage distance of 1.058, and consists of analyst capability and programmer capability. Required software reliability, execution time constraint, and main storage constraint made up the third cluster at a linkage distance of 1.26. The fourth cluster with a linkage distance of 8.37, consists of TYPE2 & CONT. At a linkage distance of 214.8, TYPE and YEAR made up the fifth cluster. The last cluster was formed at a linkage distance of 9770 and consists of nominal man months required, estimated man months required, and actual man months required. Several of the clusters mentioned above consisted of the same variables that defined a factor, which shows that both analysis's were able to determine a correlation between the same variables that made up a factor or cluster.

A plot of linkage distances across steps is shown in Figure 7. Looking at the plot, it shows that many clusters were formed at essentially the same linkage distance (near zero). This plot would be useful to determine how many clusters to retain and interpret if it

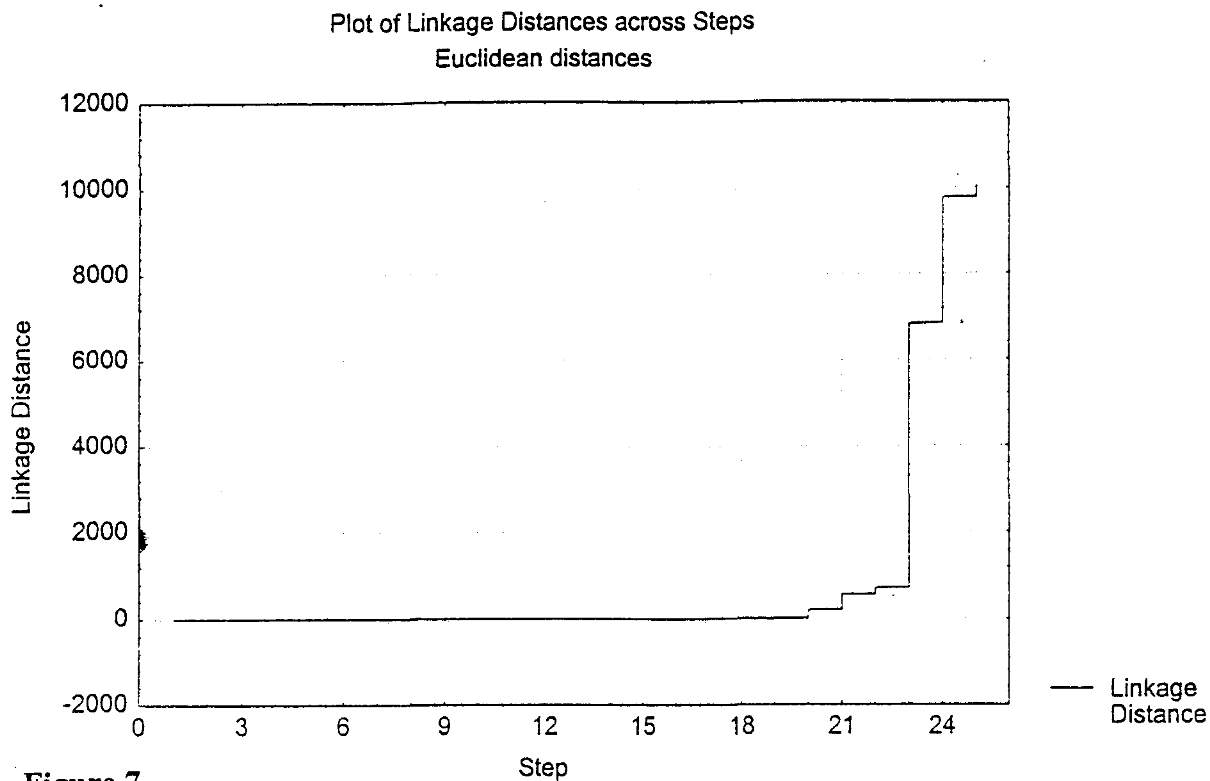


Figure 7

showed a large plateau after step 3. Since the graph is flat from the beginning, one cannot determine a cut off point, so we are left to make our own interpretations.

Using K-means clustering, the procedure would move objects around from cluster to cluster with the goal of minimizing the within-cluster variance and maximizing the between-cluster variance. In performing this procedure using 4 and 6 clusters, the results were less than satisfactory. In both cases, the program identified several clusters with only one variable in it, and in the extreme case assigned 20 variables to a cluster. Upon reviewing the results, no useful interpretations could be made and this method was dropped from the investigation.

Results

After performing these two statistical analysis methods to try to reduce the 26 variables to smaller, more manageable, and interpretable number of clusters or variables. Of the two analyses', the factor analysis method seemed to be more suitable, since it provided more information to interpret the results. Cluster analysis clustered variables with respect to those that had the most similarities with each other and did not account for the variances of other variables, nor could we measure how each variable within a cluster was correlated to each other. In factor analysis, six factors were found that would produce the most meaningful results, and accounted for 67.5% of the total variance among the variables. After taking attempt to interpret the 6 factors, some factors were named, others not, and patterns in each factor was interpretable.

Factor 1

This factor could be named restrictions since it deals with required software liability, execution time constraint, and main storage constraint. All three factors have positive loadings meaning when the value for required software liability is high, then the values for the other two variables most likely would be high too.

Factor 2
This factor could be named man-months since it deals with nominal man months, estimated man months, and actual man months. We must be careful to make interpretations from this factor because the values for these variables were considerably higher than the rest of the variables, which could affect the analysis significantly. I would ignore this factor since I am not sure about it.

Factor 3
This factor consists of computer turnaround time, modern programming practices, and use of software tools. All three variables had positive factor loadings, which suggests that if one variable has a higher value, then the others will. This pattern seems to make sense since if you use more modern programming practices and use software tools, the computer turnaround time would be quicker. A name for this factor could be productivity.

Factor 4
Factor 4 consists of virtual machine volatility, virtual machine experience, and programming language experience. All three factors also have positive factor loadings, which suggests that virtual machine volatility increases with virtual machine experience and programming language experience. A name for this factor could not be chosen, since this factor covers two different areas.

Factor 5
This factor consists of applications experience and required development schedule with positive loadings. This also suggests that with increased applications experience, the development schedule value increases. Not enough information was provided to name this factor with sufficient certainty.

Factor 6
The last factor consists of analyst capability and programming capability with negative loadings. We can say that this factor has to do with capability, and must realize that the variables have meaning opposite of capability and a possible name could be ineptly.

We must remember that the factor analytic technique and its interpretation are highly complicated that only individuals trained in advanced statistics could hope to understand it. But as we have seen above, it is possible to gain an appreciation of the system without a detailed knowledge of its theoretical underpinnings and computational intricacies.

Table 1

			PROJECT ATTRIBUTES					COMPUTER ATTRIBUTES					PERSONNEL ATTRIBUTES					
	TYPE	YEAR	LANG	RELY	DATA	CPLX	AAF	TIME	STOR	VIRT	TURN	TYPE	ACAP	AEXP	PCAP	VEXP	LEXP	CONT
1	BUS	72	COB	0.88	1.16	0.7	1	1	1.06	1.15	1.07	MAX	1.19	1.13	1.17	1.1	1	NOM
2	BUS	76	COB	0.88	1.16	0.85	0.85	1	1.06	1	1.07	MAX	1	0.91	1	0.9	0.95	NOM
3	BUS	77	PLI	1	1.16	0.85	1	1	1	0.87	0.94	MID	0.86	0.82	0.86	0.9	0.95	NOM
4	BUS	79	COB	0.75	1.16	0.7	0.76	1	1	0.87	1	MID	1.19	0.91	1.42	1	0.95	NOM
5	BUS	69	FTN	0.88	0.94	1	1	1	1	0.87	1	MAX	1	1	0.86	0.9	0.95	NOM
6	BUS	74	PLI	0.75	1	0.85	1	1	1.21	1	1	MID	1.46	1	1.42	0.9	0.95	HI
7	BUS	79	COB	0.75	1	1	1	1	1	0.87	0.87	MAX	1	1	1	0.9	0.95	HI
8	CTL	73	MOL	1.15	0.94	1.3	1	1.66	1.56	1.3	1	MIN	0.71	0.91	1	1.21	1.14	NOM
9	CTL	78	FTN	1.15	0.94	1.3	1	1.3	1.21	1.15	1	MIN	0.86	1	0.86	1.1	1.07	NOM
10	CTL	75	MOL	1.4	0.94	1.3	0.63	1.11	1.56	1	1.07	MIN	0.86	0.82	0.86	0.9	1	NOM
11	CTL	76	MOL	1.4	0.94	1.3	0.63	1.11	1.56	1	1.07	MIN	0.86	0.82	0.86	0.9	1	NOM
12	CTL	78	JOV	1.15	0.94	1.3	1	1.11	1.06	1	1	MIN	0.86	0.82	0.86	1	0.95	NOM
13	CTL	78	FTN	1.15	0.94	1.6	0.96	1.11	1.06	1.15	1	MID	0.71	1	0.7	1.1	1	HI
14	CTL	77	MOL	1.15	0.94	1.65	1	1.3	1.56	1.15	1	MIC	0.86	1	0.7	1.1	1.07	NOM
15	CTL	76	MOL	1.4	0.94	1.3	1	1.3	1.06	1.15	0.87	MIN	0.86	1.13	0.86	1.21	1.14	NOM
16	CTL	75	MOL	1.4	1	1.3	0.6	1.3	1.56	1	0.87	MIN	0.86	1	0.86	1	1	NOM
17	CTL	77	MOL	1.4	1	1.3	0.53	1.3	1.56	1	0.87	MIN	0.86	0.82	0.86	1	1	NOM
18	HMI	70	FTN	1.15	1.16	1.15	1	1.3	1.21	1	1.07	MAX	0.86	1	1	1	1	LO
19	HMI	78	FTN	1.15	1.08	1	0.84	1.11	1.21	0.87	0.94	MAX	0.71	0.91	1	1	1	NOM
20	HMI	74	JOV	1.4	1.08	1.3	0.96	1.11	1.21	1.15	1.07	MAX	0.71	0.82	1.08	1.1	1.07	NOM
21	HMI	77	PLI	1	1.16	1.15	1	1.06	1.14	0.87	0.87	MAX	0.86	1	1	1	1	HI
22	HMI	76	JOV	1.15	1	1	0.92	1.27	1.06	1	1	MAX	0.86	0.82	0.86	0.9	1	HI
23	HMI	78	HOL	1.15	1	1	0.98	1.08	1.06	1	1	MAX	0.86	0.82	0.86	0.9	1	LO
24	HMI	79	FTN	0.88	1	0.85	1	1.06	1.06	1	0.87	MIN	1	1.29	1	1.1	0.95	NOM
25	HMI	78	JOV	1.15	1.16	1.3	1	1.15	1.06	1	0.87	MAX	0.86	1	0.86	1.1	1	NOM
26	HMI	77	HOL	0.94	1	0.85	1	1.07	1.06	1.15	1.07	MAX	0.86	1	0.86	1.1	1	NOM
27	HMI	77	FTN	1.15	0.94	1.15	1	1.35	1.21	1	0.87	MIN	1	1	1	1	1	HI
28	HMI	72	MOL	1.15	1.08	1.3	1	1.11	1.21	1.15	1.07	MIN	0.86	1	0.86	1.1	1.07	NOM
29	HMI	79	FTN	0.88	1	1	1	1	1	1	1	MAX	1.1	1.29	0.86	1	1	HI
30	HMI	79	PSC	0.88	1	1	1	1	1	1	1	MAX	1	1.29	0.86	1	1	HI
31	SCI	75	MOL	1.4	1.08	1	0.81	1.48	1.56	1.15	1.07	MIN	0.86	0.82	0.86	1.1	1.07	NOM
32	SCI	72	FTN	0.88	1.08	0.85	0.67	1	1	1	1	MAX	0.71	0.82	1	1	1	NOM
33	SCI	76	FTN	1.4	1.08	1.3	0.96	1.48	1.56	1.15	0.94	MIN	0.86	0.82	0.86	0.9	1	HI
34	SCI	77	FTN	1.15	1.08	1	0.96	1.06	1	1	0.87	MIN	1	1	1	1	1	NOM
35	SCI	68	MOL	0.75	0.94	1.3	1	1.06	1.21	1.15	1	MID	1	0.91	1	1.1	1	NOM
36	SCI	79	FTN	0.88	1.08	0.85	0.81	1	1	0.87	0.87	MAX	1.19	1	1.17	0.9	0.95	NOM
37	SCI	78	FTN	0.88	0.94	0.7	0.56	1	1.06	1	1	MIN	0.86	0.82	0.86	1	1	NOM
38	SCI	77	PLI	1	1	1.15	1	1	1	0.87	0.87	MAX	0.71	0.91	1	0.9	0.95	NOM
39	SCI	64	FTN	1	1	1.15	1	1	1	0.87	1	MAX	0.71	0.82	0.7	1	0.95	HI
40	SCI	74	MOL	1	0.94	1.3	0.83	1	1	1	0.87	MIN	0.86	0.82	1.17	1	1	NOM
41	SCI	76	FTN	0.88	0.94	1	1	1	1	0.87	0.87	MAX	1	0.82	0.7	0.9	0.95	HI
42	SCI	78	FTN	0.88	1.04	1.07	0.43	1	1.06	0.87	1.07	MAX	0.86	1	0.93	0.9	0.95	HI
43	SCI	78	FTN	1	1.04	1.07	0.98	1	1.21	0.87	1.07	MAX	0.86	1	1	0.9	0.95	HI
44	SCI	78	FTN	0.88	1.04	1.07	0.98	1.06	1.21	0.87	1.07	MAX	1	1	1	0.9	0.95	HI
45	SCI	77	FTN	0.88	1.04	1.07	0.91	1	1.06	0.87	1.07	MAX	1	1	1	0.9	0.95	HI
46	SCI	78	FTN	0.88	1.04	1.07	0.78	1	1.06	0.87	1.07	MAX	1	1	0.86	0.9	0.95	NOM
47	SCI	78	FTN	0.75	0.94	1.3	1	1	1	0.87	0.87	MAX	0.71	0.82	0.7	1.1	1.07	NOM
48	SUP	76	FTN	0.88	1	0.85	0.67	1	1	0.87	1	MAX	1.19	0.91	1.17	0.9	0.95	NOM
49	SUP	76	JOV	1	1	0.85	1	1	1	1	0.87	MID	0.71	1	0.7	1.1	1	HI
50	SUP	76	MOL	1.15	1	1	1	1.3	1.21	1	0.87	MIN	0.86	1	0.86	1.1	1	NOM
51	SUP	70	COB	0.88	0.94	1	1	1	1	1	1.15	MAX	1.19	1	1.42	1	0.95	LO
52	SUP	71	MOL	0.88	0.94	0.85	1	1	1.06	1.15	1	MIN	1	1	1	1.1	1.07	NOM
53	SUP	78	MOL	0.88	0.94	1.15	1	1.11	1.21	1.3	1	MIC	0.71	1	0.7	1.1	1.07	NOM
54	SUP	78	FTN	1	0.94	1	1	1	1.06	1.15	0.87	MIC	1	0.82	1	1	0.95	HI
55	SUP	72	FTN	0.88	0.94	0.7	1	1	1	0.87	0.87	MAX	0.86	0.82	1.17	0.9	0.95	NOM
56	SYS	71	MOL	1.15	0.94	1.3	1	1.3	1.21	1	1	MAX	0.86	0.91	1	1.1	1.07	NOM
57	SYS	74	MOL	1	0.94	1.15	0.87	1.11	1.21	1.3	1	MIN	1	1	1	1.1	1.07	LO
58	SYS	76	MOL	1.4	0.94	1.3	1	1.66	1.21	1	1	MIN	0.71	1	0.7	0.9	0.95	NOM
59	SYS	77	HOL	1	0.94	1.15	0.9	1.06	1.06	1	0.87	MID	1	1	1	1	1	NOM
60	SYS	73	MOL	1.15	0.94	1.3	1	1.11	1.06	1	1	MAX	0.86	1	0.86	1.1	1.07	NOM
61	SYS	78	PSC	1	0.94	1.15	1	1	1	0.87	0.87	MAX	0.86	1	0.86	0.9	1	HI
62	SYS	78	MOL	0.88	0.94	1.3	1	1.11	1.21	1.15	1	MIC	0.78	0.82	0.7	1.21	1.14	NOM
63	SYS	79	MOL	1	0.94	1.15	1	1	1	1	0.87	MIN	0.71	0.82	0.86	1	1	NOM

PROJECT ATTRIBUTES				MAN-MONTHS			
MDOP	TOOL	SCED	RVOL	NOM	EST	ACT	(E-A)/A
1.24	1.1	1.04	1.19	814	2218	2040	9
1.1	1	1	1	2102	1770	1600	11
0.91	0.91	1	1	711	245	243	1
1.24	1	1.04	1.19	178	212	240	-12
1.24	1	1	1	89	39	33	18
1.24	1.1	1	1.19	13.7	30	43	-30
0.91	0.91	1	1	24	9.8	8	23
1.1	1.1	1.08	1.38	114	869	1075	-19
0.91	1	1	1.19	166	397	423	-6
1	1	1	1.38	90	214	321	-33
1	1	1	1.38	102	243	218	11
0.91	1	1.08	1.19	213	238	201	18
0.82	1	1	1	127	108	79	37
1.1	1.24	1.23	1.19	10.3	60	73	-18
0.91	1	1.23	1.19	14.3	52	61	-15
1	1	1	1.19	13.5	38	40	-5
1	1	1	0.91	6	10.7	9	19
1.24	1.1	1.08	1.19	2840	11056	1400	690
0.91	0.91	1	1	10694	7764	6600	18
1.24	1	1.08	1.19	1698	6536	6400	2
0.91	0.91	1	1	2132	1836	2455	-25
0.91	1	1.23	1	780	733	724	1
1	1	1.23	1	498	443	539	-18
0.82	0.83	1	1	463	326	453	-28
0.82	0.91	1.08	1.62	220	430	523	-18
0.91	1.1	1.08	1.19	292	339	387	-12
0.82	1.1	1.08	1.19	41	89	88	1
1.1	1.1	1	1	47	133	98	36
0.91	0.91	1.23	0.91	7	7	7.3	-4
0.91	0.91	1.23	0.91	6.4	5.8	5.9	-2
1	1	1	1	306	962	1063	-10
1.1	1.1	1	1	1527	869	702	24
0.91	0.91	1	1	234	529	605	-13
0.91	1.1	1.23	1.19	114	201	230	-13
1.24	1.24	1	1.19	61	161	82	96
1	0.91	1.04	1	49	33	55	-40
1	1	1	0.91	130	44	47	-6
0.82	0.91	1	1	55	20	12	67
0.91	1.1	1	1	22	8.4	8	5
1.1	1	1	1	8.4	8.1	8	1
0.91	0.91	1	1	18.4	4.7	6	-22
0.95	0.95	1.04	1	72	46	45	2
1	1	1.04	1	106	102	83	23
1.1	1	1.04	1	114	130	87	49
1	0.95	1.04	1	122	100	106	-6
1	1	1.04	1	223	166	126	32
1.1	1	1.04	1	86	33	36	-8
1.1	1	1.04	1.09	1858	1542	1272	21
0.82	0.91	1	1.19	469	168	156	8
1	1	1	1.19	127	193	176	10
1.24	1.1	1.04	1.38	36	114	122	-7
1.24	1.1	1	1.38	29	55	41	34
1	1.1	1.08	1	19.4	22	14	57
0.91	1.1	1	1.19	15	14	20	-30
1.1	1	1	1	22	7.5	18	-58
1.1	1.1	1.08	1.38	146	537	958	-44
1.1	1.1	1.23	1	72	239	237	1
0.91	1	1	1	133	145	130	12
0.91	1	1	0.91	78	68	70	-3
1.1	1.1	1.08	1.19	23.6	60	57	5
0.82	1	1	1	106	47	50	-6
0.91	1.24	1	1	36	42	38	11
0.82	1	1	1	44	17	15	13

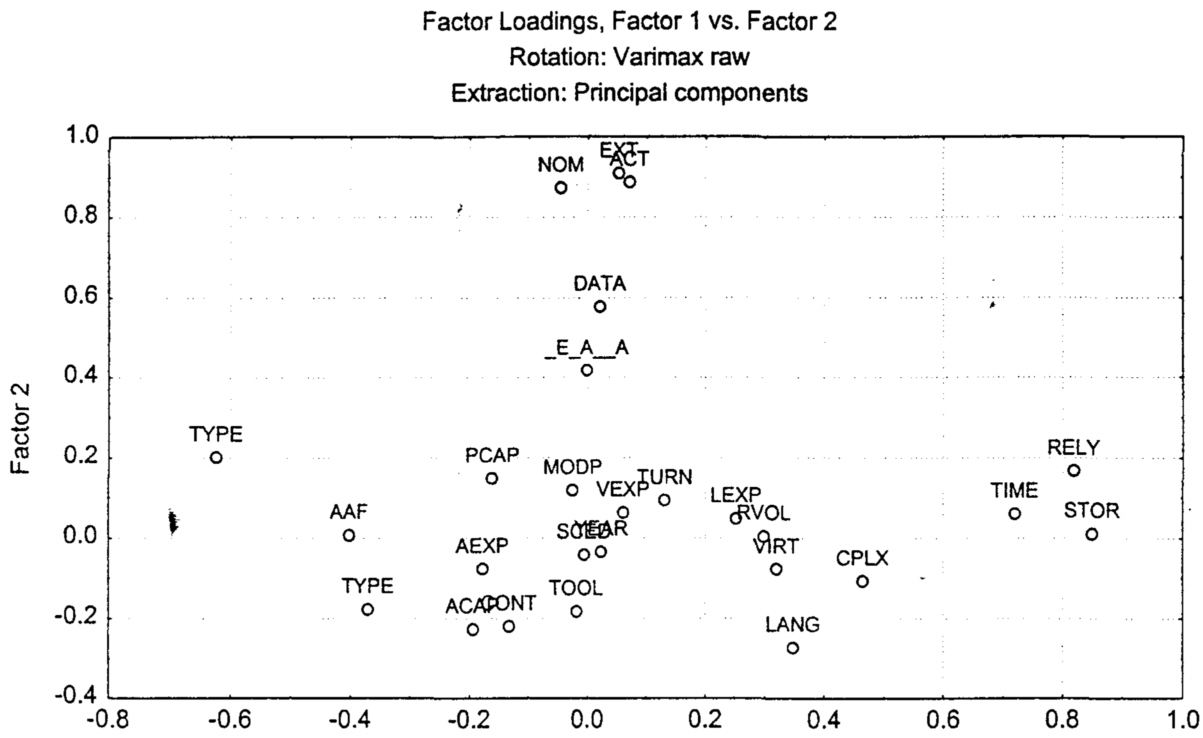

Factor Loadings, Factor 1 vs. Factor 2
Rotation: Varimax raw
Extraction: Principal components
Factor 2
Factor 1
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
NOM
EXT
ACT
DATA
TYPE
PCAP
MODP
VEXP
TURN
LEXP
RVOL
TIME
RELY
STOR
AAF
AEXP
SCED
YEAR
VIRT
CPLX
TYPE
TOOL
ACAP
CONT
LANG

Figure 2

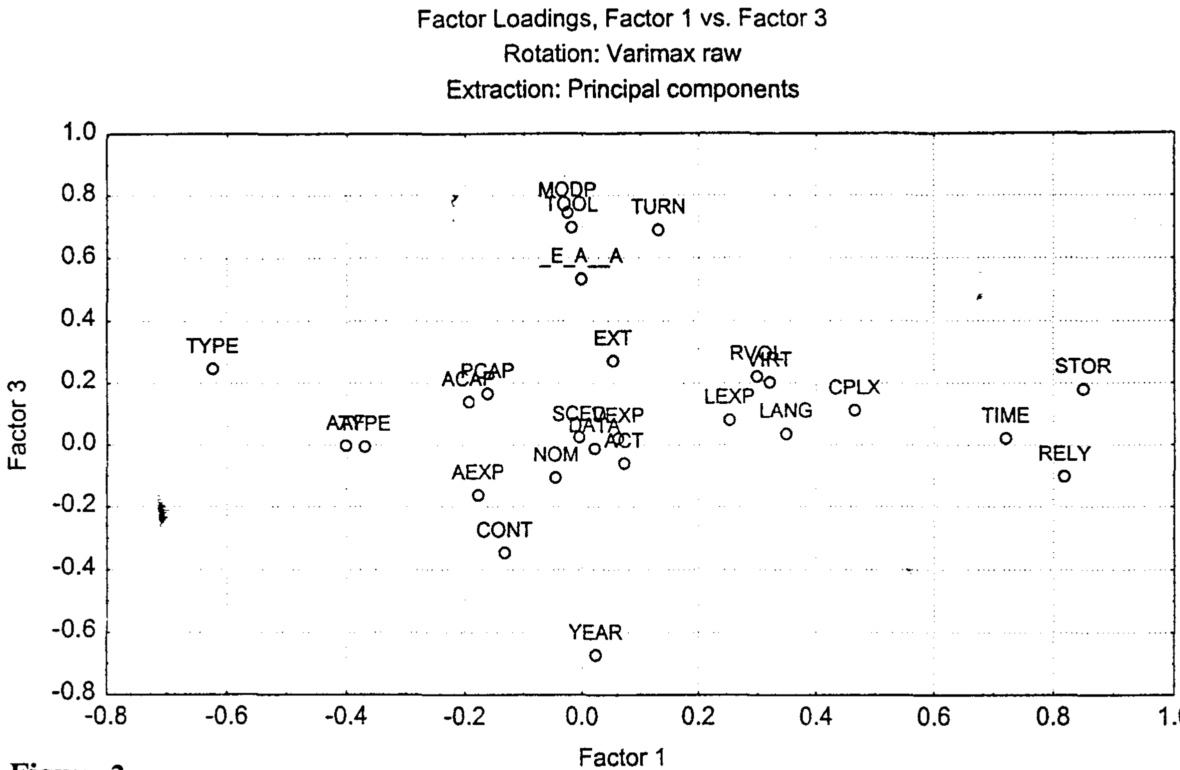
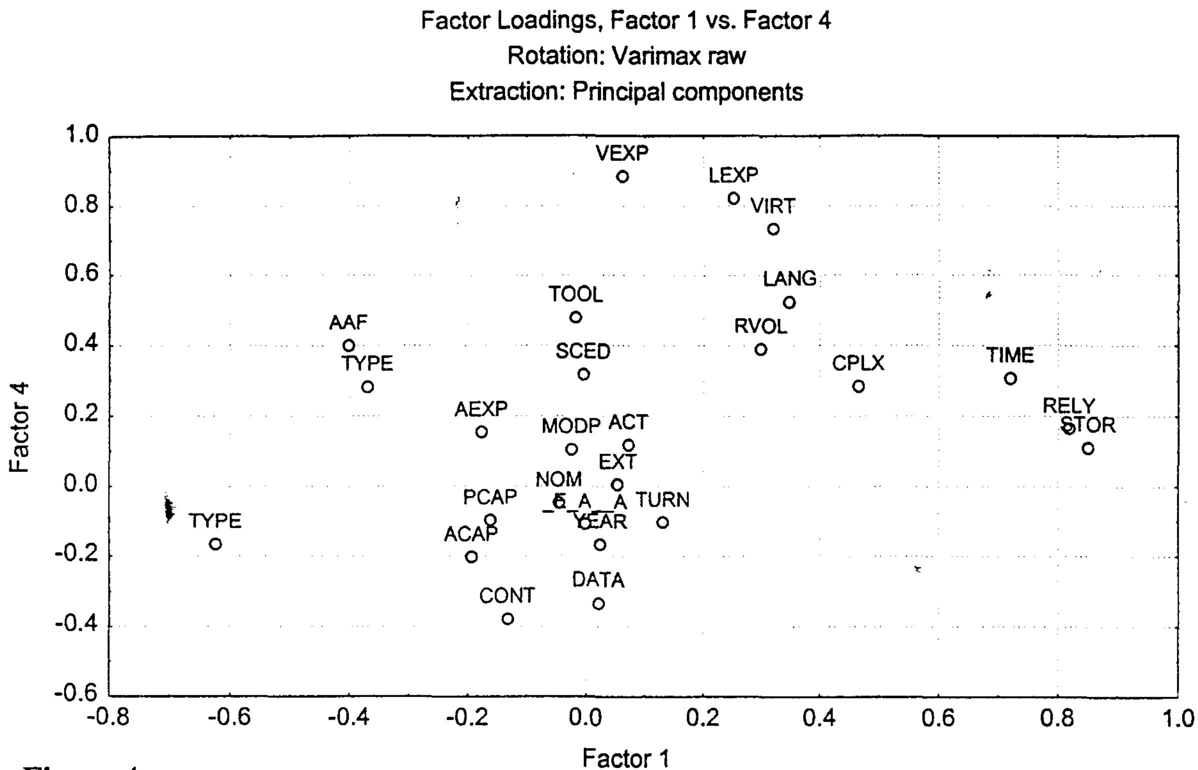
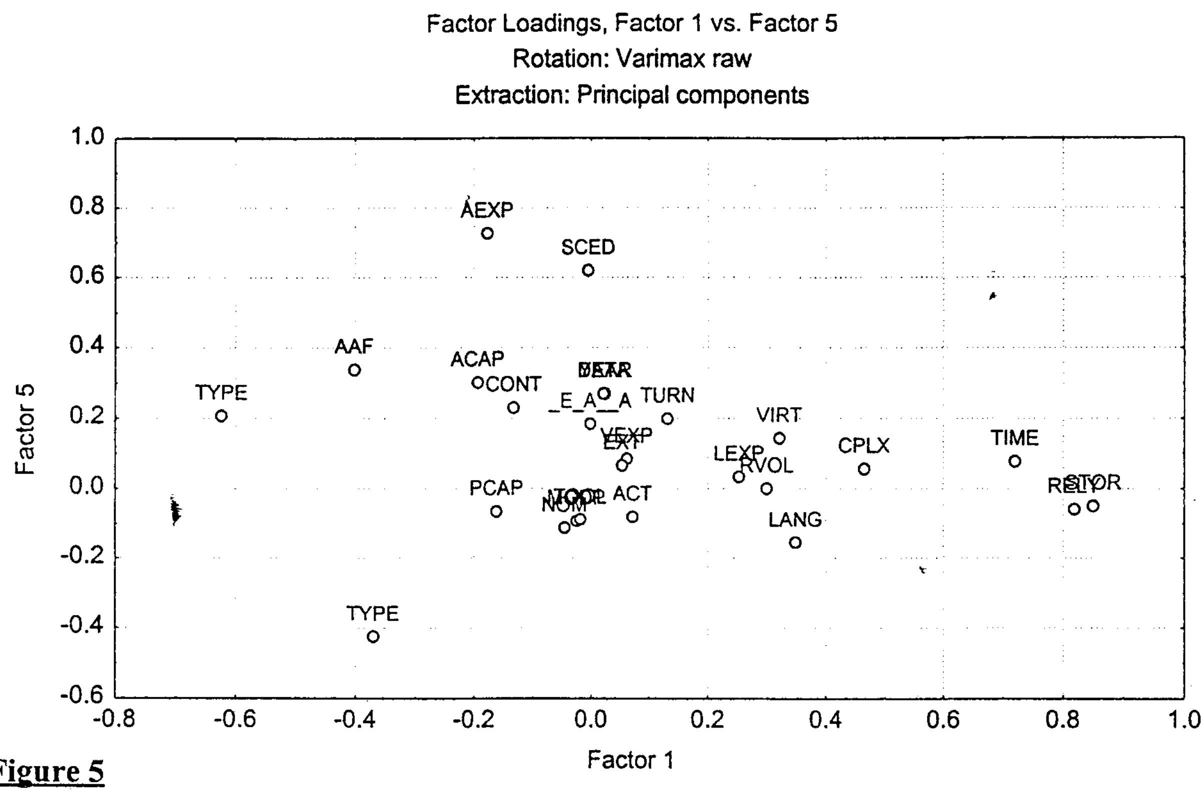

Factor Loadings, Factor 1 vs. Factor 3
Rotation: Varimax raw
Extraction: Principal components
Factor 3
Factor 1
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
MODP
TOOL
TURN
TYPE
EXT
STOR
CPLX
LEXP
LANG
TIME
RELY
NOM
AEXP
CONT
YEAR
ACT

Figure 3



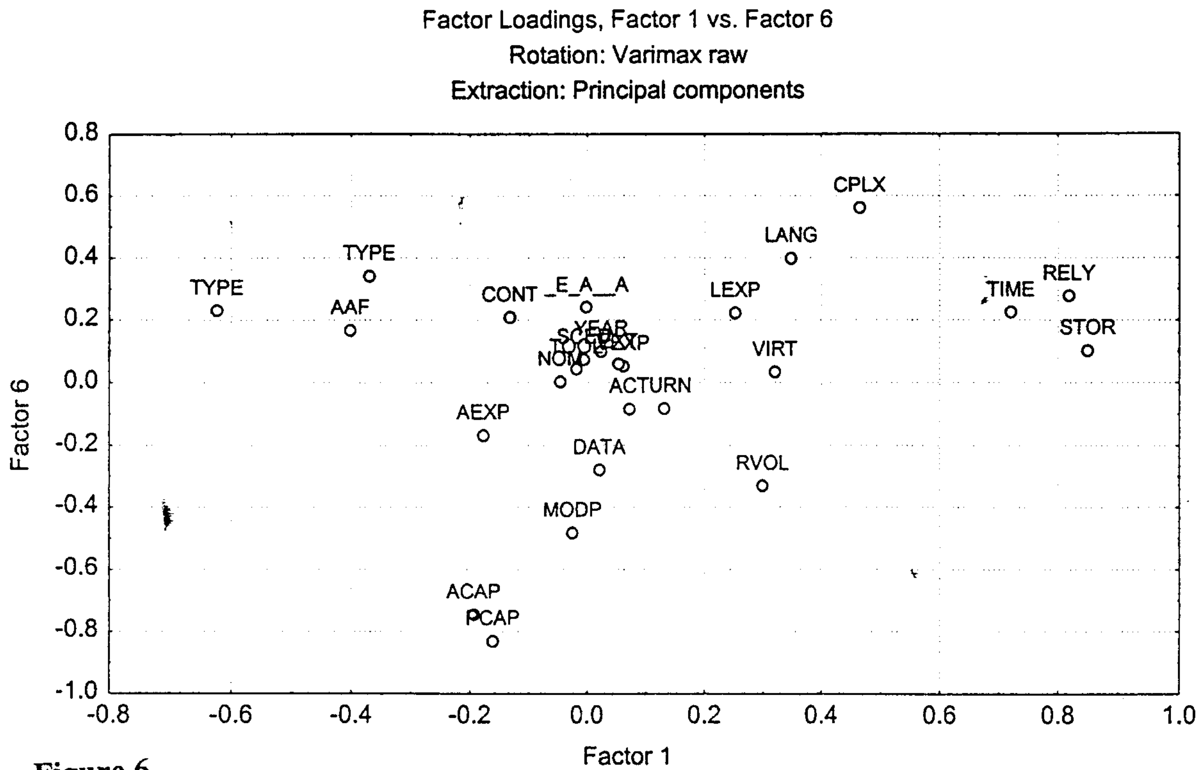
Factor Loadings, Factor 1 vs. Factor 4
Rotation: Varimax raw
Extraction: Principal components
Factor 4
Factor 1
VEXP
LEXP
VIRT
LANG
TOOL
AAF
RVOL
TYPE
SCED
CPLX
TIME
RELY
STOR
AEXP
MODP
ACT
EXT
NOM
PCAP
TURN
TYPE
YEAR
ACAP
DATA
CONT
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0

Figure 4



Factor Loadings, Factor 1 vs. Factor 5
Rotation: Varimax raw
Extraction: Principal components
Factor 5
Factor 1
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
AEXP
SCED
AAF
ACAP
TYPE
CONT
TURN
VIRT
CPLX
TIME
LEXP
RVOL
PCAP
ACT
LANG
TYPE

Figure 5



Factor Loadings, Factor 1 vs. Factor 6
Rotation: Varimax raw
Extraction: Principal components
Factor 6
Factor 1
TYPE
TYPE
AAF
CONT
E_A
A
YEAR
SCED
TOOL
PEXP
NOM
AEXP
ACTURN
DATA
MODP
ACAP
PCAP
LEXP
LANG
CPLX
VIRT
RVOL
TIME
RELY
STOR

Figure 6

Table 4

STAT. CLUSTER ANALYSIS	Amalgamation Schedule (egmt.sta) Single Linkage Euclidean distances						
linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7
.4791660	VEXP	LEXP					
.6299206	VEXP	LEXP	TOOL				
.6843975	VIRT	VEXP	LEXP	TOOL			
.7175654	VIRT	VEXP	LEXP	TOOL	SCED		
.7808329	VIRT	VEXP	LEXP	TOOL	SCED	TURN	
.7905062	DATA	VIRT	VEXP	LEXP	TOOL	SCED	TURN
.9342911	DATA	VIRT	VEXP	LEXP	TOOL	SCED	TURN
1.057781	ACAP	PCAP					
1.085818	DATA	VIRT	VEXP	LEXP	TOOL	SCED	TURN
1.111261	TIME	STOR					
1.262537	RELY	TIME	STOR				
1.263883	DATA	VIRT	VEXP	LEXP	TOOL	SCED	TURN
1.276127	DATA	VIRT	VEXP	LEXP	TOOL	SCED	TURN
1.321401	DATA	VIRT	VEXP	LEXP	TOOL	SCED	TURN
1.434399	RELY	TIME	STOR	DATA	VIRT	VEXP	LEXP
1.551290	RELY	TIME	STOR	DATA	VIRT	VEXP	LEXP
8.366600	TYPE2	CONT					
9.770302	RELY	TIME	STOR	DATA	VIRT	VEXP	LEXP
20.83267	LANG	RELY	TIME	STOR	DATA	VIRT	VEXP
214.7650	TYPE	YEAR					
569.5744	TYPE	YEAR	LANG	RELY	TIME	STOR	DATA
716.9180	TYPE	YEAR	LANG	RELY	TIME	STOR	DATA
6839.413	NOM	ACT					
9770.280	NOM	ACT	EXT				
10052.88	TYPE	YEAR	LANG	RELY	TIME	STOR	DATA

STAT. CLUSTER ANALYSIS	Amalgamation Schedule (egmt.sta) Single Linkage Euclidean distances						
linkage distance	Obj. No. 8	Obj. No. 9	Obj. No. 10	Obj. No. 11	Obj. No. 12	Obj. No. 13	Obj. No. 14
.4791660							
.6299206							
.6843975							
.7175654							
.7808329							
.7905062							
.9342911	MODP						
1.057781							
1.085818	MODP	AEXP					
1.111261							
1.262537							
1.263883	MODP	AEXP	ACAP	PCAP			
1.276127	MODP	AEXP	ACAP	PCAP	AAF		
1.321401	MODP	AEXP	ACAP	PCAP	AAF	RVOL	
1.434399	TOOL	SCED	TURN	MODP	AEXP	ACAP	PCAP
1.551290	TOOL	SCED	TURN	MODP	AEXP	ACAP	PCAP
8.366600							