

Analysis of ISSP Environment II Survey Data Using Variable Clustering

Loretta Davidson and Gongzhu Hu

Abstract Social informatics, as a sub-field of the general field of informatics, deals with processing and analysis of data for social studies. One of the social data repositories is the International Social Survey Program (ISSP) the provides cross-national surveys on various social topic. Previous studies of this data often used a subset of available variables and sometimes a reduced number of records, and most of these analyses have focused on predictive techniques such as regression. In this paper, we analyze the Environment II module of this data set using variable clustering to produce meaningful clusters related to questionnaire sections and provide information to reduce the number of demographic variables considered in further analysis. Case level clustering was attempted, but did not produce adequate results.

Key words: Social informatics, variable clustering, case level clustering

1 Introduction

The International Social Survey Program (ISSP) [2], started in 1984, is a continuing annual program of cross-national (currently members from 43 countries) collaboration on surveys covering topics important for social science research. It contains annual modules that provide cross-national data sets on topics of social interests. Some of the most surveyed topics are listed in Table 1.

Loretta Davidson

Data Mining Program, Central Michigan University, Mt. Pleasant, MI 48859, USA. e-mail: davidl1j@cmich.edu

Gongzhu Hu

Department of Computer Science, Central Michigan University, Mt. Pleasant, MI 48859, USA. e-mail: hu1g@cmich.edu

Table 1 ISSP Modules (1985 – 2012)

Topic	Years
Role of government	1985, 1990, 1996, 2006
Family, changing gender roles	1988, 1994, 2002, 2012
Religion	1991, 1998, 2008
Environment	1993, 2000, 2010

Environment was the subject of the 1993 and 2000 ISSP modules and is currently the topic for 2010 [1]. This paper is concerned with analyzing the Environment II (2000) data set using data mining techniques of variable clustering and case clustering. Previous analyses of the ISSP Environment 1993 and/or 2000 data have applied correlation, regression, factor analysis, and structural equation modeling techniques on a subset of the available variables [3, 4, 6, 7, 11]. In addition, these studies have focused on causal relationships. The analysis presented in this paper seeks to maximize the use of input variables by allowing variable clustering to segment variables and choose the most important variable from each cluster. Rather than taking a predictive modeling approach, this paper focuses on pattern recognition by implementing clustering of case level data.

2 Data Description

The ISSP Environment II data set, description, questionnaires, and codebook are available at the Interuniversity Consortium for Political and Social Research (ICPSR) at University of Michigan [8]. The raw data set consists of 31,402 cases and 209 variables. ISSP data is also accessible via GESIS Leibniz Institute for the Social Sciences that provides the variable overview, source questionnaire, and monitoring report documents. The Environment II questionnaire consists of 69 questions covering the following areas as stated in the overview and source questionnaire [5]:

- Left-right dimension
- Postmaterialism
- Attitudes towards science and nature
- Willingness to make trade-offs for environment
- Environmental efficacy
- Scientific and environmental knowledge
- Dangers of specific environmental problems
- Environmental protection, locus of control, effort
- Positive trade-off of environmentalism
- Trust information sources on causes of pollution
- Respondent's behaviors and actions to protect the environment
- Belief about God

- Type of area where respondent lives
- Grid-group theory

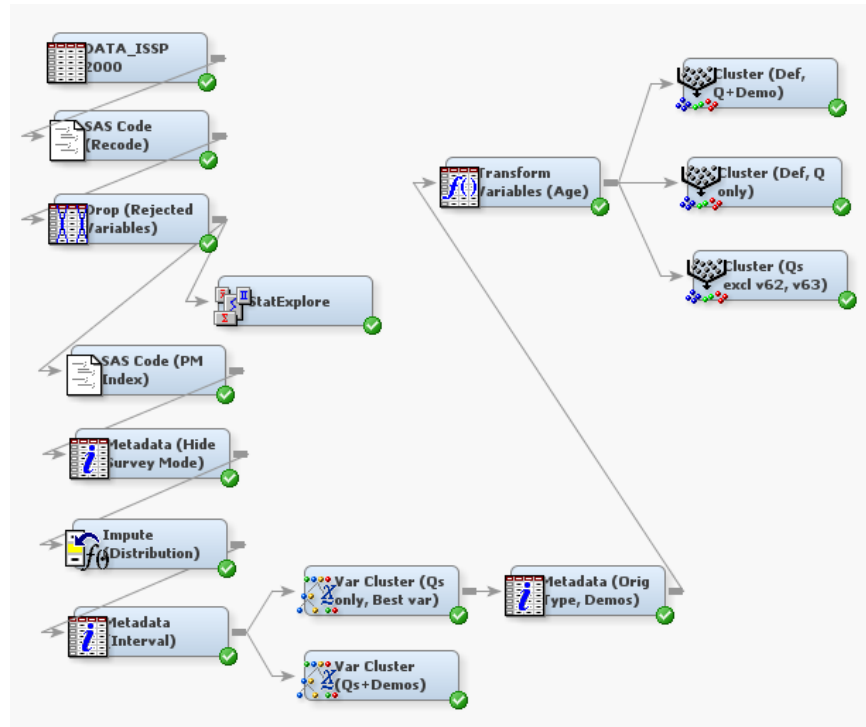


Fig. 1 SAS Enterprise Miner Diagram for ISSP Environment II data analysis.

Most of the questions use a five-point or four-point ordinal scale although some are binary or nominal. Of the 209 variables, 118 were excluded from our analysis because they were country specific or for reasons related to usability, comparability, and redundancy. The remaining variables include 69 questionnaire variables, 19 demographic variables, a respondent ID, a weight variable, and a survey mode variable.

3 Methodology

We used the SAS Enterprise Miner software to run our analysis that consists of three parts: data preparation, variable clustering, and case clustering. The Enterprise Miner diagram developed for the analysis is shown in Fig. 1. In this diagram, each box is called a *node* representing a software module to perform a specific task. An

arrow from node *A* to node *B* means the output produced by *A* is fed as an input to node *B*. We shall discuss the methods of each part in detail in this section.

3.1 Data Cleaning and Preparation

Initial data cleaning focused on recoding problematic variables. The knowledge questions for Chile contained an extra response category not in the master questionnaire. This creates problems for comparison with other countries and the distribution of these variables differs from other countries. Since merging the incorrect category with another category would distort the data, the responses for Chile for this set of variables were removed. Spain included an extra category for three questions concerning the role of the individual, government, and business. The extra category was “neither” which seemed equivalent to not answering the question, so these values were recoded as missing. The last recode involved a demographic variable for the number of people in the household. Categories for 6 or more people varied by country, so all categories for 6 people and above were collapsed into one category. After the initial recoding, the **Drop** node was used to remove the 118 excluded variables from the data set.

A second SAS Code node was used to create a Postmaterialist index following the method described in [6]. Postmaterialism values are assessed using two questions which ask the respondent to choose the 1st and 2nd highest priorities for his/her country from a list of four choices:

1. Maintain order in the nation (Materialist)
2. Give people more say in government decisions (Postmaterialist)
3. Fight rising prices (Materialist)
4. Protect freedom of speech (Postmaterialist)

The combination of a person’s rankings of the 1st and 2nd highest priorities indicates his/her materialist or postmaterialist preferences and is coded as an integer 1 – 4 shown in Table 2. The code, as the value of a new variable **PM_INDEX**, is interpreted as

1. Strong materialist
2. Weak materialist
3. Weak postmaterialist
4. Strong postmaterialist

It replaces the answers to the two original questions.

The next step of the data preparation uses a **Metadata** node to change variable roles and types. In this case, metadata was used to hide the survey mode variable which captures how the survey was administrated. Although this variable was not used in the analysis, we kept it in the data set for possible future use.

Table 2 postmaterialist index

Highest Priority		Code
1st	2nd	
Materialist	Materialist	1
Materialist	Postmaterialist	2
Postmaterialist	Materialist	3
Postmaterialist	Postmaterialist	4

The **Impute** and second **Metadata** nodes were used to prepare the data for the **Var Cluster** node to perform variable clustering. The imputation method used for both interval and class (ordinal, nominal) data computes random percentiles of a variable's distribution to replace missing data. The **Metadata** node following the **Impute** node changes all of the variable types to interval. The **Var Cluster** node manages non-interval inputs by creating a new variable for each nominal or ordinal level. It was suggested changing variable types when many of the inputs are non-interval since the efficiency of the algorithm diminishes when there are over 100 variables [12].

3.2 Variable Clustering

Variable clustering is a method for segmenting variables into similar groups. It can be a useful tool for variable reduction. We uses two **Var Cluster** nodes each with its own objective. The **Var Cluster** node labeled "Qs only" is used to assess whether the questionnaire variables group according to topic. If so, the best variable from each variable cluster can be exported for use in case level clustering. The second **Var Cluster** node labeled "Qs+Demos" is used to analyze how the demographic variables behave when clustered with the questionnaire variables. If demographic variables pair with each other or questionnaire variables, perhaps some can be excluded from further analysis. Using variable clustering prior to case clustering will reduce the number of variables used in cluster analysis.

The variable clustering algorithm is a divisive hierarchical clustering technique that aims to maximize the explained variance [12]. At each stage the weakest cluster is selected for splitting until the stopping criteria is reached. Correlation was used for the clustering source matrix and variation proportion equal to 0.40 was the stopping criteria. The algorithm continues to split clusters until the variation explained by each cluster is at least equal to the stopping criteria.

The best variable from each cluster was exported from the **Var Cluster** "Qs only" node. The best variable has the lowest $1 - R^2$ ratio. In the following **Metadata** node, these variables are changed back to their original type and the demographic variables added to the analysis are changed to have an input role. At this stage, *Age* was the

only continuous variable. The Transform Variables node changes it into a new nominal variable using 10 quantiles.

3.3 Clustering

Table 3 Clusters for Questionnaire Variables

Cluster	Description	Number of Variables.
1	Perceived threats	8
2	Environmental concern (unfavorable)	3
3	Trust in information sources except business/industry	5
4	Knowledge	3
5	Willingness to sacrifice	4
6	Actions (binary - group, money, petition, demonstration)	4
7	Attitudes - sacredness nature; God; science (unfavorable)	4
8	Locus of control and effort (Govt to others)	3
9	*Mixed* - country effort, trust bus/ind, private enterprise	3
10	Values/Grid-Group (egalitarian, anti-individualism)	3
11	Actions (ordinal - recycling, cut back on driving)	2
12	Attitudes - science and economic growth (favorable)	2
13	Perceived threat - nuclear power stations	2
14	*Mixed* - evolution T/F, population growth, intl agreements	3
15	Attitudes - modern life and economic growth (unfavorable)	3
16	Values/Grid-Group (egalitarian, communitarian)	2
17	Values/Grid-Group (hierarchy)	2
18	Demographic - describe where you live	1
19	Environmental Efficacy and Values/Grid-Group (fatalism)	4
20	Effort - business/industry, people, both	1
21	Effort - people, government, both	1
22	Knowledge - antibiotics kill bacteria not viruses	1
23	Values/Grid-Group - world is getting better	1
24	*Mixed* - animal testing and poor countries less effort	2
25	Values - Materialist-Postmaterialist index	1

The Ward method was used in the Cluster node. It follows the traditional hierarchical clustering algorithm that starts with n clusters for the n data records where each cluster containing a single data record. The algorithm repeatedly *merges* two clusters selected based on an objective function until a specified measure is reached (such as a desired number of clusters, or a merging threshold). The objective function used in the Ward algorithm is to minimize the *total error sum of squares*

when clusters C_a and C_b are selected to be merged, given in Equation (1).

$$\mathcal{F}(C_a, C_b) = ESS(C_{ab}) - ESS(C_a) - ESS(C_b) \quad (1)$$

where C_{ab} is the combined cluster after merging C_a and C_b , and $ESS(\cdot)$ is the *error sum of squares* of a given cluster, as defined in Equation (2) for the data set C .

$$ESS(C) = \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{m}_C\|^2 \quad (2)$$

where $n = |C|$, $\mathbf{x}_i \in C$, and \mathbf{m}_C is the mean of the data records in C .

The **Cluster** node defaults apply rank encoding for ordinal data and GLM encoding for nominal data. Rank encoding uses the adjusted relative frequency for each level to create a new variable on a 0 to 1 scale. The nominal encoding technique create a new dummy variable for each level of the original variable. The initial cluster seed method was left as default which is the k -means algorithm [10].

Three **Cluster** nodes were run in the analysis: questionnaire variables and demographic variables, questionnaire variables only, questionnaire variables excluding two that relate to demographic information.

4 Results

4.1 Variable Clustering

The variable clustering for questionnaire and demographic variables included 86 input variables and produced 35 variable clusters. The total proportion of variation was 0.5582. Five clusters contained a single demographic variable which may indicate more importance or independence. Four clusters contained both demographic and questionnaire variables, while three clusters were all demographic variables. After reviewing the results, eight demographic variables were excluded from further analysis which left ten remaining demographic variables.

The variable clustering of questionnaire variables contained 68 input variables and produced 25 variable clusters. The total proportion of variation explained was 0.5333. Most clusters are centered on a questionnaire topic and variables had similar type. A description for each variable cluster is in Table 3. Three clusters contained variables of mixed types without a clear topic connection.

The make-up of mixed groups can be further investigated by using cross tabulation to find the intersection of variable levels with the highest frequency. For example, cluster 24 contained the attitude questions “It is right to use animals for medical testing if it might save human lives” and “Poorer countries should be expected to make less effort than richer countries to protect the environment” [5]. Both questions used a 5-point Agree/Disagree response scale. The cross tabulation reveals that 15% of respondents are in the “Agree (2)” category for both questions

and that 13% “Agree (2)” to the animal testing, but “Disagree (4)” to less effort for poorer countries. Thus, this cluster can be described as agreement to animal testing with conflicted agreement for poor countries making less effort.

For each variable, the `Var Cluster` node output provides statistics for R^2 with own cluster, R^2 with next cluster, and $1 - R^2$ ratio. These statistics can be examined with plots to assess the strength or weakness of variables in a cluster. Fig. 2 shows R^2 with next cluster plotted against R^2 with own cluster. A high R^2 with own cluster and low R^2 with next cluster is desired.

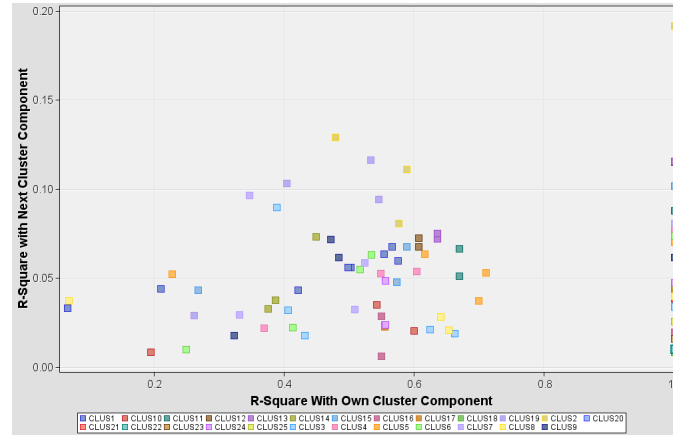


Fig. 2 R^2 with next vs. R^2 with own

Fig. 3 shows R^2 with next cluster plotted against $1 - R^2$ ratio. Lower $1 - R^2$ ratio indicates a better fit within the cluster.

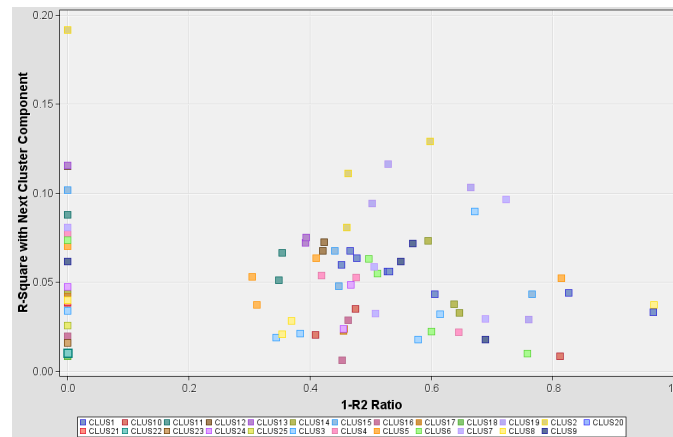


Fig. 3 R^2 with next vs. $1 - R^2$ ratio

The Var Cluster node results provide a cluster plot, shown in Fig. 4, which visualizes the relationship between clusters and variables according to relative distance, size, and orientation. Highlighted in the lower right side of Fig. 4 are clusters for willingness to sacrifice (cluster 5), actions (clusters 6 and 11), and postmaterialist values (cluster 25). This may support the research in [6] that identifies relationships between values, willingness to sacrifice, and actions.

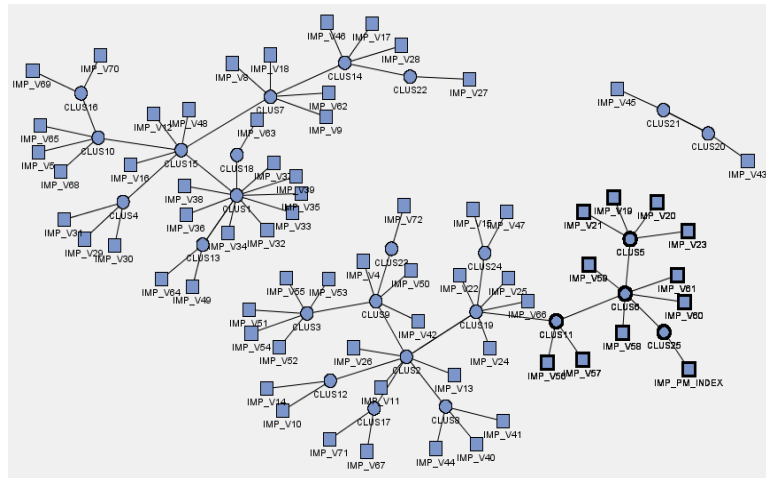


Fig. 4 Cluster Plot.

4.2 Case Clustering

A total of 35 variables were considered in the case level cluster analysis. Questionnaire variables exported from the variable clustering accounted for 25 of the variables and 10 were supplementary demographic variables. Of these 35 variables, 18 were ordinal, 14 were nominal, and 3 were binary.

The first Cluster node with both demographic and questionnaire variables produced poor results. The algorithm terminated at the default maximum number of clusters instead of finding an optimal number of clusters. The cubic clustering criterion plot in Fig. 5 plots Cubic Clustering Criterion (CCC) against the number of clusters.

The plot does not show a local CCC peak which would indicate an optimal number of clusters. The cubic clustering criterion “is obtained by comparing the observed R^2 to the approximate expected R^2 using an approximate variance-stabilizing transformation” [12]. The output also provides information from decision tree modeling with the cluster segment as the target variable. The variable importance output showed that of the 13 variables with importance greater than 0.5, 8 were

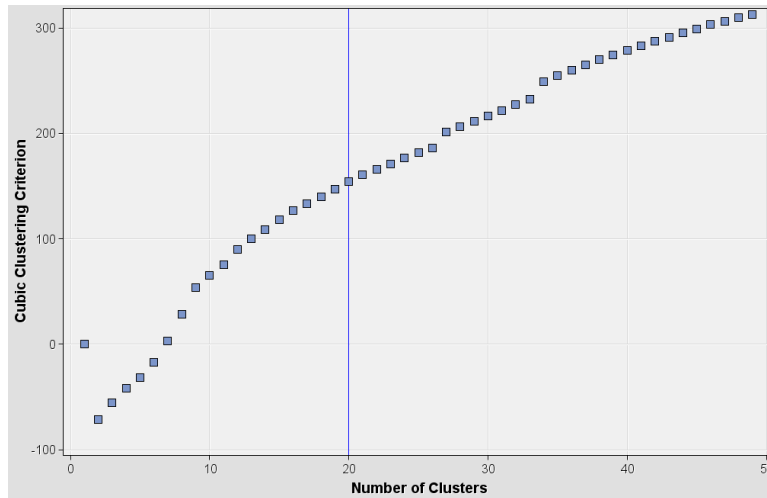


Fig. 5 Cubic Clustering Criterion Plot.

demographic variables. The variable importance measure is computed using the sum of squares error with the interpretation that values closer to one indicate greater importance.

A second **Cluster** node was run with supplementary demographic variables excluded from the analysis. Again, the algorithm terminated at the default maximum number of clusters and did not provide an optimal number of clusters. The CCC plot did not display any peaks. The variable importance output, which only had 3 variables with importance greater than .50, consisted of two binary variables and the nominal Postmaterialist index variable.

A third **Cluster** node was run which dropped two additional variables from the analysis. These two variables were part of the questionnaire but pertained to subjects that could be considered demographic information: belief in God and description of where the respondent lives. Although these variables did not reveal an important influence, I was interested in the results of cluster analysis only containing variables related to environmental topics. The results were the same as for the previous two cluster analyses. It is interesting to note that the top 5 variables according to variable importance consisted of all the binary and nominal input variables. These variables are described in Table 4.

A rerun of the last cluster node with user-specified 7 clusters produced clusters that can be described mainly in terms of the variables listed in Table 4; however, clusters defined by these inputs do not provide novel or insightful distinctions between groups.

Table 4 Important Variables in Cluster Analysis

Variable	Description	Importance
IMP_PM_INDEX	Materialist-Postmaterialist Index	1
IMP_V45	More effort to look after enviro: People or Govt or Both equally	0.70894
IMP_V60	Last 5 yrs given money to an enviro group: Yes/No	0.69236
IMP_V43	More effort to look after enviro: Business/Industry or People or Both equally	0.67948
IMP_V41	Enviro protection by Business/Insustry: Decide themselves or Govt pass laws	0.52280

5 Related Work

There are many research perspectives regarding social science survey data. Some researchers are interested in how values influence the formation of attitudes while others want to know if attitudes can predict behavior. The ISSP Environment II survey data allows for a wide range of analyses by including questions related to values, cultural theory, attitudes, knowledge, risk perception, willingness to sacrifice, and actions.

A study by [7] concluded that pro-environmental attitudes were poor predictors of pro-environmental behaviors based on regression analysis of 1993 ISSP Environment I data for New Zealand. This analysis was from a marketing perspective which reasoned that the effort to measure attitudes is fruitless unless a strong causal link between attitudes and behaviors is established.

In the study [3], six questions from the 1993 ISSP Environment I survey were included to compare values and pro-environmental actions in Western (Netherlands, United States) and Asian (Japan, Thailand, Philippines) countries. This study used factor analysis on Schwartz general values and progress/environment preferences (ISSP questions) and studied the relationship between them. The value and progress/environment factors were then used to predict three sets of pro-environmental behaviors using regression analysis. Differences in both value construction and predictors of behavior were found between Asian and Western countries.

A complex analysis by [11] used structural equation modeling to study the relationships between values (Schwartz values, Postmaterialism) and attitudes reflecting environmental concern, between concern, perceived threats, perceived behavior control and willingness to sacrifice, between willingness to sacrifice and pro-environmental behavior, and between values and behavior. This study used Schwartz harmony value data in addition to ISSP Environment II data for all available countries. The social science background motivating the analysis includes Schwartz values, Postmaterialism, and Value-Belief-Norm theory.

A study by [6] of the 1993 Environment I data for Norway compared the predictive ability of Postmaterialism and cultural theory on attitudes. An index was

created from two questions to measure the degree of Postmaterialist preferences. Postmaterialism is a value theory developed by Inglehart that identifies a shift in individual values in countries with economic stability [9]. The values shift from being dependent on basic material security to quality of life issues and individual liberty [6, 3, 11]. The Norwegian questionnaire also included eight cultural theory questions that were grouped into four cultural bias dimensions. Cultural theory was developed by Douglas and is a typology that groups individuals according to “group” and “grid” dimensions. Group refers to the degree of membership in a social unit and grid refers to an individual’s ability to negotiate relationships [13]. There are four dimensions of cultural bias depending on whether the group and grid factors are strong or weak. These grid-group dimensions are labeled hierarchy, egalitarianism, individualism, and fatalism [13]. The study by [6] performed factor analysis on twenty-nine attitude questions in the ISSP Environment I data and the resulting factors were used as dependent variables in regression analysis with the Postmaterialist index and cultural bias dimensions as independent variables.

6 Conclusion

While the variable clustering produced meaningful clusters, the case clustering did not. The questionnaire variables are mostly ordinal type which may not be optimal for case level clustering in this situation if there are not marked distinctions among the response levels. Since the variable clustering produced good results, perhaps the cluster components could be used in predictive analysis with action or attitude components as dependent variables. Cluster components were not used in the present analysis since their interpretation is not as straightforward, but using cluster components in future analyses would maximize the contribution from all variables. In addition, this study only focused on the second wave of the ISSP Environment module. The strength of the variable clustering technique for this type of data could be assessed by repeating the analysis with 1993 data. One would need to consider the differences between the 1993 and 2000 questionnaires in comparing the results.

Overall, the variable clustering supports a distinction among questionnaire topics. Given this information, analyses should be broadened to include input variables from more areas instead of focusing on a handful of selected variables. Perhaps this would lead to a more comprehensive understanding of research results from social science survey data.

References

1. Archive and data. <http://www.issp.org/page.php?pageId=4>. International Social Survey Programme
2. International Social Survey Programme. <http://www.issp.org> (2010)

3. Aoyagi-Usui, M., Vinken, H., Kuribayashi, A.: Pro-environmental attitudes and behaviors: An international comparison. *Human Ecology Review* **10**(1), 23–31 (2003)
4. Franzen, A.: Environmental attitudes in international comparison: An analysis of the ISSP surveys 1993 and 2000. *Social Science Quarterly* **84**(2), 297–308 (2003)
5. ISSP Environment II data documentation. <http://www.gesis.org/en/services/data/survey-data/issp/modules-study-overview/environment/2000/> (2009)
6. Grendstad, G., Selle, P.: Cultural theory, postmaterialism, and environmental attitudes. In: R.J. Ellis, M. Thompson (eds.) *Culture Matters: Essays in Honor of Aaron Wildavsky*, pp. 151–168. Westview Press (1997)
7. Hini, D., Gendall, P., Kearns, Z.: The link between environmental attitudes and behaviour. *Marketing Bulletin* **6**, 22–31 (1995)
8. ICPSR study No. 4104 International Social Survey Program: Environment II, 2000. <http://www.icpsr.umich.edu> (2009)
9. Inglehart, R.: The silent revolution in post-industrial societies. *American Political Science Review* **65**, 991–1017 (1971)
10. Lloyd, S.P.: Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**(2), 129–137 (1981)
11. Oreg, S., Katz-Gerro, T.: Predicting proenvironmental behavior cross-nationally: Values, the theory of planned behavior, and value-belief-norm theory. *Environment and Behavior* **38**(4), 462–483 (2006)
12. SAS Institute Inc.: *Enterprise Miner Help Documents* (2009)
13. Thompson, M., Ellis, R., Wildavsky, A.: *Cultural Theory*, pp. 1–18. Westview Press (1990)