Chapter 10
Conjoint Use of Variables Clustering and PLS Structural Equations Modeling

Valentina Stan and Gilbert Saporta

Abstract In PLS approach, it is frequently assumed that the blocks of variables satisfy the assumption of unidimensionality. In order to fulfill at best this hypothesis, we use clustering methods of variables. We illustrate the conjoint use of variables clustering and PLS structural equations modeling on data provided by PSA Company (Peugeot Citroën) on customers' satisfaction. The data are satisfaction scores on 32 manifest variables given by 2,922 customers.

10.1 Clustering of Variables

There are two main methodologies: hierarchical methods and direct partitioning methods. Hierarchical methods are either agglomerative or divisive. Partitioning methods usually require that the number of groups should be defined beforehand and will not be used here.

A good partition is such that the variables of the same class are correlated as much as possible.

We will use here algorithms which provide clusters which are as unidimensional as possible, and where correlations between variables of the same clusters are larger than correlations between variables of different clusters. This means that blocks of variables should be as homogeneous as possible, but are not independent.

One may distinguish two cases, depending on whether the sign of the correlation coefficient is important or not (i.e. if negative values show a disagreement between variables).

V. Stan
Groupe ESSCA Angers, 1 Rue Lakanal – BP 40348, 49003 Angers, Cedex 01, France
e-mail: valentina_titu@yahoo.fr

G. Saporta
Conservatoire National des Arts et Métiers, Chaire de Statistique Appliquée, case 441, 292 rue Saint Martin, 75141 Paris, Cedex 03, France
e-mail: gilbert.saporta@cnam.fr

V. Esposito Vinzi et al. (eds.), *Handbook of Partial Least Squares*, Springer Handbooks of Computational Statistics, DOI 10.1007/978-3-540-32827-8_11,

10.1.1 Agglomerative Hierarchical Clustering Methods

10.1.1.1 Methods Derived from Clustering of Statistical Units (Nakache and Confais 2005)

Various dissimilarity measures can be used, based on the usual correlation coefficient like:

$1 - \mathrm{r}_{\mathrm{jj}'} \mathrm{or} 1 - |\mathrm{r}_{\mathrm{jj}'}|$ if the sign of the correlation is not important; $\mathrm{s}_{\mathrm{jj}'} = \cos^{-1}(\mathrm{r}_{\mathrm{jj}'})$.

Then we use the following strategies of aggregation: single linkage, average linkage, complete linkage, Ward's criteria etc.

10.1.1.2 The VARHCA Method (Vigneau and Qannari 2003)

Let $C_1, C_2 \ldots .C_k$ be k blocks (or clusters) of manifest variables and $Y_1, Y_2 \ldots .Y_k$ the standardized latent variables (first principal component) associated respectively with each cluster. Manifest variables are centred, but not necessarily standardized. The following hierarchical procedure aims at locally optimizing the criterion T defined by:

$$T = n \sum_{r=1}^{k} \sum_{j=1}^{p} \delta_{rj} \operatorname{cov}^2 \left(x_j, Y_r\right) \quad where \quad \delta_{rj} = \begin{cases} 1 \; if \; x_j \in C_r \\ 0 \; otherwise \end{cases}$$

– At the first level of the hierarchy, each variable forms a cluster by itself; then, $T_0 = \sum_{j=1}^{p} \operatorname{var}\left(x_j\right)$;

– At level i, one merges the two clusters giving the minimal variation of T:

$\Delta T = T_{i-1} - T_i = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A \cup B)}$ where $\lambda_1^{(A)}, \lambda_1^{(B)}, \lambda_1^{(A \cup B)}$ are the largest eigenvalues of the covariance matrices of the variables in clusters A, B and A $\cup$ B.

10.1.2 Cutting Trees

The resulting tree should be cut at a suitable level to get a partition. We use here a criterion of unidimensionality of the groups to obtain this cut. Starting from the root of the tree, we first realize a cut in 2 classes and verify the hypothesis of unidimensionality by using the Cronbach's α or the Dillon–Goldstein's ρ. If these values are close to 1, then the hypothesis of unidimensionality is accepted. Otherwise, we proceed to a cut at the following level of the tree, and so on. We repeat the procedure until we obtain classes satisfying the unidimensionality criteria.

10.1.3 Divisive Methods

SAS VARCLUS procedure is one of the best known. At first step one performs a PCA with all manifest variables. If there is only one principal component with an eigenvalue greater than 1, there is only one cluster.

Otherwise one considers the first two principal components: each manifest variable is associated with the principal component to which it is the closest, in regard to the squared linear correlation coefficient, thus forming two groups of variables. If the second eigenvalue of a group is greater than 1, this group is divided in turn, according to the same method, and so on, until each group has only one principal component.

10.2 Application to Structural Equation Modeling

Let p variables be observed upon n units. The p variables are partitioned in J subsets or blocks of k_j variables which are presumed to be pertinent for describing the phenomenon. Each of these blocks is designed to describe a theme of the general phenomenon. We shall designate these blocks by X_j and we shall consider them as matrices with dimension $(n \times k_j)$ (Tenenhaus et al. 2005).

In the following, we shall always suppose that each block is associated with only one latent variable (unidimensionality). In order to obtain unidimensional blocks, we propose to use some of the clustering methods, previously presented in Sect. 10.1. Therefore we can identify the blocks by the same name as their latent variable. The latent variable corresponding to the X_j block will be designated by ξ_j.

In the following, we study the specific case where there are no pre-defined causal relationships between the latent variables. We use the blocks obtained by each method to build the causality scheme.

With the help of experts we propose relationships between latent variables with the aim of explaining the general satisfaction of the customers, and we therefore establish the inner model. To choose the best model from many, we use the global quality criterion developed by Amato et al. (2004):

$$GoF = \sqrt{\overline{communality} \times \overline{R^2}}$$

where $\overline{communality}$ is the average of the communality of each block and measures the quality of the external model. $\overline{R^2}$ is the average of R^2 for each endogenous latent variable.

The R^2 measures the quality of the inner model and is calculated for each endogenous variable according to latent variables which explain it.

The software used is PLSX module of SPAD.

10.3 Practical Application

10.3.1 The Questionnaire

The data obtained are satisfaction scores scaled between 1 and 10 on 32 services for a car. 2,922 customers participated. Manifest variables are the followings (Table 10.1).

Table 10.1 Manifest variables

Variable	
General satisfaction	Sat01h
General quality	Sat02h
Quality–price ratio	Sat03h
Absence of small, irritating defects	Sat04h
Absence of noise and vibrations	Sat05h
General state of the paintwork	Sat06h
Robustness of commands, buttons	Sat33h
Solidity and robustness	Sat08h
Lock, door and window mechanisms	Sat09h
Inside space and seat modularity	Sat34h
Inside habitability	Sat11h
Dashboard: quality of materials and finishing	Sat12h
Insider: quality of mat. and finishing	Sat13h
Front seat comfort	Sat14h
Driving position	Sat15h
Visibility from driver's seat	Sat16h
Radio–CD-ROM	Sat17h
Heating–ventilation	Sat18h
Boot capacity	Sat19h
Security	Sat20h
Braking	Sat21h
Acceleration	Sat22h
Handling	Sat23h
Suspension comfort	Sat24h
Silence in rolling	Sat25h
Maniability	Sat26h
Direction	Sat27h
Gears	Sat28h
Mechanic reliability	Sat29h
Oil consumption	Sat30h
Mechanic's efficiency in solving problems	Sat31h
Maintenance cost and repairs	Sat32h

10.3.2 Clustering Variables

We have used $1-r_{jj'}$ as distance. We have applied 6 clustering methods of variables: single linkage, average linkage, complete linkage, Ward's criterion, VARCLUS and VARHCA. Single linkage and average linkage did not provide well separated clusters, so they are eliminated.

For Ward's criterion, the tree shows that a partition in 8 classes is reasonable and for complete linkage in 6 classes. The partition obtained by cutting VARHCA tree into 7 clusters is here exactly the same as the partition given by VARCLUS. The Cronbach's α coefficients show that the obtained blocks are unidimensional.

In the following, we present the blocks for complete linkage, Ward's criterion VARCLUS and VARHCA:

In Table 10.2 we can observed that the blocks "solidity" and "driving quality" are identical for all methods. "General satisfaction" has the same composition for complete linkage, VARCLUS and VARHCA, but partition issued from Ward's criterion is more logical, according to experts. By comparison with the other methods, complete linkage groups in a single block the variables which form the blocks "interior design," "driving comfort," "interior comfort" in Ward's criterion, VARCLUS and VARHCA. For VARCLUS and VARHCA, the variables which are associated to the block "maintenance" in Ward's criterion and complete linkage, are in the same block with "quality–price ratio". Complete linkage is the only method which realizes a distinct block for the variable "quality-price ratio."

The tree for Ward's criterion (partition in 8 classes):

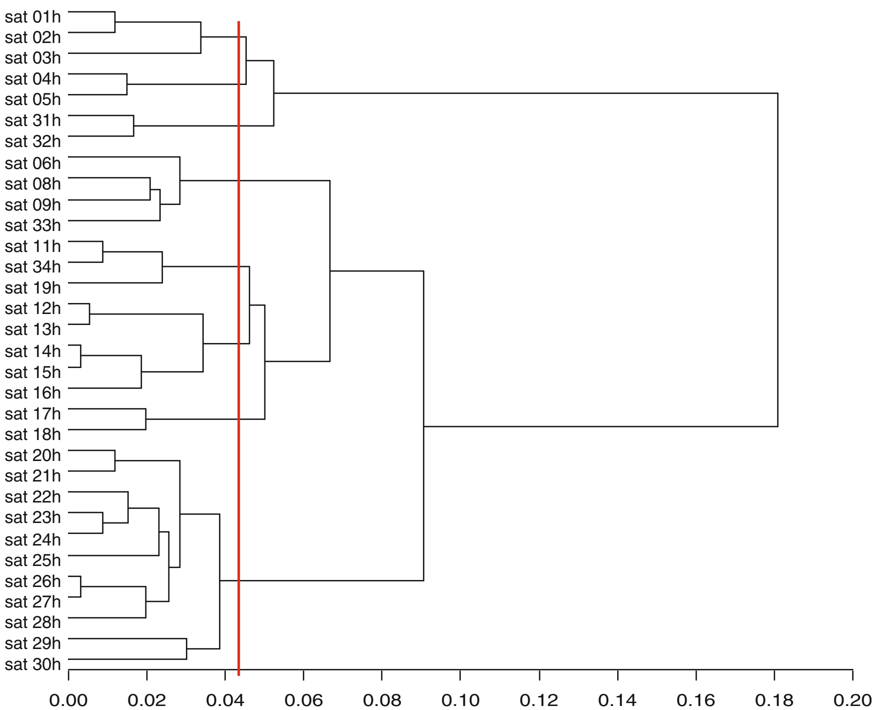
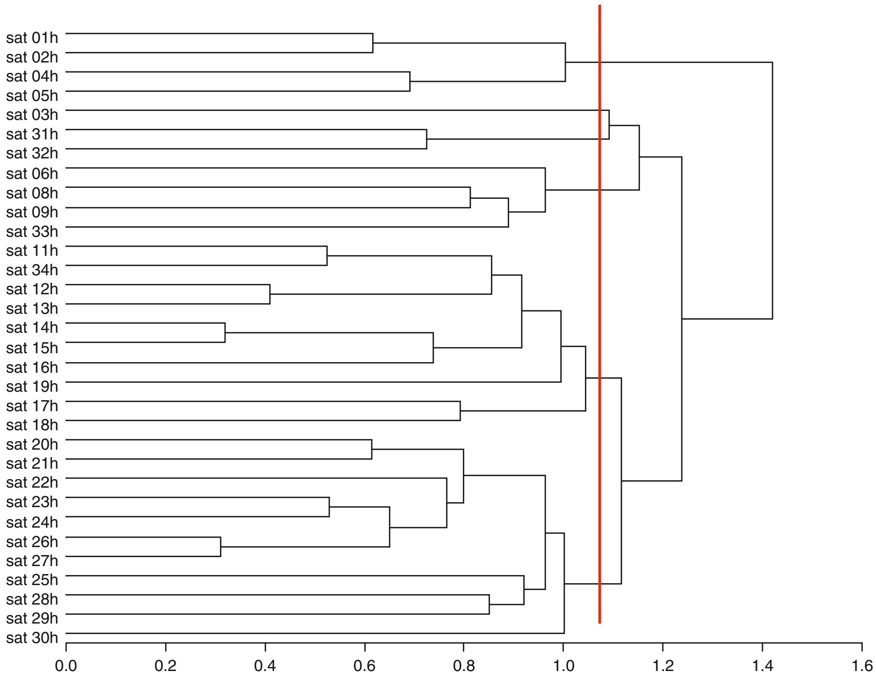


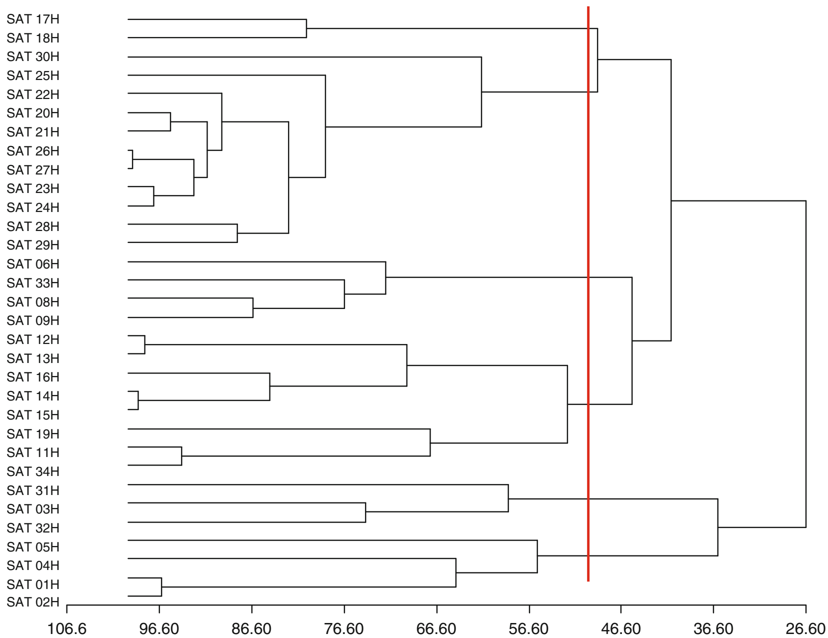
Table 10.2 Blocks of manifest variables after Ward's criterion, complete linkage, VARCLUS and VARHCA

Ward's criterion						Complete linkage						VARCLUS and VARHCA					
Block	MV	Block	MV	Block	MV	Block	MV	Block	MV	Block	MV	Block	MV	Block	MV	Block	MV
General satisfaction (Gs)	Sat01h	Solidity (Sd)	Sat06h	Driving quality (Dq)	Sat20h	General satisfaction (Gs)	Sat01h	Comfort (Cf)	Sat11h	Driving quality (Dq)	Sat20h	General satisfaction (Gs)	Sat01h	Solidity (Sd)	Sat06h	Driving quality (Dq)	Sat20h
	Sat02h		Sat08h		Sat21h		Sat02h		Sat34h		Sat21h		Sat02h		Sat08h		Sat21h
	Sat03h		Sat09h		Sat22h		Sat04h		Sat12h		Sat22h		Sat04h		Sat09h		Sat22h
Construct quality (Cq)	Sat04h		Sat33h		Sat23h		Sat05h		Sat13h		Sat23h		Sat05h		Sat33h		Sat23h
	Sat05h	Driving comfort (Dc)	Sat12h		Sat24h	Quality – price ratio (Qpr)	Sat03h		Sat14h		Sat24h	Quality – price (Qp)	Sat03h	Driving comfort (Dc)	Sat12h		Sat24h
Maintenance (Mn)	Sat31h		Sat13h		Sat25h	Maintenance (Mn)	Sat31h		Sat15h		Sat26h		Sat31h		Sat13h		Sat25h
	Sat32h		Sat14h		Sat26h		Sat32h		Sat16h		Sat27h		Sat32h		Sat14h		Sat26h
Interior design (Id)	Sat11h		Sat15h		Sat27h	Solidity (Sd)	Sat06h		Sat19h		Sat25h	Interior design (Id)	Sat11h		Sat15h		Sat27h
	Sat34h		Sat16h		Sat28h		Sat08h		Sat17h		Sat28h		Sat19h		Sat16h		Sat28h
	Sat19h	Interior comfort (Ic)	Sat17h		Sat29h		Sat09h		Sat18h		Sat29h		Sat34h	Interior comfort (Ic)	Sat17h		Sat29h
			Sat18h		Sat30h		Sat33h				Sat30h				Sat18h		Sat30h

The tree for complete linkage (partition in 6 classes):



The tree for VARHCA et VARCLUS (partition in 7 classes)



10.3.3 PLS Structural Models

The clustering techniques provide blocks but not the relationships between them.

With the help of experts we then propose relations between blocks, so as to explain the latent variable "general satisfaction". The following figures give the 3 causality schemes (Figs. 10.1–10.3):

The values of Amato's criterion *GoF* are:

– For Ward's criterion: $GoF = 0.48$
– For complete linkage: $GoF = 0.42$
– For VARCLUS: $GoF = 0.47$

Ward's clustering gives the best result and will be selected.

10.3.4 Results and Interpretations

10.3.4.1 The Measurement Model

After convergence of the PLS algorithm, one obtains the final weights which allow us to link the manifest variables with the latent variables. An example for "general satisfaction":

$$Gs \;=\; 0,22\,Sat01h + 0,57\,Sat02h + 0,48\,Sat03h.$$

This table presents only correlations larger than the mean of the absolute values (0.3333) (Table 10.3).

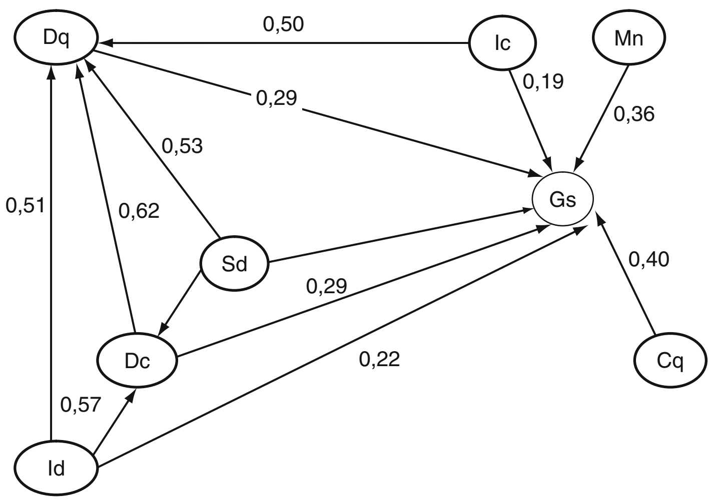


Fig. 10.1 Causality scheme after Ward's clustering

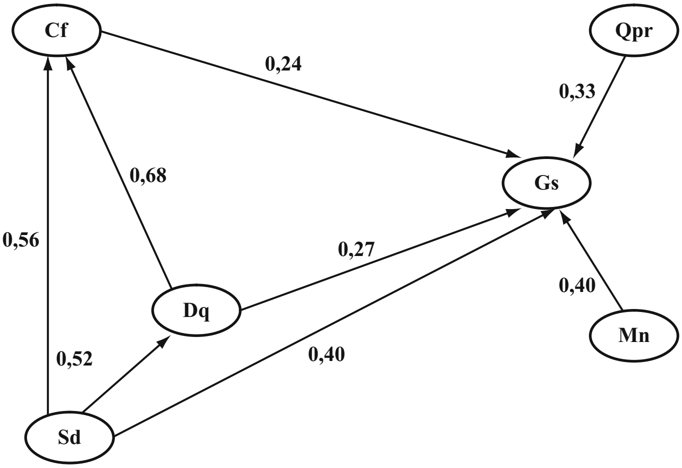


Fig. 10.2 Causality scheme after complete linkage clustering

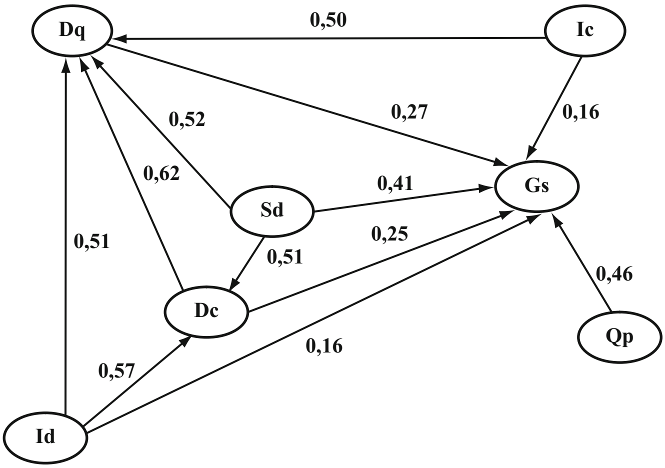


Fig. 10.3 Causality scheme after VARCLUS or VARCHA clustering

Analyzing the correlations, we observe that all latent variables are well correlated with their own manifest. So, the manifest variables "describe" their latent appropriately and the blocks are therefore validated.

10.3.4.2 The Structural Model

The R^2 coefficients between connected latent variables are:

$$R^2\,(Driving\ comfort; Sd,\ Id) = 0.42$$
$$R^2\,(Driving\ quality; Sd,\ Id,\ Dc,\ Ic) = 0.5$$
$$R^2\,(General\ satisfaction; Cq,\ Mn,\ Sd,\ Id,\ Dc,\ Ic,\ Dq) = 0.27$$

Table 10.3 Correlations between manifest and latent variables

Variables	General satisfaction	Construct quality	Maintenance	Solidity	Interior design	Driving comfort	Interior comfort	Driving quality
Sat01h	0.6442	0.3588						
Sat02h	0.8706	0.4011		0.3731				
Sat03h	0.7397							
Sat04h	0.3667	0.8780						
Sat05h		0.8449						
Sat31h		0.3828	0.8739					
Sat32h			0.8332					
Sat06h				0.6534		0.3428		
Sat08h				0.7867	0.3558	0.4223		0.4605
Sat09h				0.7057		0.3493		0.3707
Sat33h				0.7061		0.3420		
Sat11h				0.3597	0.8801	0.5249		0.4442
Sat34h				0.4039	0.8286	0.4816		0.4088
Sat19h					0.7015	0.3651		0.3774
Sat12h				0.4308	0.4684	0.7711	0.3480	0.4782
Sat13h				0.4305	0.4502	0.7903	0.3396	0.4522
Sat14h				0.3756	0.4351	0.8122	0.3461	0.4786
Sat15h				0.3914	0.4611	0.8283	0.3851	0.5367
Sat16h				0.3444	0.3971	0.6595	0.3403	0.4455
Sat17h						0.3508	0.8110	0.3895
Sat18h				0.3434	0.3506	0.4086	0.8665	0.4514
Sat20h				0.4589	0.4924	0.5299	0.4713	0.7315
Sat21h				0.3909	0.3453	0.3952	0.3760	0.6739
Sat22h					0.3349	0.3944	0.3349	0.6757
Sat23h				0.3737	0.3685	0.4458	0.3690	0.7716
Sat24h				0.3685	0.3647	0.4789	0.3379	0.7362
Sat25h						0.3840		0.6218
Sat26h				0.3908	0.3724	0.4791	0.3593	0.7837
Sat27h				0.3902	0.3594	0.4880	0.3751	0.7841
Sat28h				0.3573		0.4048		0.6396
Sat29h								0.5690
Sat30h								0.4743

For "general satisfaction," the R^2 coefficient generated by the other latent variables is 27%, and we consider that as satisfactory because there are 2,922 individuals (Table 10.4).

The correlations between the latent variables are given below:

Analyzing the correlations between the latent variables, we can see that to improve "driving quality", the producer should concentrate on "driving comfort" (correlation coefficient = 0.62), on the "solidity" (0.53) and on the "interior design" (0.51).

In order to obtain a good "driving comfort", the producer could concentrate on "interior design" (0.57) and on "solidity" (0.51).

Given the causality scheme, the determination of "general satisfaction" is a complex procedure in which almost all the latent variables are directly involved.

Table 10.4 The correlations between latent variables

	General satisfaction	Construct quality	Maintenance	Solidity	Interior design	Driving comfort	Interior comfort	Driving quality
General satisfaction	1.0000							
Construct quality	0.4041	1.0000						
Maintenance	0.3576	0.3503	1.0000					
Solidity	0.3722	0.3407	0.2914	1.0000				
Interior design	0.2237	0.0988	0.1979	0.4217	1.0000			
Driving comfort	0.2928	0.1539	0.2266	0.5119	0.5729	1.0000		
Interior comfort	0.1854	0.1233	0.2301	0.3951	0.3812	0.4542	1.0000	
Driving quality	0.2943	0.2023	0.3071	0.5257	0.5085	0.6180	0.5029	1.0000

"Construct quality" is the most important variable for the "general satisfaction" (correlation coefficient $= 0.40$) and the less important is the "interior comfort" (0.19).

Consequently, in order to increase the general satisfaction, the producer should concentrate first on the "construct quality" and then on the "solidity", "maintenance", "driving quality", "driving comfort", "interior design" and "interior comfort".

The equation is as follows:

$$Gs = 0.26\,Cq + 0.19\,Mn + 0.15\,Sd + 0.03\,Id + 0.10\,Dc - 0.03\,Ic + 0.04\,Dq.$$

10.4 Conclusions

Variables clustering provide a simple way of obtaining unidimensional blocks in structural equation modeling, when prior knowledge of blocks is not available.

It must be underlined that this study did not follow the logical sequence of steps of the PLS approach: the construction of a model by experts, the construction of a questionnaire using this model, and the collection of customer data using this questionnaire.

In our case, the process is inverted: we have tried to build a model using data that had already been collected. This fact has obviously effects on the final results which cannot be measured.

By means of clustering methods of variables, we established the external model. According to Amato's criterion, Ward's clustering was choosen as the best technique for our data set. But we observe that the values of this criterion for the 3 models are very close.

For the chosen model, a hierarchy of the influence of the latent variables on general satisfaction can be established using the structural model:

I. Construct quality; II. Solidity; III. Maintenance; IV. Driving quality, V. Driving comfort, VI. Interior design, VII. Interior comfort.

The results obtained are satisfactory: $R^2 = 27\%$ for a large sample of almost 3,000 respondents.

References

Amato, S., Esposito Vinzi, V., & Tenenhaus, M. (2004). A global goodness-of-fit index for PLS structural equation modeling. *Oral Communication to PLS Club, HEC School of Management, France, March 24*.

Nakache, J. P., & Confais J. (2005) Approche pragmatique de la classification. *Ed. Technip, Paris*.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis, 48*(1), 159–205.

Vigneau, E., & Qannari, E. M. (2003). Clustering of variables around latent component – application to sensory analysis. *Communications in Statistics, Simulation and Computation, 32*(4), 1131–1150.