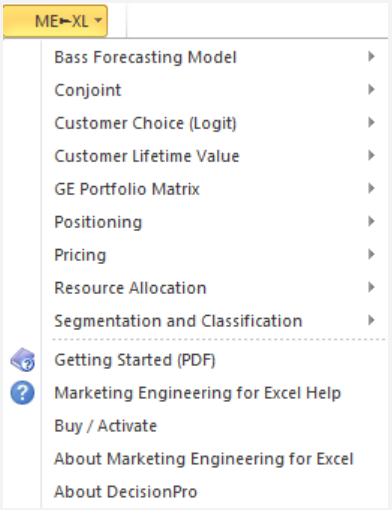
Tutorial

Segmentation and Classification

Marketing Engineering for Excel is a *Microsoft Excel* add-in. The software runs from within Microsoft Excel and only with data contained in an Excel spreadsheet.

After installing the software, simply open *Microsoft Excel*. A new menu appears, called "*ME ▸ XL.*" This tutorial refers to the "*ME ▸ XL/Segmentation and Classification*" submenu.



Overview

Segmentation and classification is an analytic technique that helps firms compare and group customers who share common characteristics (i.e., segmentation variables) into homogeneous segments and identify ways to target particular segments of customers in a market on the basis of external variables (i.e., discriminant variables).

Segmentation refers to the process of classifying customers into homogenous groups (segments), such that each group of customers shares enough characteristics in common to make it viable for the firm to design specific offerings or products for it. This application identifies customer segments using needs-based variables called basis variables. Cluster analysis helps firms to:

- ✓ better understand their customers.
- ✓ identify different segments in a market.
- ✓ choose attractive customer segments for classification with its marketing programs.

Getting Started

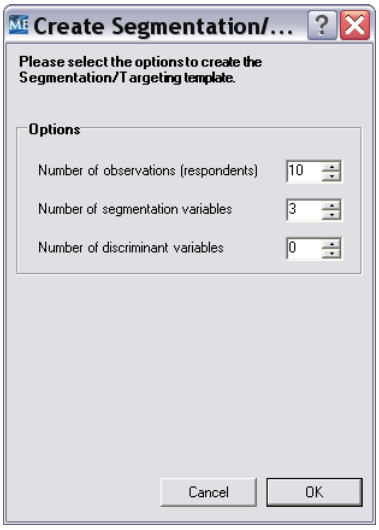
To apply segmentation and classification analysis, you can use your own data directly or use a template preformatted by the ME►XL software.

The next section explains how to create an easy-to-use template to enter your own data.

If you want to run a segmentation and classification analysis immediately, open the example file "*OfficeStar Data (Segmentation).xls*" and jump to "Step 3: Running analysis" (p. 6). By default, the example files install in "My *Documents/My Marketing Engineering/.*"

Step 1 Creating a template

In Excel, if you click on ME►XL → SEGMENTATION AND CLASSIFICATION → CREATE TEMPLATE, a dialog box appears. This box represents the first step in creating a template to run the segmentation and classification analysis software.



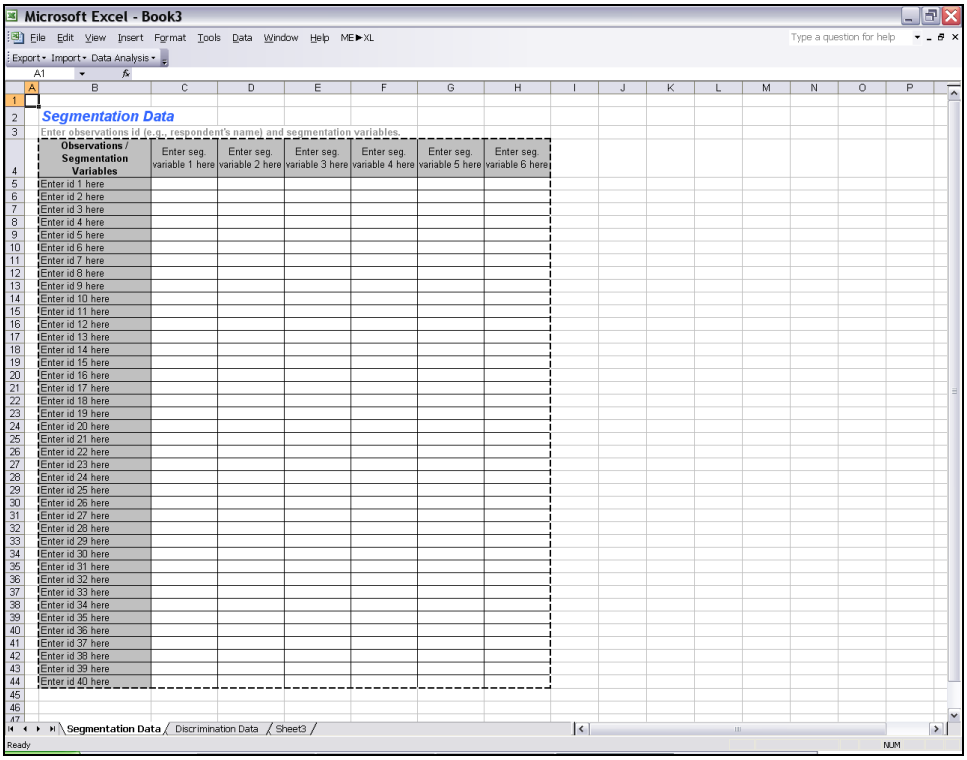
The dialog box requests three pieces of information to design the template:

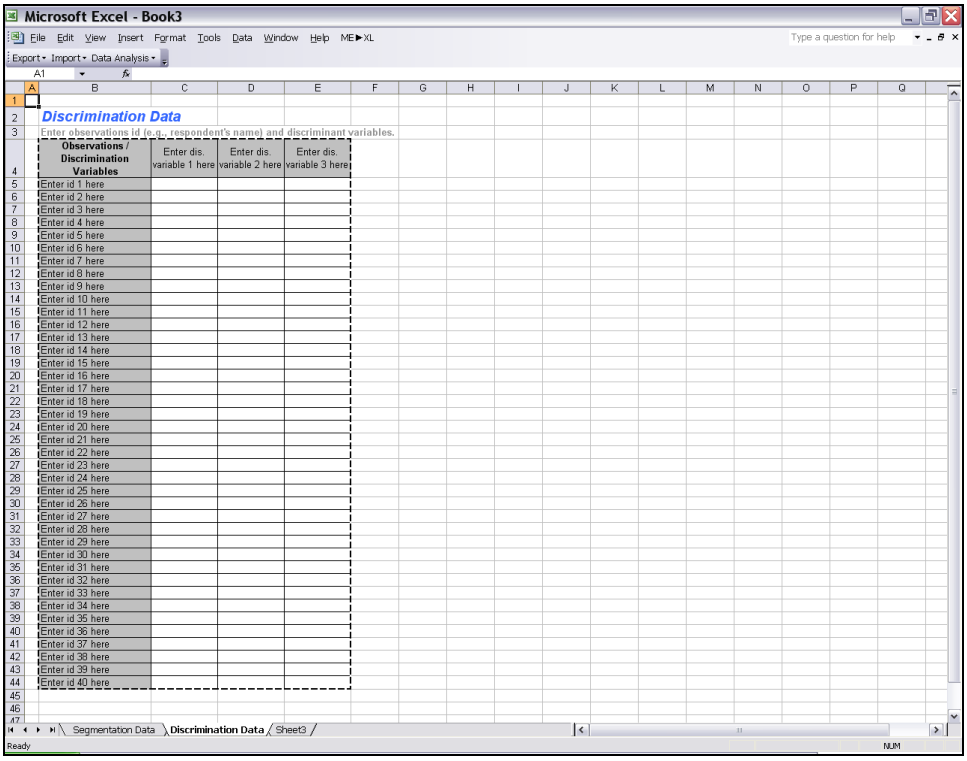
- **Observations** (respondents) indicate the number of customers or respondents in the data that need to be clustered.
- **Segmentation variables** help us assess the similarity between two respondents. These variables serve as the basis for segmentation and are often called **basis variables**. They might include customer's needs, wants, expectations, or preferences.
- **Discriminant variables**, also called **descriptors**, are optional variables that can describe the segments formed on the basis of the segmentation variables. These include demographic variables, such as educational level, gender, income, media consumption, and the like.

It is not always clear whether a specific variable should be treated as a segmentation variable or discriminant variable. This choice might depend on the context, the managerial question, or the product category.

When in doubt, ask yourself the following questions: (1) Would this piece of information tell me what that customer wants, in which case it should be treated as segmentation variable, or (2) does this piece of information tell me who that customer is and therefore should be treated as discriminant variable? For example, "gender" would fall in the second category most of the time, whereas "need for timely information" usually falls in the former category.

After specifying the number of observations and variables, click OK to proceed. The software generates a template that contains either one or two sheets, depending on whether you have included discriminant data.

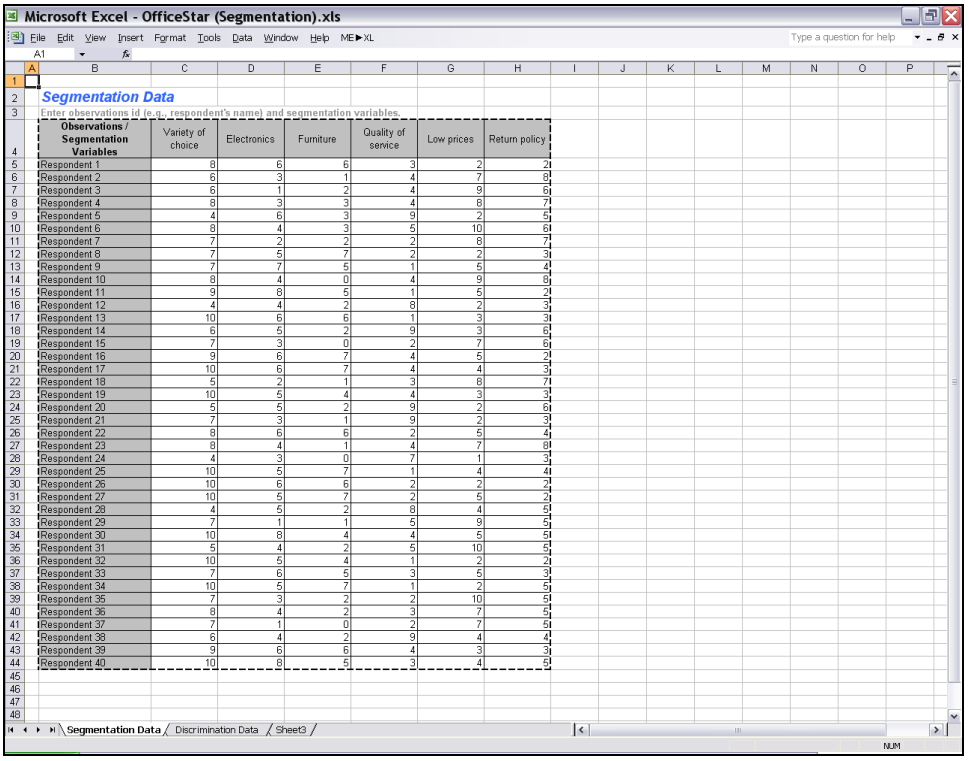


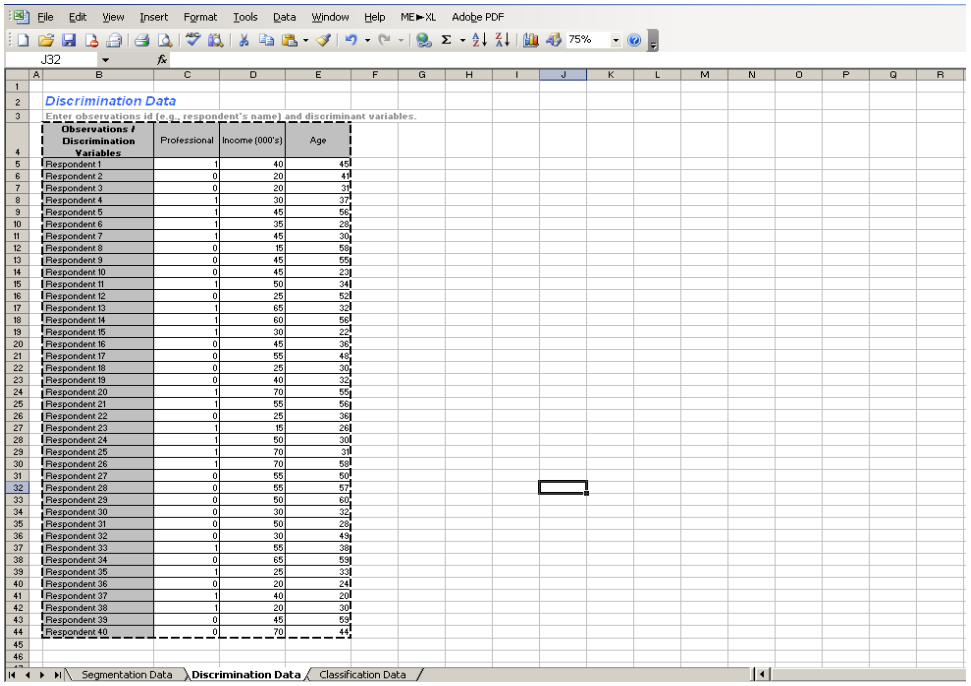


Step 2 Entering your data

In this tutorial, we use the example file "*OfficeStar Data (Segmentation).xls,*" which appears by default in "*My Documents/My Marketing Engineering/.*"

To view a proper data format, open that spreadsheet in Excel. A snapshot is shown below.





A typical segmentation spreadsheet contains one or two spreadsheets that contain segmentation and/or discrimination data.

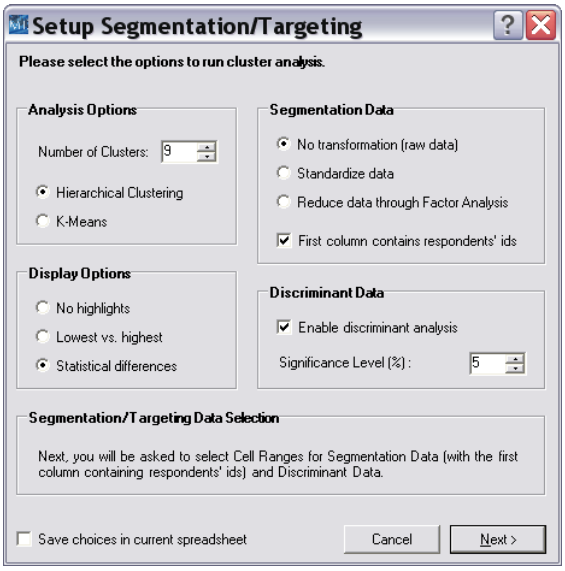
- **Segmentation data** are required for the segmentation model. This data set contains the respondent identifier and a column for each segmentation variable collected in the study. The data within each column must be

scaled using the same scale (e.g., 1–10), but each column can have a different scale (e.g., 1–10 for satisfaction, 1–5 for convenience). Typically, segmentation variables are numerical values (interval or ratio scale). The data set contains one row per respondent in your study. If you must use basis variables that are nominal (e.g., "male" "female"), then you can apply latent class segmentation analysis (see appendix).

- **Discriminant data** constitute an optional data set, depending on whether your study has collected discrimination data. Recall that discrimination data enables you to differentiate one customer from another (e.g., age, income, gender). Again, data within a column must be scaled using the same scale, but different columns may use different scales. Typically, discriminant variables are numerical (interval or ratio scale) or nominal ("male", "female"). Each respondent in your study appears in a separate row.

Step 3 Running segmentation analyses

After you enter your data in an Excel spreadsheet with the appropriate format, click on ME ▸ XL → SEGMENTATION AND CLASSIFICATION → RUN SEGMENTATION. The dialog box that appears indicates the next steps required to perform a segmentation analysis of your data.



Analysis options

You may specify the number of segments (clusters) to develop during the analysis. For the segmentation method, you can choose either K-means or hierarchical clustering.

- **Hierarchical clustering** builds up or breaks down the data, customer by customer (row by row).
- **K-means** partitioning breaks the data into a pre-specified number of segments and then reallocates or swaps customers to improve some measure of effectiveness.

Usually, a segmentation analysis consists of two steps. First, you run the analysis with a large number of segments (up to 9). Second, on the basis of a dendogram analysis (discussed subsequently), you can determine the number of segments to retain for further analysis.

Segmentation data

This section enables you to specify how to pre-process the data, and whether the first column of the data has respondent identifiers.

- **No transformation.** This button indicates you want to use the original data.
- **Standardize data.** This option scales all variables to 0 mean and unit variance before the analysis. Choosing this option is a good idea if you have measured the variables on different scales.
- **Reduce data through Factor Analysis.** This button combines related variables into unique factors.

Display options

In this section, you specify how you want the cluster data presented.

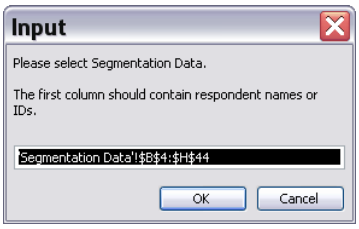
- **No highlights.** The results are displayed in plain tabular format.
- **Lowest vs. highest.** For each variable, colors highlight the value of the cluster with the highest (green) and lowest (red) values.
- **Statistical differences.** For each variable, colors highlight clusters whose values are statistically different from the overall mean at a 95% confidence level. Those that are different from the mean at a 99% confidence level appear in italics.

Discriminant data

Decide whether you want to include a discriminant analysis. Check this button if you wish to perform discriminant analysis, and indicate the level of statistical significance you wish to use.

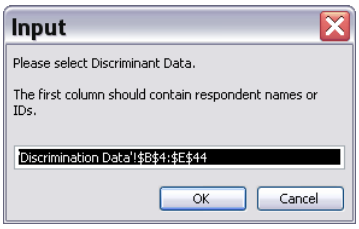
The *Save choices in current worksheet* option allows you to save cell range selections when you perform Run Analysis. If you are using your own data or have modified a Marketing Engineering for Excel template, you should choose this check box to save your selections.

After selecting all the options, you must select the cells containing the data. When you click Next, the following dialog box appears:



The software requests a range for the segmentation data. If you are using a *Marketing Engineering for Excel* template, the software preselects the cell ranges.

If you have specified the inclusion of discriminant data, the following dialog box appears, which allows you to select your discrimination data. The cell ranges might be pre-selected.



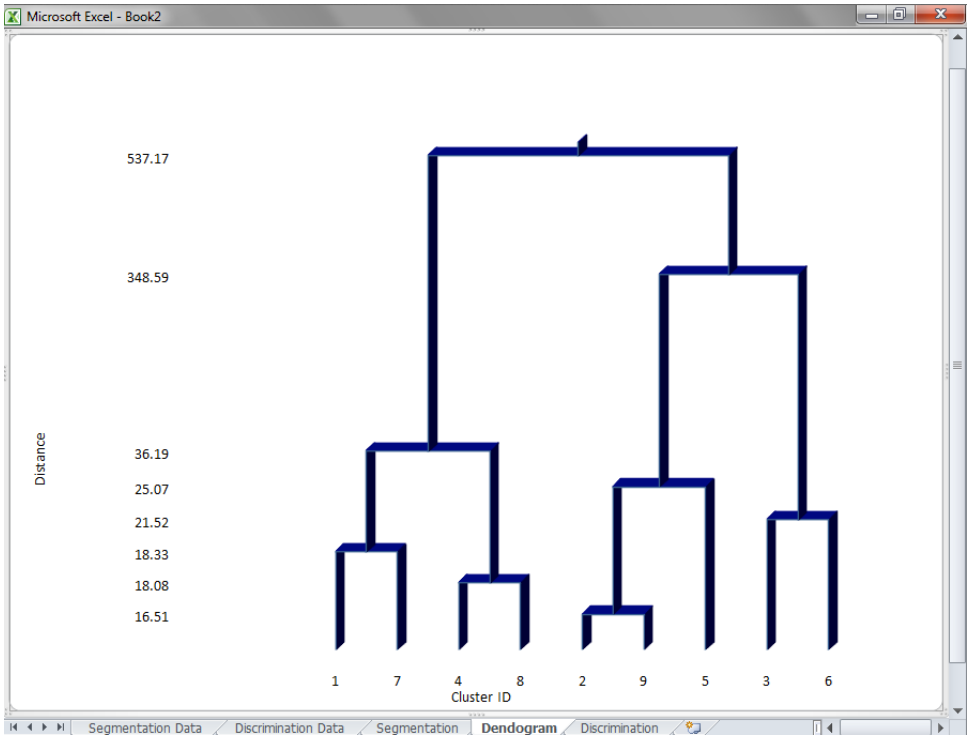
The newly generated workbook contains the results of your segmentation analysis.

Step 4 Interpreting the segmentation results

The workbook generated by segmentation analysis may contain several worksheets, depending on whether your study has included discriminant data.

Dendogram

Dendograms provide graphical representations of the loss of information generated by grouping different clusters (or customers) together. The dendogram is generated only if you choose the Hierarchical Clustering option in the Setup dialog box.



At one extreme (upper part of the dendogram), all customers group into one cluster, and the loss of information is maximum, because they all receive undifferentiated treatment, regardless of their characteristics.

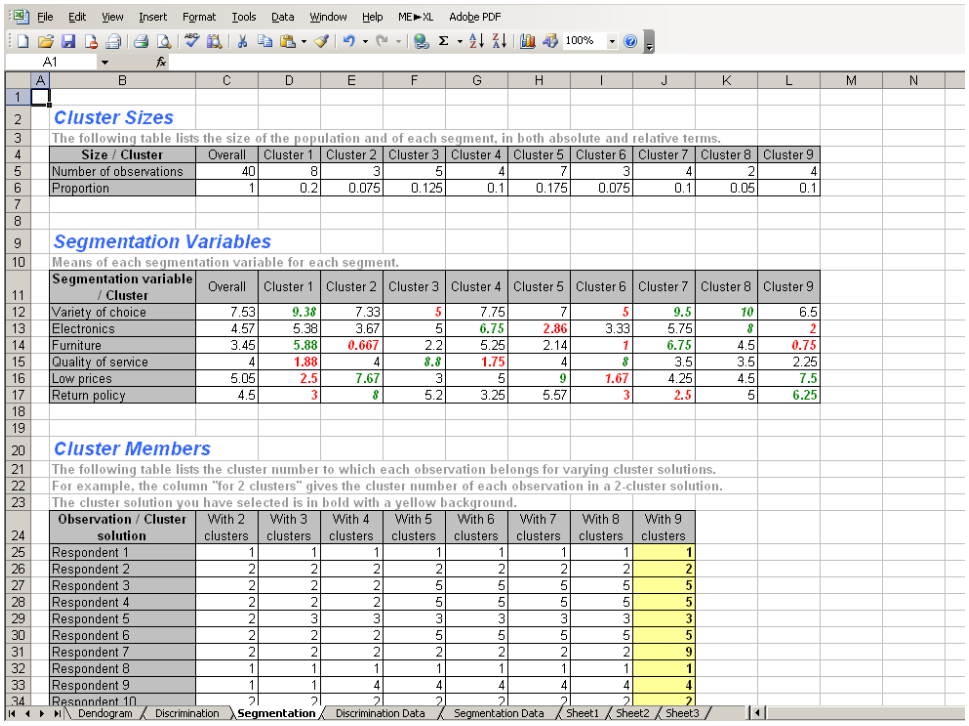
At the other extreme (lower part of the dendogram), customers appear in separate, small clusters, and only those customers very similar to one another group together ("similar" or "close" in this context refers to the distance between two customers in terms of the segmentation variables).

When reviewing a dendogram, look for significant distances or "jumps" in the distances. For example, the *OfficeStar* example contains a very large jump when moving from three to two clusters. Grouping these three clusters into two generates a significant loss of information; in other words, it results in grouping within the same cluster customers who are very dissimilar. In the preceding example, a three-cluster solution seems to be the best approach.

A dendogram is simply a graphical representation of the clustering output. For a more detailed understanding of cluster members and attributes, you must analyze the other tabs in the segmentation output as well.

Segmentation

The tab contains the statistical output of the cluster process and shows cluster sizes (number of members), cluster means, and the placement of each member in clusters (highlighted in yellow). This tab also provides columns that represent individual members and where they would be clustered in a 2–9 cluster solution.



Cluster Sizes

The following table lists the size of the population and of each segment, in both absolute and relative terms.

Size / Cluster	Overall	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Number of observations	40	8	3	5	4	7	3	4	2	4
Proportion	1	0.2	0.075	0.125	0.1	0.175	0.075	0.1	0.05	0.1

Segmentation Variables

Means of each segmentation variable for each segment.

Segmentation variable / Cluster	Overall	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Variety of choice	7.53	9.38	7.33	5	7.75	7	5	9.5	10	6.5
Electronics	4.57	5.38	3.67	5	6.75	2.86	3.33	5.75	8	2
Furniture	3.45	5.88	0.667	2.2	5.25	2.14	1	6.75	4.5	0.75
Quality of service	4	1.88	4	8.8	1.75	4	8	3.5	3.5	2.25
Low prices	5.05	2.5	7.67	3	5	9	1.67	4.25	4.5	7.5
Return policy	4.5	3	8	5.2	3.25	5.57	3	2.5	5	6.25

Cluster Members

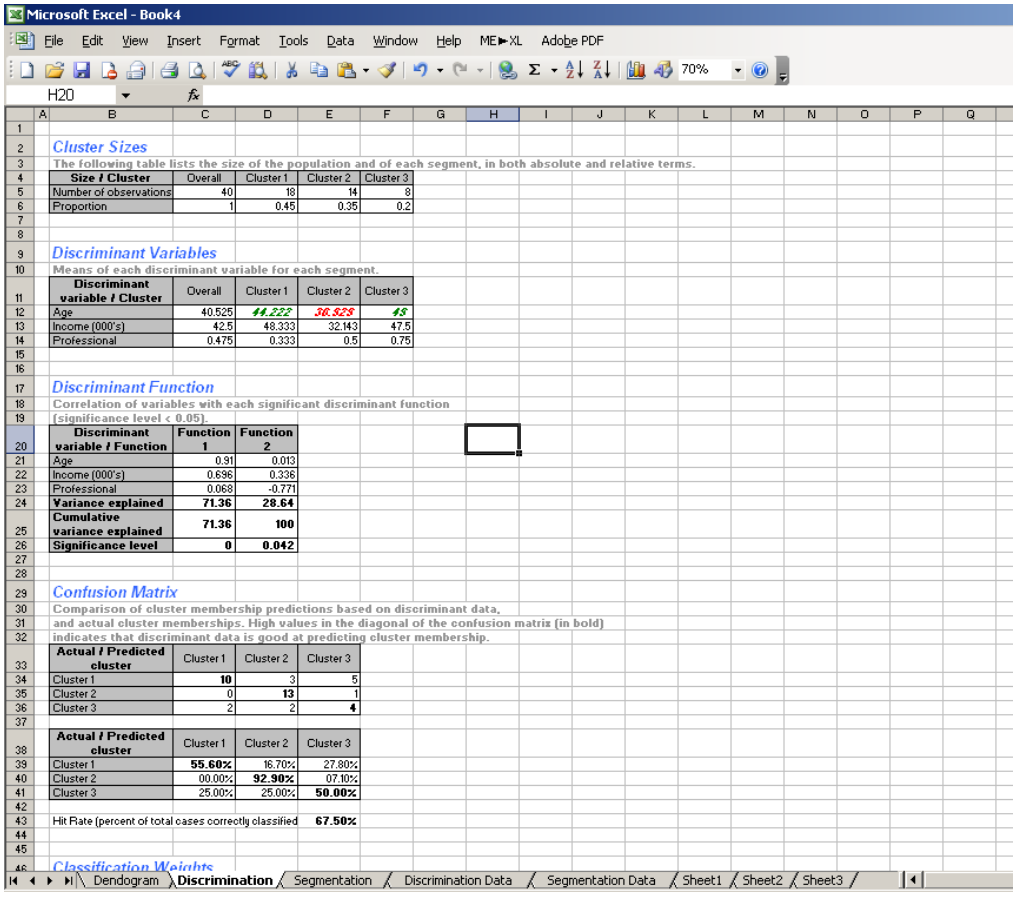
The following table lists the cluster number to which each observation belongs for varying cluster solutions.
For example, the column "for 2 clusters" gives the cluster number of each observation in a 2-cluster solution.
The cluster solution you have selected is in bold with a yellow background.

Observation / Cluster solution	With 2 clusters	With 3 clusters	With 4 clusters	With 5 clusters	With 6 clusters	With 7 clusters	With 8 clusters	With 9 clusters
Respondent 1	1	1	1	1	1	1	1	**1**
Respondent 2	2	2	2	2	2	2	2	**2**
Respondent 3	2	2	2	5	5	5	5	**5**
Respondent 4	2	2	2	5	5	5	5	**5**
Respondent 5	2	3	3	3	3	3	3	**3**
Respondent 6	2	2	2	5	5	5	5	**5**
Respondent 7	2	2	2	2	2	2	2	**9**
Respondent 8	1	1	1	1	1	1	1	**1**
Respondent 9	1	1	4	4	4	4	4	**4**
Respondent 10	2	2	2	2	2	2	2	**2**

Discrimination

This optional spreadsheet reflects the output of the discrimination analysis. The matrices included on this sheet are as follows:

- **Cluster sizes** depicts the number of respondents who appear in each cluster, along with the proportion of the whole population that each cluster represents.
- **Discriminant variables** depict the means of each discriminant variable for each cluster.
- **Discriminant function** reflects the correlation of the variables with each significant discriminant function and thus indicates the predictive ability of each discriminant function.
- **Confusion matrix** depicts how well the discriminant data predict correct clusters. Two matrices are available, one showing the actual data counts and the other showing percentages for these same data.
- **Classification weights** and **classification coefficients** are intermediary results required to run further classification analyses on external data. These matrices are of no particular interest as is, and cannot be easily interpreted, but are necessary to carry over further classification analyses.



Microsoft Excel - Book4

File Edit View Insert Format Tools Data Window Help ME►XL Adobe PDF

H20

Cluster Sizes

The following table lists the size of the population and of each segment, in both absolute and relative terms.

Size / Cluster	Overall	Cluster 1	Cluster 2	Cluster 3
Number of observations	40	18	14	8
Proportion	1	0.45	0.35	0.2

Discriminant Variables

Means of each discriminant variable for each segment.

Discriminant variable / Cluster	Overall	Cluster 1	Cluster 2	Cluster 3
Age	40.525	44.222	36.929	48
Income (000's)	42.5	48.333	32.143	47.5
Professional	0.475	0.333	0.5	0.75

Discriminant Function

Correlation of variables with each significant discriminant function (significance level < 0.05).

Discriminant variable / Function	Function 1	Function 2
Age	0.91	0.013
Income (000's)	0.696	0.336
Professional	0.068	-0.771
Variance explained	**71.36**	**28.64**
Cumulative variance explained	**71.36**	**100**
Significance level	**0**	**0.042**

Confusion Matrix

Comparison of cluster membership predictions based on discriminant data, and actual cluster memberships. High values in the diagonal of the confusion matrix (in bold) indicates that discriminant data is good at predicting cluster membership.

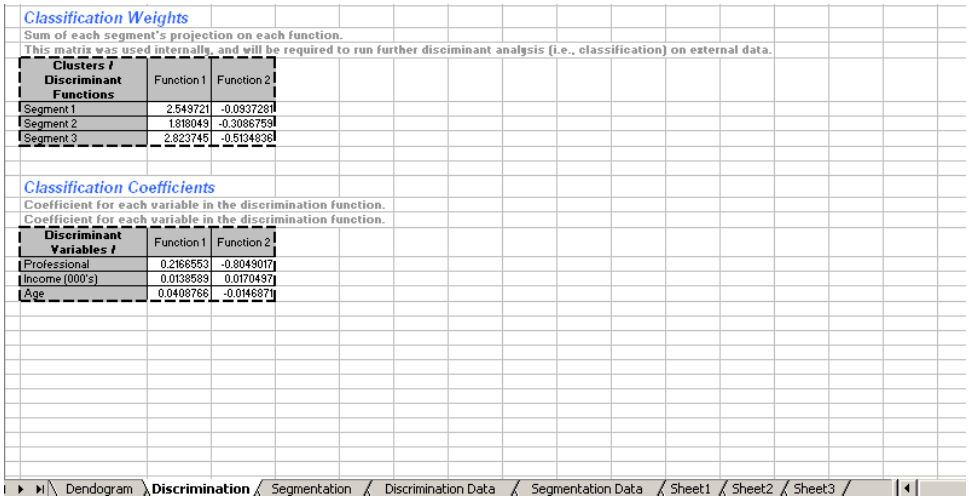
Actual / Predicted cluster	Cluster 1	Cluster 2	Cluster 3
Cluster 1	**10**	3	5
Cluster 2	0	**13**	1
Cluster 3	2	2	**4**

Actual / Predicted cluster	Cluster 1	Cluster 2	Cluster 3
Cluster 1	**55.60%**	16.70%	27.80%
Cluster 2	00.00%	**92.90%**	07.10%
Cluster 3	25.00%	25.00%	**50.00%**

Hit Rate (percent of total cases correctly classified) **67.50%**

Classification Weights

Dendogram | Discrimination | Segmentation | Discrimination Data | Segmentation Data | Sheet1 | Sheet2 | Sheet3



Classification Weights

Sum of each segment's projection on each function.

This matrix was used internally, and will be required to run further discriminant analysis (i.e., classification) on external data.

Clusters / Discriminant Functions	Function 1	Function 2
Segment 1	2.549721	-0.0937281
Segment 2	1.818049	-0.3086759
Segment 3	2.823745	-0.5134836

Classification Coefficients

Coefficient for each variable in the discrimination function.

Coefficient for each variable in the discrimination function.

Discriminant Variables /	Function 1	Function 2
Professional	0.2166553	-0.8049017
Income (000's)	0.0138589	0.0170497
Age	0.0408766	-0.0146871

Segmentation and discriminant data

These tabs contain the original segmentation and discriminant data used for the segmentation analysis, included in the output for your convenience. The original spreadsheet used for the analysis remains intact, so you can modify it for subsequent analysis runs. The data preserved with this tab always reflect the data represented in the dendogram and segmentation tabs.

Step 5 Running classification analyses

Introduction

If you ran segmentation analysis with discriminant data, the software estimated the best way to predict to which cluster an individual is most likely to belong based solely on discriminant data. This is very useful to predict whether young people (age as a discriminant factor) are more likely to be more price sensitive (price sensitivity as a segmentation variable); or if businesses in certain industries require more support than others.

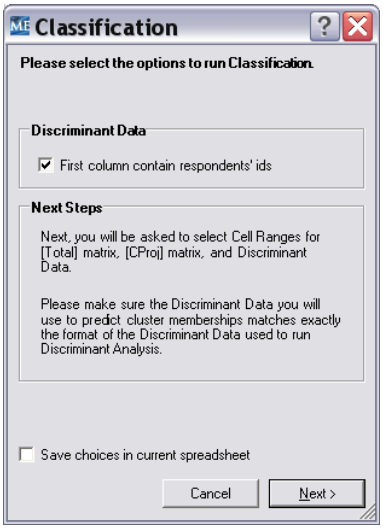
The ability of recouping segment membership based on discriminant variables is best summarized by the confusion matrix and hit rate (see above).

Once this discriminant analysis has been applied to the original dataset, it can be applied again to external customers for whom discriminant data—but no segmentation data—is available. The process of classifying customers among segments, based on a preceding segmentation analysis, but using discriminant data only, is called **classification analysis**.

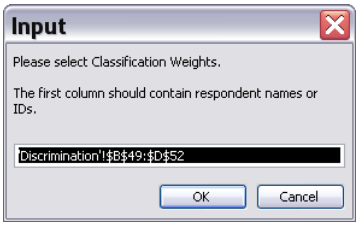
Classification analysis is usually applied to new customers, for whom segmentation data is not available. For learning purpose, you can also apply it to discriminant data of customers for whom segmentation data is available, and see how well segment memberships are recouped. This analysis is automatically done when you run a segmentation analysis, and its results are summarized by the confusion matrix.

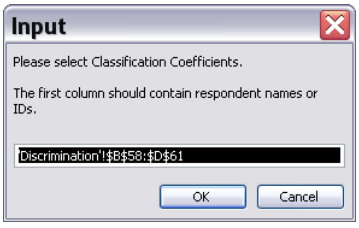
Selecting data

Click on ME ▸ XL → SEGMENTATION AND CLASSIFICATION → RUN CLASSIFICATION. The dialog box that appears indicates the next steps required to perform a classification analysis of your data.

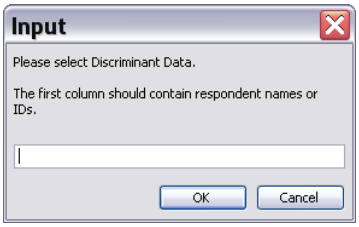


The first steps of the classification analysis consist of selecting two cell ranges: classification weights and classification coefficients. You can find that data at the bottom of the "discriminant" sheet in the analysis workbook generated by segmentation analysis. The cell ranges might be pre-selected.

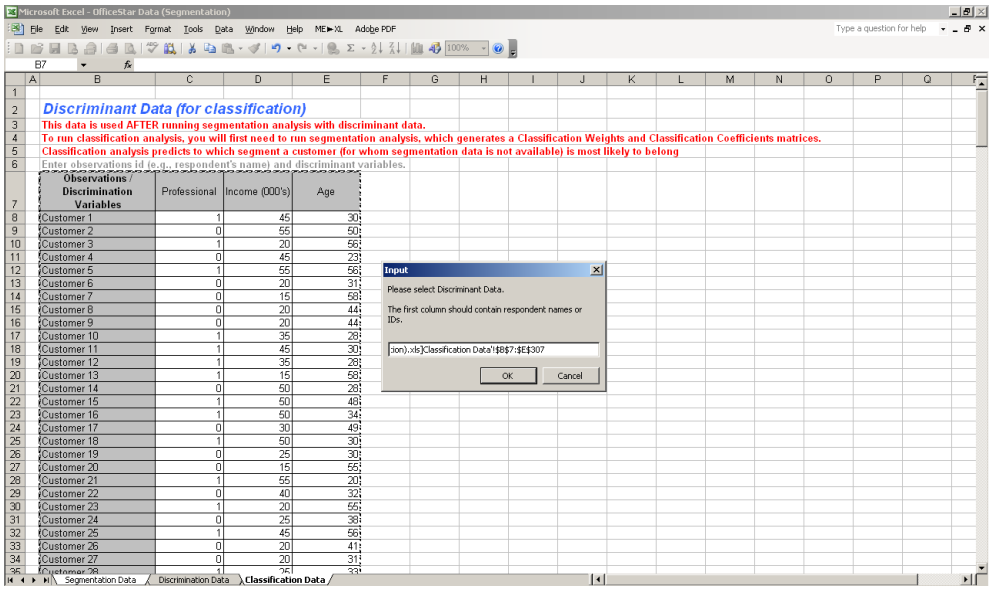




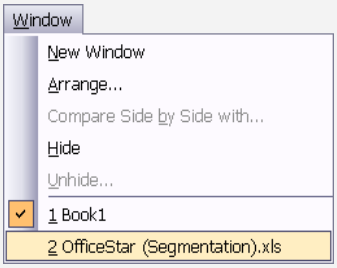
The last step is to select discriminant data. In most cases, that consists of data about new customers for whom no segmentation data is available. It is important that formatting of the discriminant data matches exactly the format of the discriminant data (both variables, orders and ranges) used for the original segmentation analysis.



Discriminant data of "new" customers is available on the original OfficeStar workbook, in the last sheet. Go back to the OfficeStar workbook, and manually select the discriminant data available for the 300 additional customers (the last sheet of the workbook, named classification data).

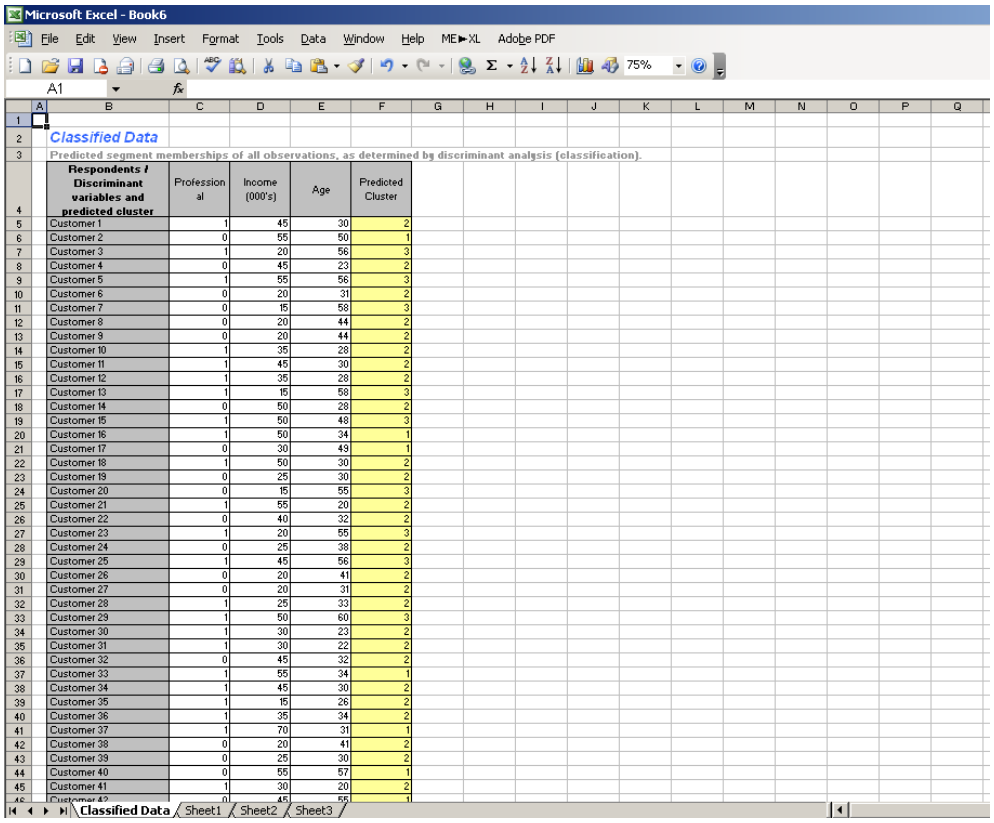


Once you are in selecting mode, Excel might not allow you to easily switch between two workbooks. If you require selecting data in different workbooks (as it is usually the case with classification analysis), simply use the Window menu of Excel to select and open another workbook.



Interpreting the results

When you click Ok, a new workbook is generated. This workbook contains the discriminant data used to run classification analysis, and the segment to which each customer is most likely to belong.



Note that this classification of customers across segments is our best guess based on discriminant analysis. It is not perfect, and some customers might be misclassified, that is, they are the closest to segment *A* in terms of needs, but their discriminant variables send us astray and predict they are more likely to belong to segment *B*.

Appendix Latent Class Segmentation

Latent Class Segmentation is intended for more advanced users, who have large data sets for segmentation. It utilizes modeling and statistical techniques that are more advanced than traditional segmentation. Unlike traditional segmentation, it embeds within its estimation structure, the determination of the number of segments (classes, clusters) that best describe the input data. It also allows for incorporation of categorical data in the determination of segment structure, which is not feasible with traditional segmentation. The illustrations below utilize the OfficeStar data, but such data is not representative of the types of data that one would typically subject to Latent Class Analysis. Latent Class segmentation analysis is an iterative process in which you should explore many different segmentation solutions, requiring you to run the model several times, in order to identify a solution that makes the most sense for addressing your segmentation objectives.

Latent class segmentation (LC) is a powerful and general tool that can be used to identify important large segments and/or distinct niche market segments. The main purpose of using latent class segmentation techniques is to reveal segment structures among customers that are unknown a priori. Thus, LC segmentation is particularly useful for determining unexpected groupings of customers that may point the way to new ways of identifying and targeting customers.

LC models do not rely on the traditional modeling assumptions, such as linearity, normal distribution, and homogeneity. Hence, they are less subject to biases associated with data not conforming to model assumptions. In addition, LC models can include variables of mixed scale types (nominal, ordinal, continuous and/or count variables) in the same analysis. The current implementation in ME>XL allows for nominal and continuous variables.

The latent class segmentation model implemented in ME►XL is the Autoclass model developed by the Ames Research Center at NASA. Autoclass uses a finite mixture Bayesian Classifier that takes a database of customers, described by a combination of real and discrete-valued attributes, and automatically finds the segments, or clusters, within that data. Discrete-valued attributes (e.g., sex, region) are modeled as Bernoulli with uniform Dirichlet conjugate priors and real-valued attributes are modeled as Normal densities with either a Uniform or Normal prior on the means, and Jeffreys prior (for a single attribute) or the inverse Wishart (for attribute subsets) for the variance. In the current implementation in ME►XL, we do not allow for attribute subsets that share a covariance structure.

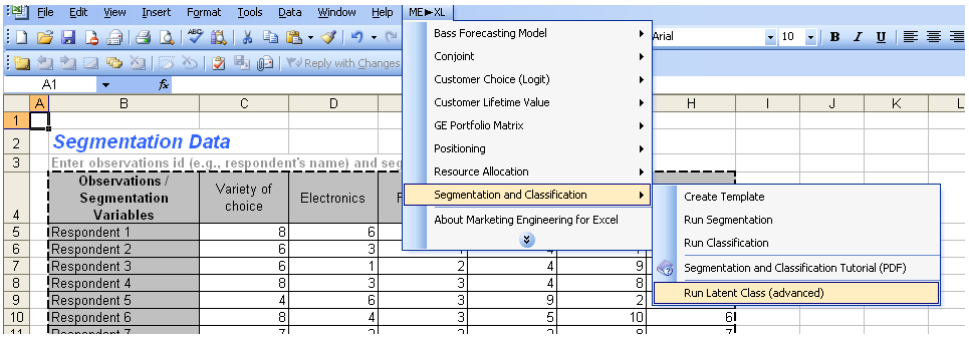
Unlike the traditional segmentation model in ME►XL, Autoclass does not require you to specify the number of segments into which the customers are partitioned—it extracts this information from the data itself. In the results, the segments are described probabilistically, so that every customer can have partial membership in the different segments and the segment definitions can overlap. ME►XL also uses a "hard assignment" to assign each customer to the segment to which that customer has the highest probability of belonging.

Compared to the traditional segmentation methods (Hierarchical and K-Means), latent class segmentation is most useful when the needs variables contain both continuous and discrete attributes (traditional methods do not allow for inclusion of discrete attributes), and we have a sufficiently large sample of respondents, especially, when the number of variables included in the model is large. To get reasonable segment structures, we recommend about 10 respondents per variable per segment. That is, if we expect 3 or 4 segments and have about 10 segmentation variables, we need about 300 to 400 respondents for analysis. However, because our method is Bayesian, it

will produce results (sometimes good results) even if this requirement is violated, **as long as there is sufficient number of observations to make the estimation feasible**. In that case, the user is cautioned to proceed carefully in assessing whether the result segment structure makes business sense.

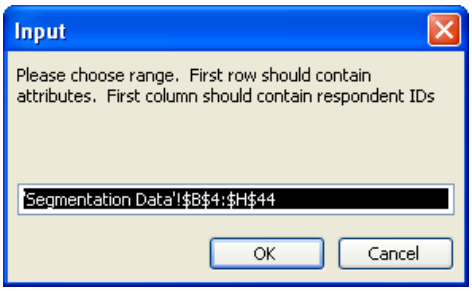
Step 1(lc) Initiate Latent Class Analysis and Load Data

As shown below, Latent Class Analysis is invoked from the ME►XL menu, within the Segmentation and Classification model. This example assumes you are using the OfficeStar (Segmentation) data example found in My Marketing Engineering, or that you have prepared your data using the ME►XL template feature.



Step 2 (lc) Select Data Range

You will be asked to select the range containing the segmentation data.



Step 3 (lc) Choose Parameters to Control the Analysis

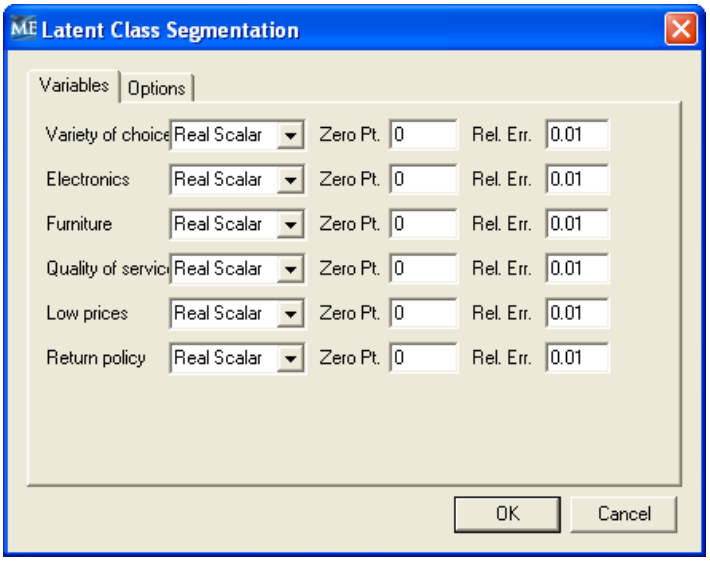
The dialog box shown below will list all of the variables you specified in your segmentation data. For each variable, you must choose the appropriate data type using the drop-down arrow in the right hand side of the first box.

Real Scalar: scalar variables that are bounded below by the ZERO_POINT. An example is height, or measured importance weights a customer associates with a variable. Typically, the ZERO_POINT is set to 0 (the default), but can take on other finite values.

Real Continuous: variables that are unbounded (i.e., those that could theoretically take values from – to + infinity). An example of such a variable

is profit. For purposes of model implementation, we specify a range for this variable, and also measurement error.

Discrete: An example discrete or nominal variable is Color, which can take the values Blue, Green, Red, and Other/Unknown. Dummy variables are nominal variables with two levels, 0 and 1. Discrete variables that have a large number of levels would need more data for estimation (each level of a discrete variable is like a variably by itself, i.e., each level introduces one unknown parameter for model estimation in each segment).



Additionally, for the continuous variable type, you may set two other (optional) parameters:

Zero Pt: This is the smallest value that the measurement process for a variable could have produced (before allowing for error).

Rel Err: This is an indicator of the measurement errors that can occur in the variables. If there is no information about measurement error, this variable can be taken as **half the minimum possible difference between measured values of a variable**. In the case of continuous real variables, we specify an absolute measurement error, and for continuous scalar variables, we specify a relative error. For example, if differences of 0.1 are not distinguishable by the measurement procedure, then **error** can be set to 0.05, and if that variable is real scalar with an average value of 10, then relative error for that variable is 0.005 (0.05/10). As a default, we have set the error to 0.1*range of the continuous variable, and we set relative error to 0.01 for real scalar variables. Large values of error or relative errors will lead to more number segments and more overlaps across segments, and small values will lead to fewer and tighter segments.

For discrete variables, you will be asked to specify a **Level** for each variable (i.e., the number of levels of the variable). ME►XL automatically calculates the number of levels. However, if you suspect that there are, in fact, greater or fewer number of levels than the one determined by ME►XL, you should check to make sure that your data sets do not have any inaccuracies.

The dialog box also contains a second tab as shown below. The **Options** tab allows you to further control the latent class analysis.

Choose the maximum number of segments: This is an optional setting. If you specify a maximum number of segments, the program will not evaluate any segmentation solution that contains more than this number of segments. As a first cut at understanding your data, we recommend you run the model without specifying a maximum number of segments, and explore whether there are unexpected and distinct segments that might be of interest in pursuing your segmentation objectives.

Set starting number of segments: This is an optional setting that enables you to start the search algorithm with a specified number of segments. Ideally, this number should be expected number of segments, and therefore, this option is most useful at later stages of your segmentation analysis, when you are reasonably confident of the number of segments you want in your final segmentation scheme. If you set the maximum number of segments and the starting number of segments to an identical value, then you can explore multiple solutions that contain an identical number of desired segments.

Number of non-duplicate solutions to save: This is the number of distinct segmentation solutions that you wish to explore, set to a default of 10. Depending on the characteristics of the likelihood function (the function that is optimized to identify segments), there may be a number of distinct solutions, each one of which is a local optimum. If you specify a large number for this option, the program will generate very different segmentation schemes for you to explore – this may be useful for identifying niche segments in the market.

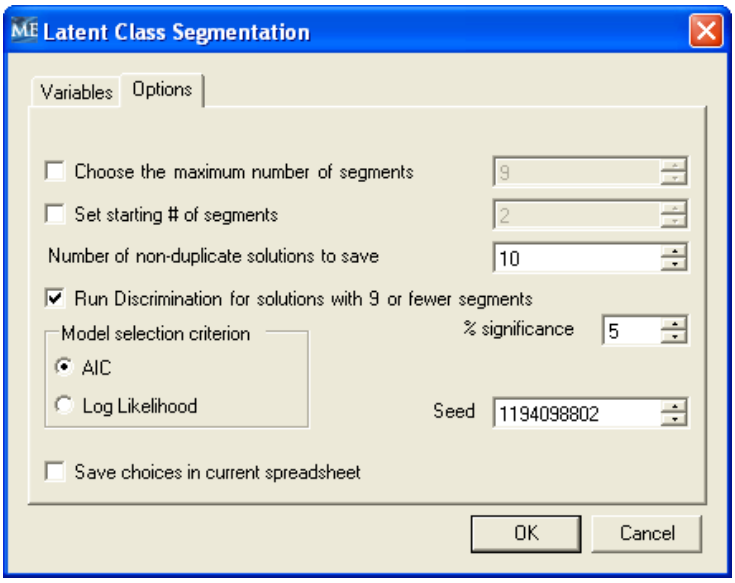
Run Discrimination for solutions with 9 or fewer classes (i.e., segments): Check this box if you would like to characterize how the different segments in a particular segmentation solution vary between themselves on some observable criteria (e.g., age, sex, income). You must have a separate data file containing discriminant analysis data for each respondent used in the segmentation analysis. Specify the significance level of the discriminant functions in the % Significance box. Typically, this number is 5. If you would like a clearer discrimination between segments, use 1%, and if you are willing accept a looser discrimination than the default, use 10%. Currently, the discrimination analysis option can be exercised for segmentation schemes that have 9 or fewer segments.

Model selection criterion: This is an important option for you to specify. We have included two different statistical criteria to help you identify the segmentation scheme that best fits your data: (1) 2LL (2*Loglikelihood) and (2) AIC – Akaike Information criterion. The Akaike criterion is a modification of the LogLikelihood criterion to account for the number of parameters estimated by the model, and is the default criterion AIC favors a more parsimonious segmentation scheme (i.e., one with fewer segments) than the LogLikelihood criterion. We recommend that you use the AIC criterion, unless you want to identify potential niche segments in your market. In that case, we also recommend that you leave the maximum number of segments unspecified.

In our context, both criteria will yield negative numbers, and the closer that a criterion value is to 0, the better a segmentation scheme is in representing your data. The difference in the values of a criterion corresponding to two different segmentation solutions represents the relative likelihoods of those solutions. For example, if the AIC values from two solutions are -2184.85 and -2193.74, the difference is 8.89, which means that the first solution is 8.89 times more likely than the second solution given the data. Informally, for model selection purposes, we suggest that differences that are less than 1% of the base likelihood (here, 100*8.89/2184.85 = 0.4%) are not statistically meaningful differences (although such differences may still be meaningful

operationally in a business context). More formally, one should do chi-square tests of significance based on the criterion values and the number of parameters in the model.

Random Seed: This is a number **generated** by the system automatically to initiate the optimization process to determine which segmentation schemes best fit the data. You can also input a random seed value to initiate the search algorithm from a specific point. The main reason that you might want to provide a random number is to re-create the same solution at a later time. The random seed corresponding to any solution that you retain is also included at the bottom of the output worksheet that contains a segmentation scheme that you have selected for further analysis.



Step 4 (Ic) Choose Solutions to Report

After model execution, ME►XL Latent Class will generate a dialog box shown below to allow you to select which solution(s) you would like to include in your reported output (If you have a very large data set with thousands of observations, and/or you do not specify a maximum number of segments, it could take several minutes or even hours for the program to execute).

In initial stages of exploration, we recommend that you select five or more segmentation schemes (especially those that differ in the number of segments) to analyze. Depending on the parameters chosen in Step 2, you will be presented with a set of useful criteria corresponding to each solution, which should allow you to select the ones that seem most appropriate. The number of solutions listed will be the **Number of duplicate solutions to save** that you specified in the previous dialog box (see above). Also, the reported solutions are listed in decreasing order of performance on the model selection criterion (2*LogLikelihood or AIC).

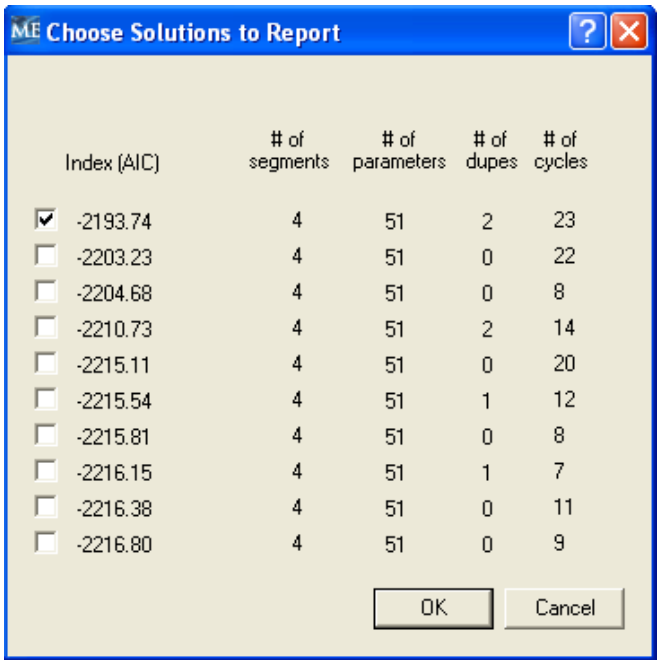
# of segments: This indicates the number of segments in a particular solution found by the program.

# of parameters: This is the number of parameters in the model (not all of them may be estimated – some could be directly from the prior distribution specified in the Bayesian model). The number of parameters increases with

the number of segments. The larger the number of parameters, relative to the number of observations, the less reliable is the statistical value of the results.

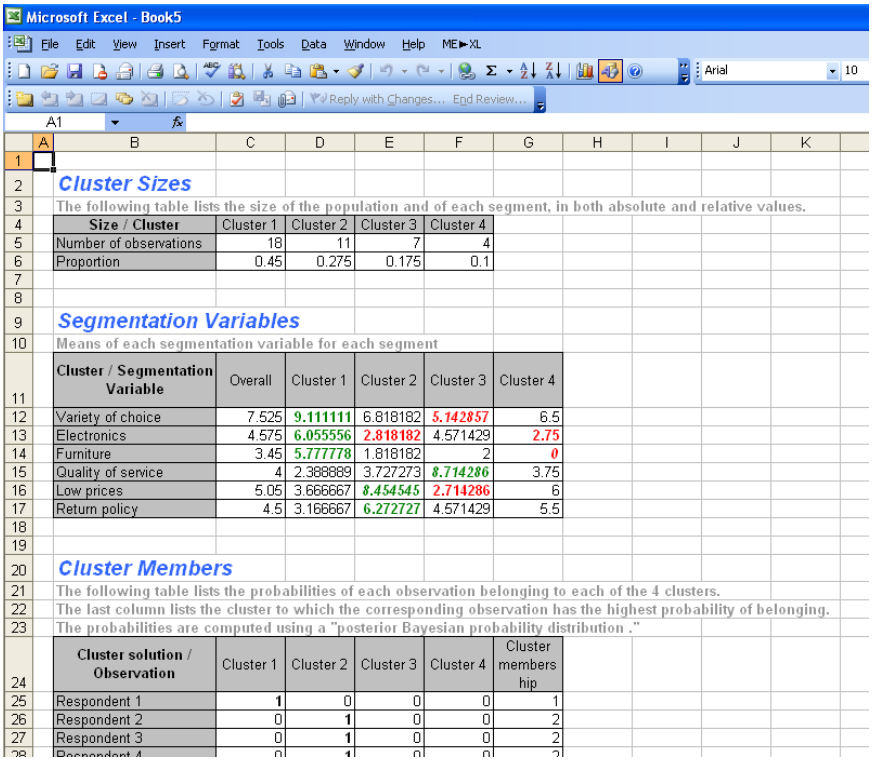
# of dupes: This indicates the number of times a particular solution was found during search (based on the relative errors of measurement specified earlier). It is an indicator of the robustness of a particular solution (i.e., that particular solution would be found by a search algorithm from many different directions). The corresponding solutions will be nearly identical, except in very large data sets, where there may be small differences between solutions that are considered to be duplicates.

# of cycles: This is a technical criterion that lists the number of convergence cycles associated with a particular solution. If this number is 200 or more, it suggests there was no convergence to a solution, and the reported solution is one where the program stopped searching further. It is best to ignore solutions that have 200 or more convergence cycles.



Step 4 (lc) Review Latent Class Output

The output from Latent Class analysis follows the same output format and characteristics as our basic segmentation. Please refer to **Step 4 Interpreting the segmentation results in the main Segmentation tutorial** for instructions on how to interpret ME►XL segmentation output.



Microsoft Excel - Book5

File Edit View Insert Format Tools Data Window Help ME►XL

A1

Cluster Sizes

The following table lists the size of the population and of each segment, in both absolute and relative values.

Size / Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of observations	18	11	7	4
Proportion	0.45	0.275	0.175	0.1

Segmentation Variables

Means of each segmentation variable for each segment

Cluster / Segmentation Variable	Overall	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Variety of choice	7.525	**9.111111**	6.818182	***5.142857***	6.5
Electronics	4.575	**6.055556**	**2.818182**	4.571429	**2.75**
Furniture	3.45	**5.777778**	1.818182	2	***0***
Quality of service	4	2.388889	3.727273	***8.714286***	3.75
Low prices	5.05	3.666667	***8.454545***	**2.714286**	6
Return policy	4.5	3.166667	**6.272727**	4.571429	5.5

Cluster Members

The following table lists the probabilities of each observation belonging to each of the 4 clusters.
The last column lists the cluster to which the corresponding observation has the highest probability of belonging.
The probabilities are computed using a "posterior Bayesian probability distribution ."

Cluster solution / Observation	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster membership
Respondent 1	**1**	0	0	0	1
Respondent 2	0	**1**	0	0	2
Respondent 3	0	**1**	0	0	2

The output from Latent Class analysis can also be used with ME►XL's Classification tool. Please refer to Section 5, **Running classification analyses** in the main Segmentation tutorial, for information about this topic.

Refer to our technical appendix at www.decisionpro.biz for further technical details about the Latent class segmentation model.

Data Limitations:

The data size that can be used for latent class segmentation is only limited by the sizes of your computer's internal memory, or the row or column size restrictions in your version of Excel, whichever is smaller.

Discriminant analysis can currently be performed only for solutions with 9 or fewer segments.