MARKET SEGMENTATION ANALYSIS

ISSUES AND SOLUTIONS (VOLUME 2)

TAPESH LAGHARI

Market Segmentation Analysis: Issues and Solutions (Volume 2)

Market Segmentation Analysis: Issues and Solutions (Volume 2)

Tapesh Laghari

THE INFO LIBRARY

Published by The InfoLibrary,
4/21B, First Floor, E-Block,
Model Town-II,
New Delhi-110009, India

Market Segmentation Analysis: Issues and Solutions (Volume 2)
Tapesh Laghari
ISBN: 978-93-5590-586-4

Table of Contents

Market Profiling and Segmentation

8.1 Identifying Key Characteristics of Market Segments

The aim of the profiling step is to get to know the market segments resulting from the extraction step. Profiling is only required when data-driven market segmentation is used. For commonsense segmentation, the profiles of the segments are predefined. If, for example, age is used as the segmentation variable for the commonsense segmentation, it is obvious that the resulting segments will be age groups. Therefore, Step 6 is not necessary when commonsense segmentation is conducted.

The situation is quite different in the case of data-driven segmentation: users of the segmentation solution may have decided to extract segments on the basis of benefits sought by consumers. Yet – until after the data has been analysed – the defining characteristics of the resulting market segments are unknown. Identifying these defining characteristics of market segments with respect to the segmentation variables is the aim of profiling. Profiling consists of characterising the market segments individually, but also in comparison to the other market segments. If winter tourists in Austria are asked about their vacation activities, most state they are going alpine skiing. Alpine skiing may characterise a segment, but alpine skiing may not differentiate a segment from other market segments.

At the profiling stage, we inspect a number of alternative market segmentation solutions. This is particularly important if no natural segments exist in the data, and either a reproducible or a constructive market segmentation approach has to be taken. Good profiling is the basis for correct interpretation of the resulting segments. Correct interpretation, in turn, is critical to making good strategic marketing decisions.

Data-driven market segmentation solutions are not easy to interpret. Managers have difficulties interpreting segmentation results correctly (Nairn and Bottomley 2003; Bottomley and Nairn 2004); 65% of 176 marketing managers surveyed in a study by Dolnicar and Lazarevski (2009) on the topic of market segmentation state that they have difficulties understanding data-driven market segmentation solutions,

and 71% feel that segmentation analysis is like a black box. A few of the quotes provided by these marketing managers when asked how market segmentation results are usually presented to them are insightful:

- ... as a long report that usually contradicts the results
- ... rarely with a clear Executive Summary
- ...in a rushed slap hazard fashion with the attitude that 'leave the details to us' ...
- The result is usually arranged in numbers and percentages across a few (up to say 10) variables, but mostly insufficiently conclusive.
- ...report or spreadsheet...report with percentages
- ...often meaningless information
- In a PowerPoint presentation with a slick handout

(quotes from the study reported in Dolnicar and Lazarevski 2009).

In the following sections we discuss traditional and graphical statistics approaches to segment profiling. Graphical statistics approaches make profiling less tedious, and thus less prone to misinterpretation.

8.2 Traditional Approaches to Profiling Market Segments

We use the Australian vacation motives data set. Segments were extracted from this data set in Sect. 7.5.4 using the neural gas clustering algorithm with number of segments varied from 3 to 8 and with 20 random restarts. We reload the segmentation solution derived and saved on page 171:

```
R> library("flexclust")
R> data("vacmot", package = "flexclust")
R> load("vacmot-clusters.RData")
```

Data-driven segmentation solutions are usually presented to users (clients, managers) in one of two ways: (1) as high level summaries simplifying segment characteristics to a point where they are misleadingly trivial, or (2) as large tables that provide, for each segment, exact percentages for each segmentation variable. Such tables are hard to interpret, and it is virtually impossible to get a quick overview of the key insights. This is illustrated by Table 8.1. Table 8.1 shows the mean values of the segmentation variables by segment (extracted from the return object using `parameters(vacmot.k6)`), together with the overall mean values. Because the travel motives are binary, the segment means are equal to the percentage of segment members engaging in each activity.

Table 8.1 provides the exact percentage of members of each segment that indicate that each of the travel motives matters to them. To identify the defining characteristics of the market segments, the percentage value of each segment for each segmentation variable needs to be compared with the values of other segments or the total value provided in the far right column.

Table 8.1 Six segments computed with the neural gas algorithm for the Australian travel motives data set. All numbers are percentages of people in the segment or in the total sample agreeing to the motives

	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Seg. 6	Total
Rest and relax	83	96	89	82	98	96	90
Change of surroundings	27	82	73	82	87	77	67
Fun and entertainment	7	71	81	60	95	37	53
Free-and-easy-going	12	65	58	45	87	75	52
Not exceed planned budget	23	100	2	49	84	73	51
Life style of the local people	9	29	30	90	75	80	46
Good company	14	59	40	58	77	55	46
Excitement, a challenge	9	17	39	57	76	36	33
Maintain unspoilt surroundings	9	10	16	7	67	95	30
Cultural offers	4	2	5	96	62	38	28
Luxury / be spoilt	19	24	39	13	89	6	28
Unspoilt nature/natural landscape	10	10	13	15	69	64	26
Intense experience of nature	6	8	9	21	50	58	22
Cosiness/familiar atmosphere	11	24	12	7	49	25	19
Entertainment facilities	5	25	30	14	53	6	19
Not care about prices	8	7	43	19	29	10	18
Everything organised	7	21	15	12	46	9	16
Do sports	8	12	13	10	46	7	14
Health and beauty	5	8	10	8	49	16	12
Realise creativity	2	2	3	8	29	14	8

Using Table 8.1 as the basis of interpreting segments shows that the defining characteristics of segment 2, for example, are: being motivated by rest and relaxation, and not wanting to exceed the planned travel budget. Also, many members of segment 2 care about a change of surroundings, but not about cultural offers, an intense experience of nature, about not caring about prices, health and beauty and realising creativity. Segment 1 is likely to be a response style segment because – for each travel motive – the percentage of segment members indicating that a travel motive is relevant to them is low (compared to the overall percentage of agreement).

Profiling all six market segments based on Table 8.1 requires comparing 120 numbers if each segment's value is only compared to the total (for each one of 20 travel motives, the percentages for six segments have to be compared to the percentage in the total column). If, in addition, each segment's value is compared to the values of other segments, $(6 \times 5)/2 = 15$ pairs of numbers have to be compared for each row of the table. For the complete table with 20 rows, a staggering $15 \times 20 = 300$ pairs of numbers would have to be compared between segments. In total this means 420 comparisons including those between segments only and between segments and the total.

Imagine that the segmentation solution in Table 8.1 is not the only one. Rather, the data analyst presents five alternative segmentation solutions containing six

segments each. A user in that situation would have to compare $5 \times 420 = 2100$ pairs of numbers to be able to understand the defining characteristics of the segments. This is an outrageously tedious task to perform, even for the most astute user.

Sometimes – to deal with the size of this task – information is provided about the statistical significance of the difference between segments for each of the segmentation variables. This approach, however, is not statistically correct. Segment membership is directly derived from the segmentation variables, and segments are created in a way that makes them maximally different, thus not allowing to use standard statistical tests to assess the significance of differences.

8.3 Segment Profiling with Visualisations

Neither the highly simplified, nor the very complex tabular representation typically used to present market segmentation solutions make much use of graphics, although data visualisation using graphics is an integral part of statistical data analysis (Tufte 1983, 1997; Cleveland 1993; Chen et al. 2008; Wilkinson 2005; Kastellec and Leoni 2007). Graphics are particularly important in exploratory statistical analysis (like cluster analysis) because they provide insights into the complex relationships between variables. In addition, in times of big and increasingly bigger data, visualisation offers a simple way of monitoring developments over time. Both McDonald and Dunbar (2012) and Lilien and Rangaswamy (2003) recommend the use of visualisation techniques to make the results of a market segmentation analysis easier to interpret. Haley (1985, p. 227), long before the wide adoption of graphical statistics, pointed out that the same information presented in tabular form is not nearly so insightful. More recently, Cornelius et al. (2010, p. 170) noted, in a review of graphical approaches suitable for interpreting results of market structure analysis, that a single two-dimensional graphical format is preferable to more complex representations that lack intuitive interpretations.

A review of visualisation techniques available for cluster analysis and mixture models is provided by Leisch (2008). Examples of prior use of visualisations of segmentation solutions are given in Reinartz and Kumar (2000), Horneman et al. (2002), Andriotis and Vaughan (2003), Becken et al. (2003), Dolnicar and Leisch (2003, 2014), Bodapati and Gupta (2004), Dolnicar (2004), Beh and Bruyere (2007), and Castro et al. (2007).

Visualisations are useful in the data-driven market segmentation process to inspect, for each segmentation solution, one or more segments in detail. Statistical graphs facilitate the interpretation of segment profiles. They also make it easier to assess the usefulness of a market segmentation solution. The process of segmenting data always leads to a large number of alternative solutions. Selecting one of the possible solutions is a critical decision. Visualisations of solutions assist the data analyst and user with this task.

8.3.1 Identifying Defining Characteristics of Market Segments

A good way to understand the defining characteristics of each segment is to produce a *segment profile plot*. The segment profile plot shows – for all segmentation variables – how each market segment differs from the overall sample. The segment profile plot is the direct visual translation of tables such as Table 8.1.

In figures and tables, segmentation variables do not have to be displayed in the order of appearance in the data set. If variables have a meaningful order in the data set, the order should be retained. If, however, the order of variables is independent of content, it is useful to rearrange variables to improve visualisations.

Table 8.1 sorts the 20 travel motives by the total mean (last column). Another option is to order segmentation variables by similarity of answer patterns. We can achieve this by clustering the columns of the data matrix:

```
R> vacmot.vdist  <- dist(t(vacmot))
R> vacmot.vclust <- hclust(vacmot.vdist, "ward.D2")
```

The `t()` around the data matrix `vacmot` transposes the matrix such that distances between columns rather than rows are computed. Next, hierarchical clustering of the variables is conducted using Ward's method. Figure 8.1 shows the result.

Tourists who are motivated by cultural offers are also interested in the lifestyle of local people. Tourists who care about an unspoilt natural landscape also show interest in maintaining unspoilt surroundings, and seek an intense experience of nature. A segment profile plot like the one in Fig. 8.2 results from:

```
R> barchart(vacmot.k6, shade = TRUE,
+   which = rev(vacmot.vclust$order))
```

Argument `which` specifies the variables to be included, and their order of presentation. Here, all variables are shown in the order suggested by hierarchical clustering of variables. `shade = TRUE` identifies so-called *marker variables* and depicts them in colour. These variables are particularly characteristic for a segment. All other variables are greyed out.

The segment profile plot is a so-called *panel plot*. Each of the six panels represents one segment. For each segment, the segment profile plot shows the cluster centres (centroids, representatives of the segments). These are the numbers contained in Table 8.1. The dots in Fig. 8.2 are identical in each of the six panels, and represent the total mean values for the segmentation variables across all observations in the data set. The dots are the numbers in the last column in Table 8.1. These dots serve as reference points for the comparison of values for each segment with values averaged across all people in the data set.

To make the chart even easier to interpret, marker variables appear in colour (solid bars). The remaining segmentation variables are greyed out. The definition of marker variables in the segment profile plot used by default in `barchart()` is suitable for binary variables, and takes into account the absolute and relative difference of the segment mean to the total mean. Marker variables are defined as variables which deviate by more than 0.25 from the overall mean. For example, a

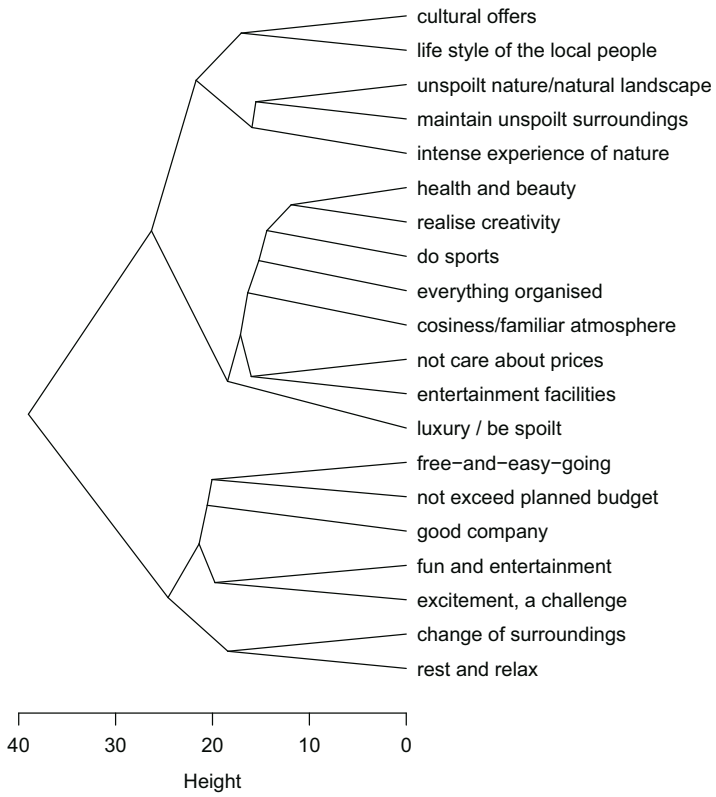


Fig. 8.1 Hierarchical clustering of the segmentation variables of the Australian travel motives data set using Ward's method

variable with a total sample mean of 0.20, and a segment mean of 0.60 qualifies as marker variable ($0.20 + 0.25 = 0.45 < 0.60$). Such a large absolute difference is hard to obtain for segmentation variables with very low sample means. A relative difference of 50% from the total mean, therefore, also makes the variable a marker variable.

The deviation figures of 0.25 and 50% have been empirically determined to indicate substantial differences on the basis of inspecting many empirical data sets, but are ultimately arbitrary and, as such, can be chosen by the data analyst and user as they see fit. In particular if the segmentation variables are not binary, different thresholds for defining a marker variable need to be specified.

Looking at the travel motive of HEALTH AND BEAUTY in Fig. 8.2 makes it obvious that this is not a mainstream travel motive for tourists. This segmentation variable has a sample mean of 0.12; this means that only 12% of all the people who participated in the survey indicated that HEALTH AND BEAUTY was a travel motive for them. For segments with HEALTH AND BEAUTY outside of the interval 0.12 ± 0.06 this vacation activity will be considered a marker variable, because 0.06 is 50% of 0.12.

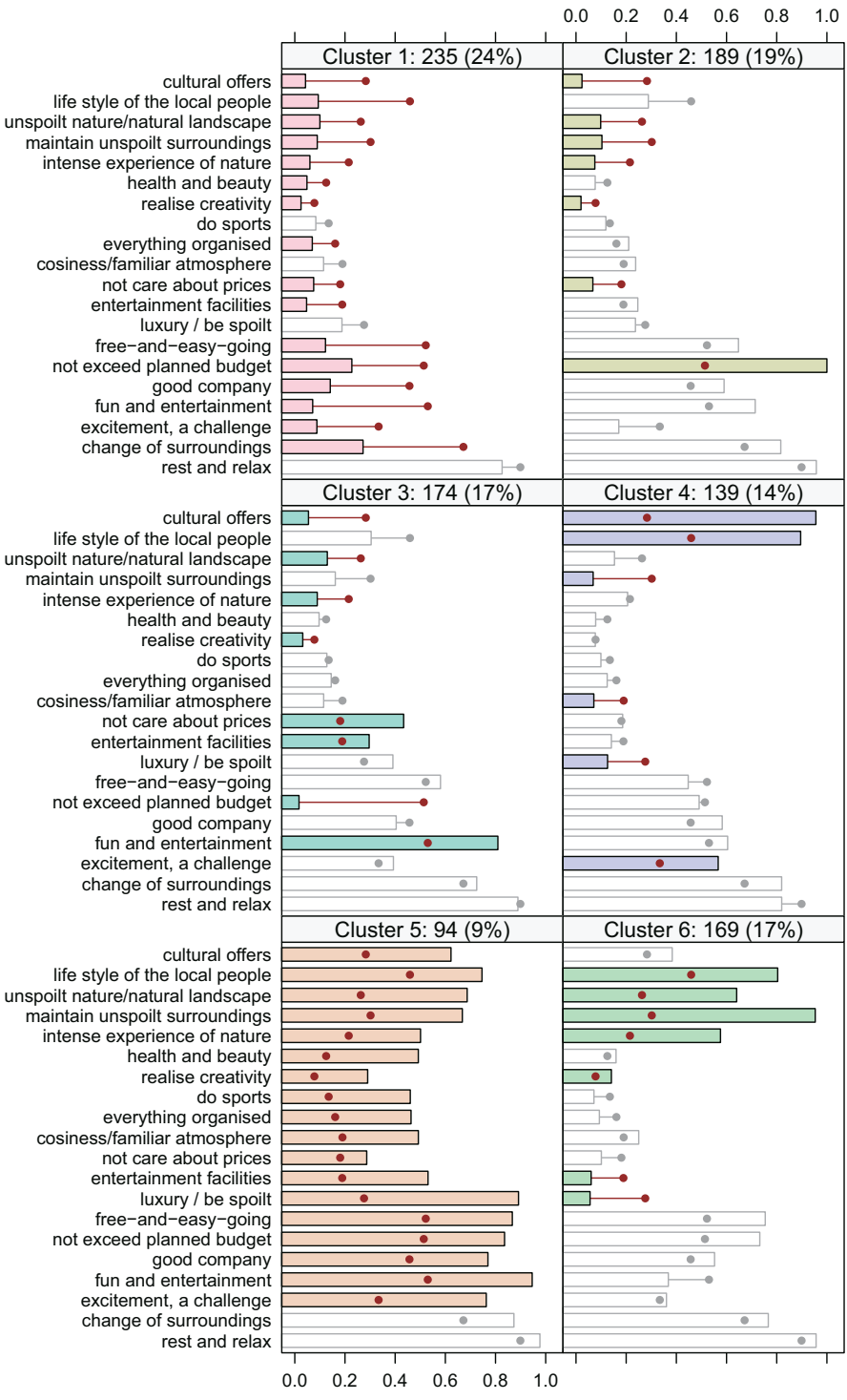


Fig. 8.2 Segment profile plot for the six-segment solution of the Australian travel motives data set

The segment profile plot in Fig. 8.2 contains the same information as Table 8.1: the percentage of segment members indicating that each of the travel motives matters to them. Marker variables are highlighted in colour. As can be seen, a segmentation solution presented using a segment profile plot (such as the one shown in Fig. 8.2) is much easier and faster to interpret than when it is presented as a table, no matter how well the table is structured. We see that members of segment 2 are characterised primarily by not wanting to exceed their travel budget. Members of segment 4 are interested in culture and local people; members of segment 3 want fun and entertainment, entertainment facilities, and do not care about prices. Members of segment 6 see nature as critical to their vacations. Finally, segments 1 and 5 have to be interpreted with care as they are likely to represent response style segments.

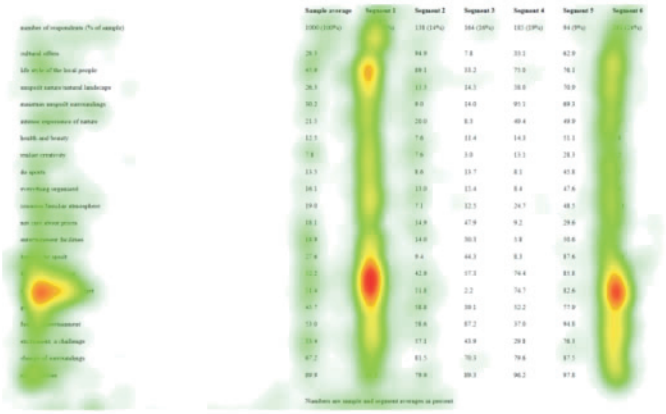
An eye tracking study conducted by Nazila Babakhani as part of her PhD studies investigated differences in people's ability to interpret complex data analysis results from market segmentation studies presented in traditional tabular versus graphical statistics format. Participants saw one of three types of presentations of segmentation results: a table; an improved table with key information bolded; and a segment profile plot. Processing time of information was the key variable of interest. Eye tracking plots indicate how long a person looked at something.

A heat map showing how long one person was looking at each section of the table or figure is shown in Fig. 8.3. We see that this person worked harder to extract information from the tables; the heat maps of the tables contain more yellow and red colouring, representing longer looking times. Longer looking times indicate more cognitive effort being invested in the interpretation of the tables. Also, the person looked at a higher proportion of the table; they were processing a larger area in the attempt to answer the question. In contrast, the heat map of the segment profile plot in Fig. 8.3 shows that the person did not need to look as long to find the answer. They also inspected a smaller surface area. The heat map suggests that it took less effort to find the information required to answer the question. It is therefore well worth spending some extra time on presenting results of a market segmentation analysis as a well designed graph. Good visualisations facilitate interpretation by managers who make long-term strategic decisions based on segmentation results. Such long-term strategic decisions imply substantial financial commitments to the implementation of a segmentation strategy. Good visualisations, therefore, offer an excellent return on investment.

8.3.2 Assessing Segment Separation

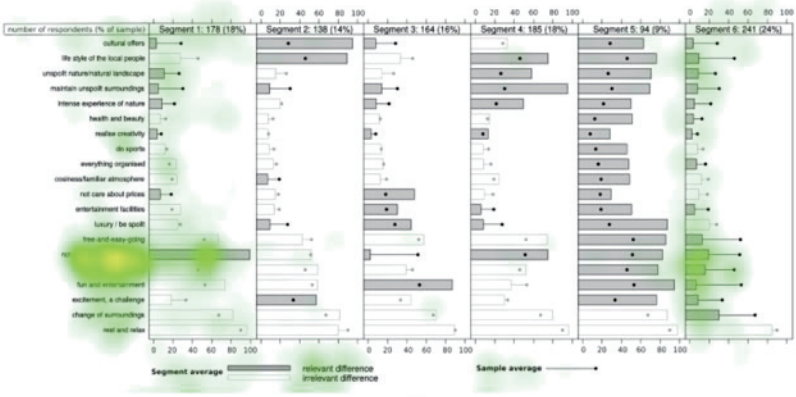
Segment separation can be visualised in a *segment separation plot*. The segment separation plot depicts – for all relevant dimensions of the data space – the overlap of segments.

Segment separation plots are very simple if the number of segmentation variables is low, but become complex as the number of segmentation variables increases. But even in such complex situations, segment separation plots offer data analysts and users a quick overview of the data situation, and the segmentation solution.



(a)

(b)



(c)

Fig. 8.3 One person's eye tracking heat maps for three alternative ways of presenting segmentation results. (**a**) Traditional table. (**b**) Improved table. (**c**) Segment profile plot

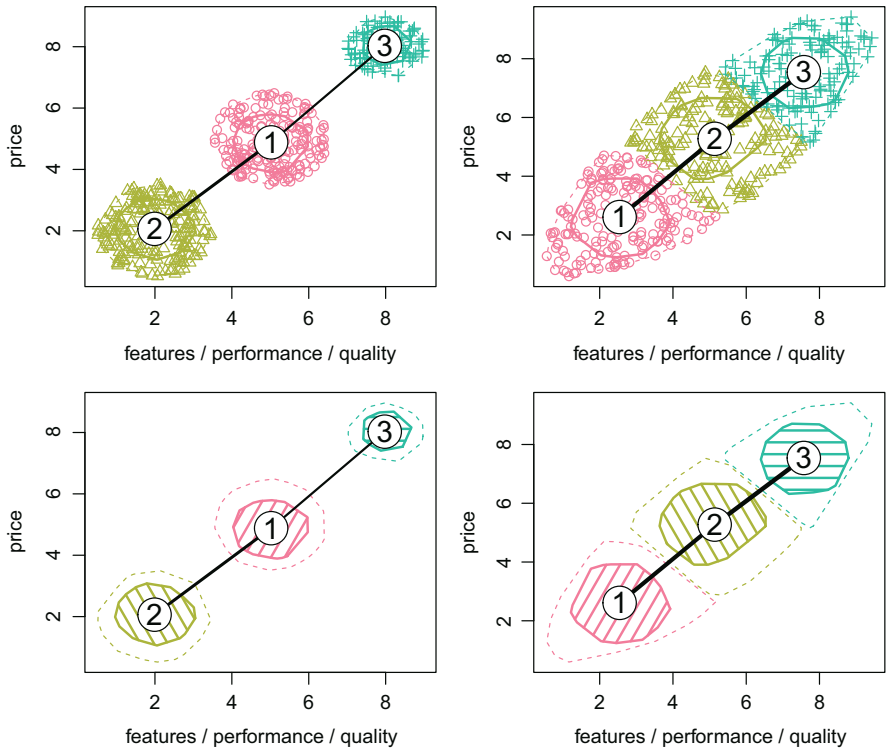


Fig. 8.4 Segment separation plot including observations (first row) and not including observations (second row) for two artificial data sets: three natural, well-separated clusters (left column); one elliptic cluster (right column)

Examples of segment separation plots are provided in Fig. 8.4 for two different data sets (left compared to right column). These plots are based on two of the artificial data sets used in Table 2.3: the data set that contains three distinct, well-separated segments, and the data set with an elliptic data structure. The segment separation plot consists of (1) a scatter plot of the (projected) observations coloured by segment membership and the (projected) cluster hulls, and (2) a neighbourhood graph.

The artificial data visualised in Fig. 8.4 are two-dimensional. So no projection is required. The original data is plotted in a scatter plot in the top row of Fig. 8.4. The colour of the observations indicates true segment membership. The different cluster hulls indicate the shape and spread of the true segments. Dashed cluster hulls contain (approximately) all observations. Solid cluster hulls contain (approximately) half of the observations. The bottom row of Fig. 8.4 omits the data, and displays cluster hulls only.

Neighbourhood graphs (black lines with numbered nodes) indicate similarity between segments (Leisch 2010). The segment solutions in Fig. 8.4 contain three

segments. Each plot, therefore, contains three numbered nodes plotted at the position of the segment centres. The black lines connect segment centres, and indicate similarity between segments. A black line is only drawn between two segment centres if they are the two closest segment centres for at least one observation (consumer). The width of the black line is thicker if more observations have these two segment centres as their two closest segment centres.

As can be seen in Fig. 8.4, the neighbourhood graphs for the two data sets are quite similar. We need to add either the observations or the cluster hulls to assess the separation between segments.

For the two data sets used in Fig. 8.4, the two dimensions representing the segmentation variables can be directly plotted. This is not possible if 20-dimensional travel motives data serve as segmentation variables. In such a situation, the 20-dimensional space needs to be projected onto a small number of dimensions to create a segment separation plot. We can use a number of different projection techniques, including some which maximise separation (Hennig 2004), and principal components analysis (see Sect. 6.5). We calculate principal components analysis for the Australian travel motives data set with the following command:

```
R> vacmot.pca <- prcomp(vacmot)
```

This provides the rotation applied to the original data when creating our segment separation plot. We use the segmentation solution obtained from neural gas on page 171, and create a segment separation plot for this solution:

```
R> plot(vacmot.k6, project = vacmot.pca, which = 2:3,
+   xlab = "principal component 2",
+   ylab = "principal component 3")
R> projAxes(vacmot.pca, which = 2:3)
```

Figure 8.5 contains the resulting plot. Argument `project` uses the principal components analysis projection. Argument `which` selects principal components 2 and 3, and `xlab` and `ylab` assign labels to axes. Function `projAxes()` enhances the segment separation plot by adding directions of the projected segmentation variables. The enhanced version combines the advantages of the segment separation plot with the advantages of perceptual maps.

Due to the overlap of market segments (and the sample size of $n = 1000$), the plot in Fig. 8.5 is messy and hard to read. Modifying colours (argument `col`), omitting observations (`points = FALSE`), and highlighting only the inner area of each segment (`hull.args = list(density = 10)`, where density specifies how many lines shade the area) leads to a cleaner version (Fig. 8.6):

```
R> plot(vacmot.k6, project = vacmot.pca, which = 2:3,
+   col = flxColors(1:6, "light"),
+   points = FALSE, hull.args = list(density = 10),
+   xlab = "principal component 2",
+   ylab = "principal component 3")
R> projAxes(vacmot.pca, which = 2:3, col = "darkblue",
+   cex = 1.2)
```

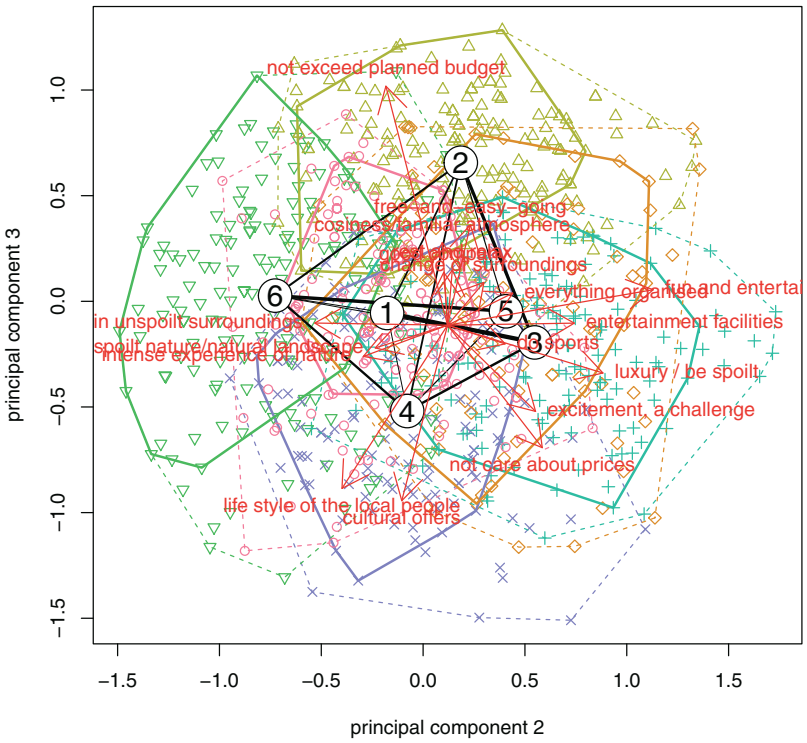


Fig. 8.5 Segment separation plot using principal components 2 and 3 for the Australian travel motives data set

The plot is still not trivial to assess, but it is easier to interpret than the segment separation plot shown in Fig. 8.5 containing additional information. Figure 8.6 is hard to interpret, because natural market segments are not present. This difficulty in interpretation is due to the data, not the visualisation. And the data used for this plot is very representative of consumer data.

Figure 8.6 shows the existence of a market segment (segment 6, green shaded area) that cares about maintaining unspoilt surroundings, unspoilt nature, and wants to intensely experience nature when on vacations. Exactly opposite is segment 3 (cyan shaded area) wanting luxury, wanting to be spoilt, caring about fun, entertainment and the availability of entertainment facilities, and not caring about prices. Another segment on top of the plot in Fig. 8.6 (segment 2, olive shaded area)

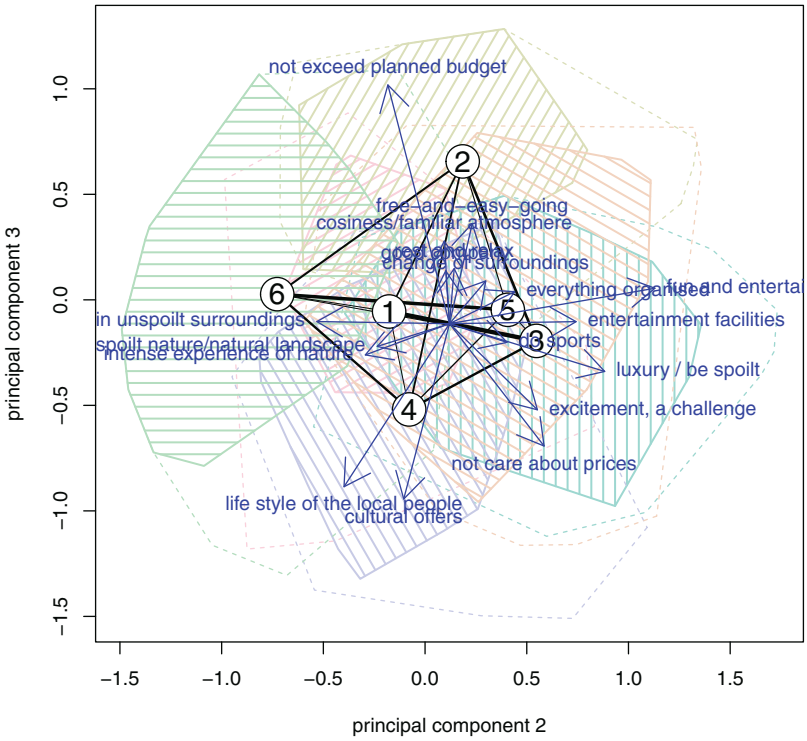


Fig. 8.6 Segment separation plot using principal components 2 and 3 for the Australian travel motives data set without observations

is characterised by one single feature only: members of this market segment do not wish to exceed their planned travel budget. Opposite to this segment, at the bottom of the plot is segment 4 (blue shaded area), members of which care about the life style of local people and cultural offers.

Each segment separation plot only visualises one possible projection. So, for example, the fact that segments 1 and 5 in this particular projection overlap with other segments does not mean that these segments overlap in all projections. However, the fact that segments 6 and 3 are well-separated in this projection does allow the conclusion – based on this single projection only – that they represent distinctly different tourists in terms of the travel motives.

8.4 Step 6 Checklist

Task	Who is responsible?	Completed?
Use the selected segments from Step 5.		☐
Visualise segment profiles to learn about what makes each segment distinct.		☐
Use knock-out criteria to check if any of the segments currently under consideration should already be eliminated because they do not comply with the knock-out criteria.		☐
Pass on the remaining segments to Step 7 for describing.		☐

References

Andriotis K, Vaughan RD (2003) Urban residents' attitudes toward tourism development: the case of Crete. J Travel Res 42(2):172–185

Becken S, Simmons D, Frampton C (2003) Segmenting tourists by their travel pattern for insights into achieving energy efficiency. J Travel Res 42(1):48–53

Beh A, Bruyere BL (2007) Segmentation by visitor motivation in three Kenyan national reserves. Tour Manag 28(6):1464–1471

Bodapati AV, Gupta S (2004) The recoverability of segmentation structure from store-level aggregate data. J Mark Res 41(3):351–364

Bottomley P, Nairn A (2004) Blinded by science: the managerial consequences of inadequately validated cluster analysis solutions. Int J Mark Res 46(2):171–187

Castro CB, Armario EM, Ruiz DM (2007) The influence of market heterogeneity on the relationship between a destination's image and tourists' future behavior. Tour Manag 28(1):175–187

Chen CH, Härdle WK, Unwin A (2008) Handbook of data visualization. Springer handbooks of computational statistics. Springer, Heidelberg

Cleveland W (1993) Visualizing data. Hobart Press, Summit

Cornelius B, Wagner U, Natter M (2010) Managerial applicability of graphical formats to support positioning decisions. Journal für Betriebswirtschaft 60(3):167–201

Dolnicar S (2004) Beyond commonsense segmentation: a systematics of segmentation approaches in tourism. J Travel Res 42(3):244–250

Dolnicar S, Leisch F (2003) Winter tourist segments in Austria: identifying stable vacation styles for target marketing action. J Travel Res 41(3):281–193

Dolnicar S, Leisch F (2014) Using graphical statistics to better understand market segmentation solutions. Int J Mark Res 56(2):97–120.

Dolnicar S, Lazarevski K (2009) Methodological reasons for the theory/practice divide in market segmentation. J Mark Manag 25(3–4):357–373

Haley RI (1985) Developing effective communications strategy – a benefit segmentation approach. Wiley, New York

Hennig C (2004) Asymmetric linear dimension reduction for classification. J Comput Graph Stat 13(4):930–945

Horneman L, Carter R, Wei S, Ruys H (2002) Profiling the senior traveler: an Australian perspective. J Travel Res 41(1):23–37

Kastellec JP, Leoni EL (2007) Using graphs instead of tables in political science. Perspect Polit 5(4):755–771

Leisch F (2008) Visualization of cluster analysis and finite mixture models. In: Chen CH, Härdle W, Unwin A (eds) Handbook of data visualization. Springer handbooks of computational statistics. Springer, Berlin

Leisch F (2010) Neighborhood graphs, stripes and shadow plots for cluster visualization. Stat Comput 20(4):457–469

Lilien GL, Rangaswamy A (2003) Marketing engineering: computer-assisted marketing analysis and planning, 2nd edn. Prentice Hall, Upper Saddle River

McDonald M, Dunbar I (2012) Market segmentation: how to do it and how to profit from it, 4th edn. Wiley, Chichester

Nairn A, Bottomley P (2003) Something approaching science? Cluster analysis procedures in the CRM era. Int J Mark Res 45(2):241–261

Reinartz W, Kumar V (2000) On the profitability of long-life customers in a noncontractual setting: an empirical investigation and implications for marketing. J Mark 64(4):17–35

Tufte ER (1983) The visual display of quantitative information. Graphics Press, Cheshire

Tufte ER (1997) Visual explanations. Graphics Press, Cheshire

Wilkinson L (2005) The grammar of graphics. Springer, New York

Describing Market Segments

9.1 Developing a Complete Picture of Market Segments

Segment profiling is about understanding differences in segmentation variables across market segments. Segmentation variables are chosen early in the market segmentation analysis process: conceptually in Step 2 (specifying the ideal target segment), and empirically in Step 3 (collecting data). Segmentation variables form the basis for extracting market segments from empirical data.

Step 7 (describing segments) is similar to the profiling step. The only difference is that the variables being inspected have *not* been used to extract market segments. Rather, in Step 7 market segments are described using *additional* information available about segment members. If committing to a target segment is like a marriage, profiling and describing market segments is like going on a number of dates to get to know the potential spouse as well as possible in an attempt to give the marriage the best possible chance, and avoid nasty surprises down the track. As van Raaij and Verhallen (1994, p. 58) state: segment . . . should be further described and typified by crossing them with all other variables, i.e. with psychographic . . . , demographic and socio-economic variables, media exposure, and specific product and brand attitudes or evaluations.

For example, when conducting a data-driven market segmentation analysis using the Australian travel motives data set (this is the segmentation solution we saved on page 171; the data is described in Appendix C.4), profiling means investigating differences between segments with respect to the travel motives themselves. These profiles are provided in Fig. 8.2. The segment description step uses additional information, such as segment members' age, gender, past travel behaviour, preferred vacation activities, media use, use of information sources during vacation planning, or their expenditure patterns during a vacation. These additional variables are referred to as *descriptor variables*.

Good descriptions of market segments are critical to gaining detailed insight into the nature of segments. In addition, segment descriptions are essential for the

development of a customised marketing mix. Imagine, for example, wanting to target segment 4 which emerged from extracting segments from the Australian travel motives data set. Step 6 of the segmentation analysis process leads to the insight that members of segment 4 care about nature. Nothing is known, however, about how old these people are, if they have children, how high their discretionary income is, how much money they spend when they go on vacation, how often they go on vacation, which information sources they use when they plan their vacation, and how they can be reached. If segment description reveals, for example, that members of this segment have a higher likelihood of volunteering for environmental organisations, and regularly read National Geographic, tangible ways of communicating with segment 4 have been identified. This knowledge is important for the development of a customised marketing mix to target segment 4.

We can study differences between market segments with respect to descriptor variables in two ways: we can use descriptive statistics including visualisations, or we can analyse data using inferential statistics. The marketing literature traditionally relies on statistical testing, and tabular presentations of differences in descriptor variables. Visualisations make segment description more user-friendly.

9.2 Using Visualisations to Describe Market Segments

A wide range of charts exist for the visualisation of differences in descriptor variables. Here, we discuss two basic approaches suitable for nominal and ordinal descriptor variables (such as gender, level of education, country of origin), or metric descriptor variables (such as age, number of nights at the tourist destinations, money spent on accommodation).

Using graphical statistics to describe market segments has two key advantages: it simplifies the interpretation of results for both the data analyst and the user, and integrates information on the statistical significance of differences, thus avoiding the over-interpretation of insignificant differences. As Cornelius et al. (2010, p. 197) put it: Graphical representations . . . serve to transmit the very essence of marketing research results. The same authors also find – in a survey study with marketing managers – that managers prefer graphical formats, and view the intuitiveness of graphical displays as critically important. Section 8.3.1 provides an illustration of the higher efficiency with which people process graphical as opposed to tabular results.

9.2.1 Nominal and Ordinal Descriptor Variables

When describing differences between market segments in one single nominal or ordinal descriptor variable, the basis for all visualisations and statistical tests is a cross-tabulation of segment membership with the descriptor variable. For the

Australian travel motives data set (see Appendix C.4), data frame `vacmotdesc` contains several descriptor variables. These descriptor variables are automatically loaded with the Australian travel motives data set. To describe market segments, we need the segment membership for all respondents. We store segment membership in helper variable `C6`:

```
R> C6 <- clusters(vacmot.k6)
```

The sizes of the market segments are

```
R> table(C6)
```

```
C6
  1   2   3   4   5   6
235 189 174 139  94 169
```

The easiest approach to generating a cross-tabulation is to add segment membership as a categorical variable to the data frame of descriptor variables. Then we can use the formula interface of R for testing or plotting:

```
R> vacmotdesc$C6 <- as.factor(C6)
```

The following R command gives the number of females and males across market segments:

```
R> C6.Gender <- with(vacmotdesc,
+   table("Segment number" = C6, Gender))
R> C6.Gender
```

```
               Gender
Segment number Male Female
             1  125    110
             2   86    103
             3   94     80
             4   78     61
             5   47     47
             6   82     87
```

A visual inspection of this cross-tabulation suggests that there are no huge gender differences across segments. The upper panel in Fig. 9.1 visualises this cross-tabulation using a stacked bar chart. The *y*-axis shows segment sizes. Within each bar, we can easily how many are male and how many are female. We cannot, however, compare the proportions of men and women easily across segments. Comparing proportions is complicated if the segment sizes are unequal (for example, segments 1 and 5). A solution is to draw the bars for women and men next to one another rather than stacking them (not shown). The disadvantage of this approach is that the absolute sizes of the market segments can no longer be directly seen on the *y*-axis. The *mosaic plot* offers a solution to this problem.

The mosaic plot also visualises cross-tabulations (Hartigan and Kleiner 1984; Friendly 1994). The width of the bars indicates the absolute segment size. The column for segment 5 of the Australian travel motives data set – containing 94

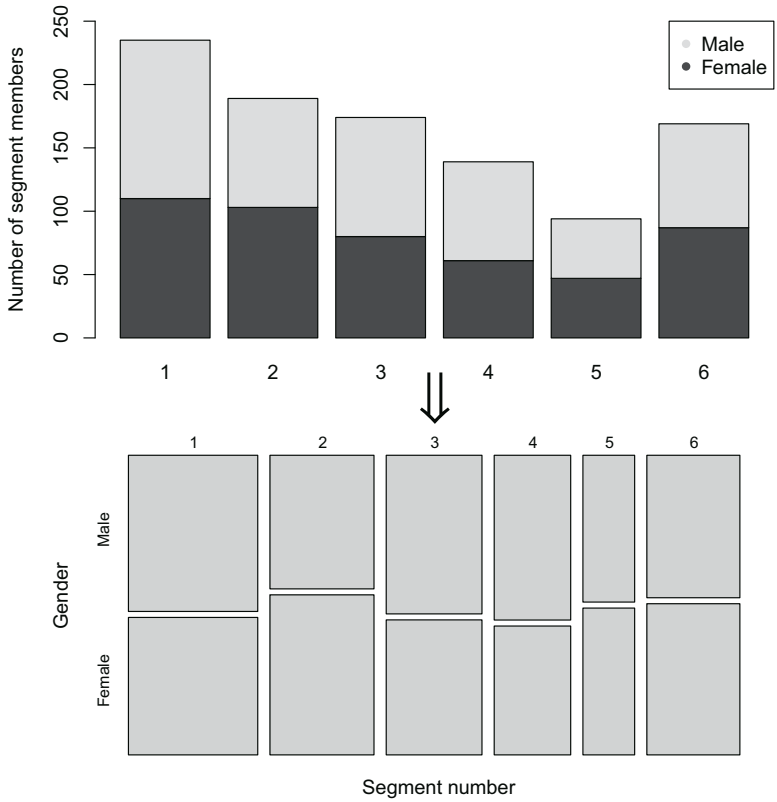


Fig. 9.1 Comparison of a stacked bar chart and a mosaic plot for the cross-tabulation of segment membership and gender for the Australian travel motives data set

respondents or 9% of the sample – is much narrower in the bottom plot of Fig. 9.1 than the column for segment 1 – containing 235 respondents or 24% of the sample.

Each column consists of rectangles. The height of the rectangles represents the proportion of men or women in each segment. Because all columns have the same total height, the height of the bottom rectangles is in the same position for two segments with the same proportion of men and women (even if the absolute number of men and women differs substantially). Because the width of the columns represents the total segment sizes, the area of each cell is proportional to the size of the corresponding cell in the table.

Mosaic plots can also visualise tables containing more than two descriptor variables and integrate elements of inferential statistics. This helps with interpretation. Colours of cells can highlight where observed frequencies are different from expected frequencies under the assumption that the variables are independent. Cell colours are based on the standardised difference between the expected and observed frequencies. Negative differences mean that observed are lower than expected

frequencies. They are coloured in red. Positive differences mean that observed are higher than expected frequencies. They are coloured in blue. The saturation of the colour indicates the absolute value of the standardised difference. Standardised differences follow asymptotically a standard normal distribution. Standard normal random variables lie within $[-2, 2]$ with a probability of $\approx$95%, and within $[-4, 4]$ with a probability of $\approx$99.99%. Standardised differences are equivalent to the standardised Pearson residuals from a log-linear model assuming independence between the two variables.

By default, function `mosaicplot()` in R uses dark red cell colouring for contributions or standardised Pearson residuals smaller than -4, light red if contributions are smaller than -2, white (not interesting) between -2 and 2, light blue if contributions are larger than 2, and dark blue if they are larger than 4. Figure 9.2 shows such a plot with the colour coding included in the legend.

In Fig. 9.2 all cells are white, indicating that the six market segments extracted from the Australian travel motives data set do not significantly differ in gender distribution. The proportion of female and male tourists is approximately the same across segments. The dashed and solid borders of the rectangles indicate that the number of respondents in those cells are either lower than expected (dashed

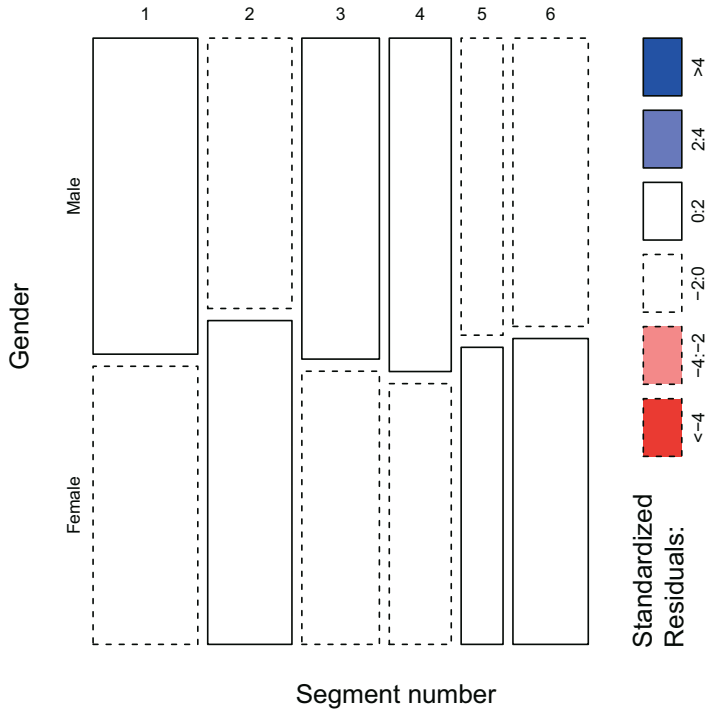


Fig. 9.2 Shaded mosaic plot for cross-tabulation of segment membership and gender for the Australian travel motives data set

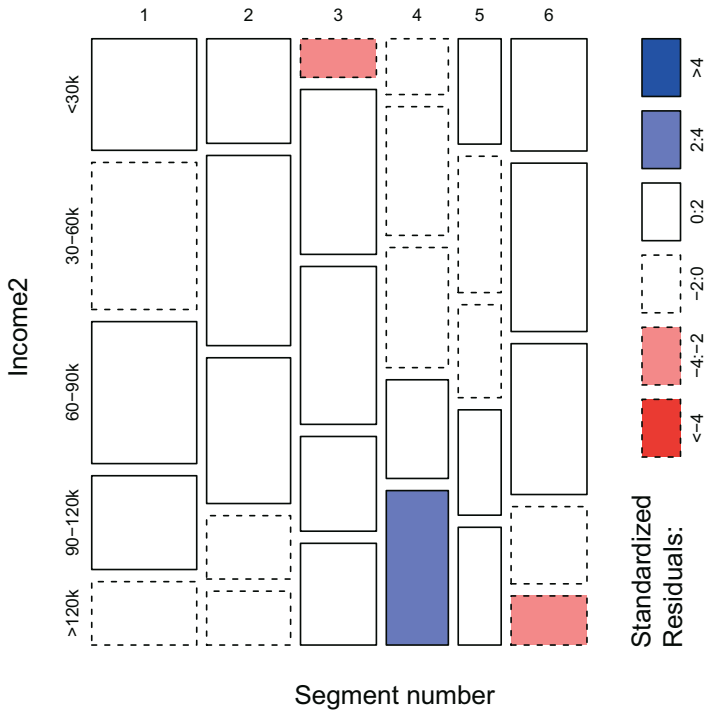


Fig. 9.3 Shaded mosaic plot for cross-tabulation of segment membership and income for the Australian travel motives data set

borders), or higher than expected (solid black borders). But, irrespective of the borders, white rectangles mean differences are statistically insignificant.

Figure 9.3 shows that segment membership and income are moderately associated. The top row corresponds to the lowest income category (less than AUD 30,000 per annum). The bottom row corresponds to the highest income category (more than AUD 120,000 per annum). The remaining three categories represent AUD 30,000 brackets in-between those two extremes. We learn that members of segment 4 (column 4 in Fig. 9.3) – those motivated by cultural offers and interested in local people – earn more money. Low income tourists (top row of Fig. 9.3) are less frequently members of market segment 3, those who do not care about prices and instead seek luxury, fun and entertainment, and wish to be spoilt when on vacation. Segment 6 (column 6 in Fig. 9.3) – the nature loving segment – contains fewer members on very high incomes.

Figure 9.4 points to a strong association between travel motives and stated moral obligation to protect the environment. The moral obligation score results from averaging the answers to 30 survey questions asking respondents to indicate how obliged they feel to engage in a range of environmentally friendly behaviours at home (including not to litter, to recycle rubbish, to save water and energy; see

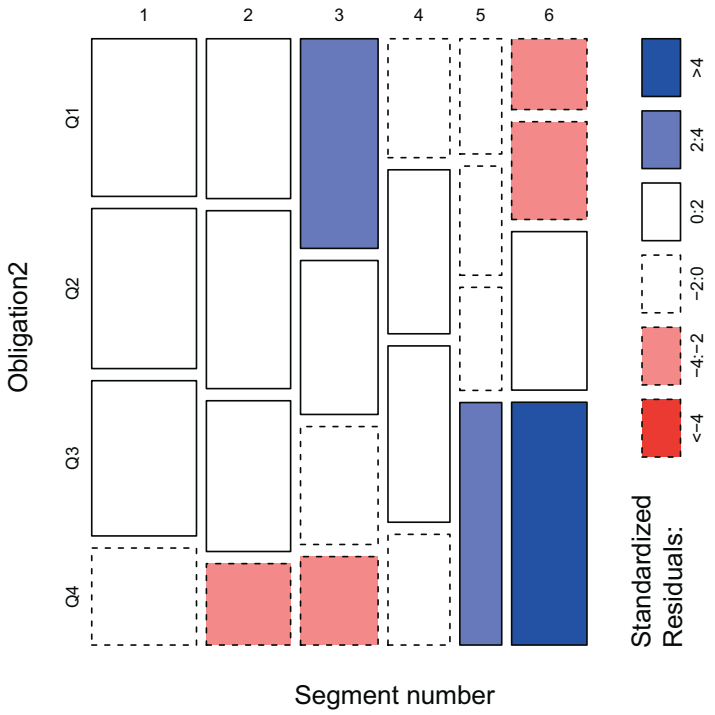


Fig. 9.4 Shaded mosaic plot for cross-tabulation of segment membership and moral obligation to protect the environment for the Australian travel motives data set

Dolnicar and Leisch 2008 for details). The moral obligation score is numeric and ranges from 1 (lowest moral obligation) to 5 (highest moral obligation) because survey respondents had five answer options. The summated score ranges from 30 to 150, and is re-scaled to 1 to 5 by dividing through 30. We provide an illustration of how this descriptor variable can be analysed in its original metric format in Sect. 9.2.2. To create the mosaic plot shown in Fig. 9.4, we cut the moral obligation score into quarters containing 25% of respondents each, ranging from Q1 (low moral obligation) to Q4 (high moral obligation). Variable `Obliged2` contains this re-coded descriptor variable.

Figure 9.4 graphically illustrates the cross-tabulation, associating segment membership and stated moral obligation to protect the environment in a mosaic plot. Segment 3 (column 3 of Fig. 9.4) – whose members seek entertainment – contains significantly more members with low stated moral obligation to behave in an environmentally friendly way. Segment 3 also contains significantly fewer members in the high moral obligation category. The exact opposite applies to segment 6. Members of this segment are motivated by nature, and plotted in column 6 of Fig. 9.4. Being a member of segment 6 implies a positive association with high

moral obligation to behave environmentally friendly, and a negative association with membership in the lowest moral obligation category.

9.2.2 Metric Descriptor Variables

R package lattice (Sarkar 2008) provides conditional versions of most standard R plots. An alternative implementation for conditional plots is available in package ggplot2 (Wickham 2009). *Conditional* in this context means that the plots are divided in sections (panels, facets), each presenting the results for a subset of the data (for example, different market segments). Conditional plots are well-suited for visualising differences between market segments using metric descriptor variables. R package lattice generated the segment profile plot in Sect. 8.3.1.

In the context of segment description, this R package can display the age distribution of all segments comparatively. Or visualise the distribution of the (original metric) moral obligation scores for members of each segment.

To have segment names (rather than only segment numbers) displayed in the plot, we create a new factor variable by pasting together the word `"Segment"` and the segment numbers from `C6`. We then generate a histogram for age for each segment. Argument `as.table` controls whether the panels are included by starting on the top left (`TRUE`) or bottom left (`FALSE`, the default).

```
R> library("lattice")
R> histogram(~ Age | factor(paste("Segment", C6)),
+   data = vacmotdesc, as.table = TRUE)
```

We do the same for moral obligation:

```
R> histogram(~ Obligation | factor(paste("Segment",C6)),
+   data = vacmotdesc, as.table = TRUE)
```

The resulting histograms are shown in Figs. 9.5 (for age) and 9.6 (for moral obligation). In both cases, the differences between market segments are difficult to assess just by looking at the plots.

We can gain additional insights by using a parallel box-and-whisker plot; it shows the distribution of the variable separately for each segment. We create this parallel box-and-whisker plot for age by market segment in R with the following command:

```
R> boxplot(Age ~ C6, data = vacmotdesc,
+   xlab = "Segment number", ylab = "Age")
```

where arguments `xlab` and `ylab` customise the axis labels.

Figure 9.7 shows the resulting plot. As expected – given the histograms inspected previously – differences in age across segments are minor. The median age of members of segment 5 is lower, that of segment 6 members is higher. These visually detected differences in descriptors need to be subjected to statistical testing.

Like mosaic plots, parallel box-and-whisker plots can the incorporate elements of statistical hypothesis testing. For example, we can make the width of the

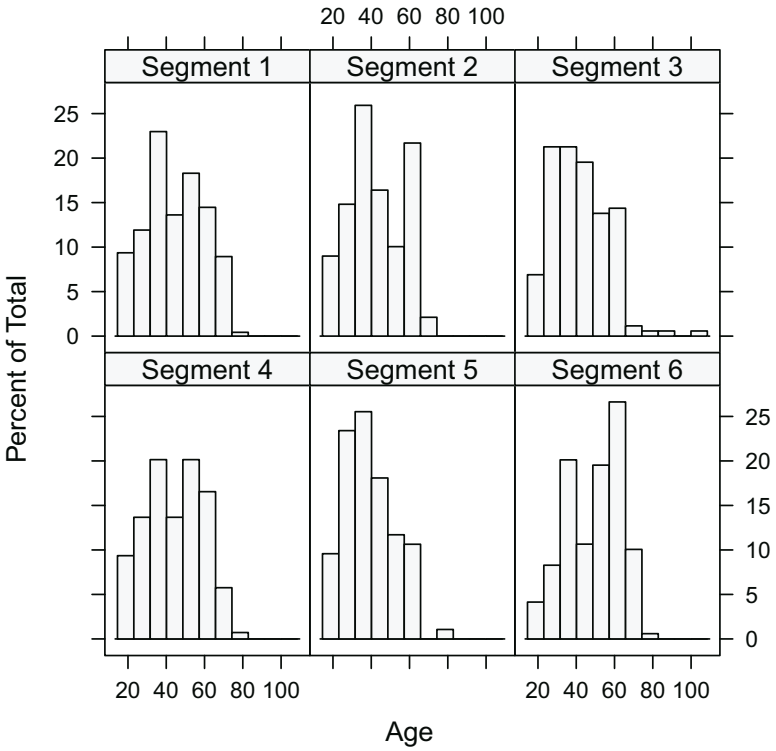


Fig. 9.5 Histograms of age by segment for the Australian travel motives data set

boxes proportional to the size of market segments (`varwidth = TRUE`), and include 95% confidence intervals for the medians (`notch = TRUE`) using the R command:

```
R> boxplot(Obligation ~ C6, data = vacmotdesc,
+    varwidth = TRUE, notch = TRUE,
+    xlab = "Segment number",
+    ylab = "Moral obligation")
```

Figure 9.8 contains the resulting parallel box-and-whisker plot. This version illustrates that segment 5 is the smallest; its box is the narrowest. Segment 1 is the largest. Moral obligation to protect the environment is highest among members of segment 6.

The notches in this version of the parallel box-and-whisker plot correspond to 95% confidence intervals for the medians. If the notches for different segments do not overlap, a formal statistical test will usually result in a significant difference. We can conclude from the inspection of the plot in Fig. 9.8 alone, therefore, that there is a significant difference in moral obligation to protect the environment between members of segment 3 and members of segment 6. The notches for those two

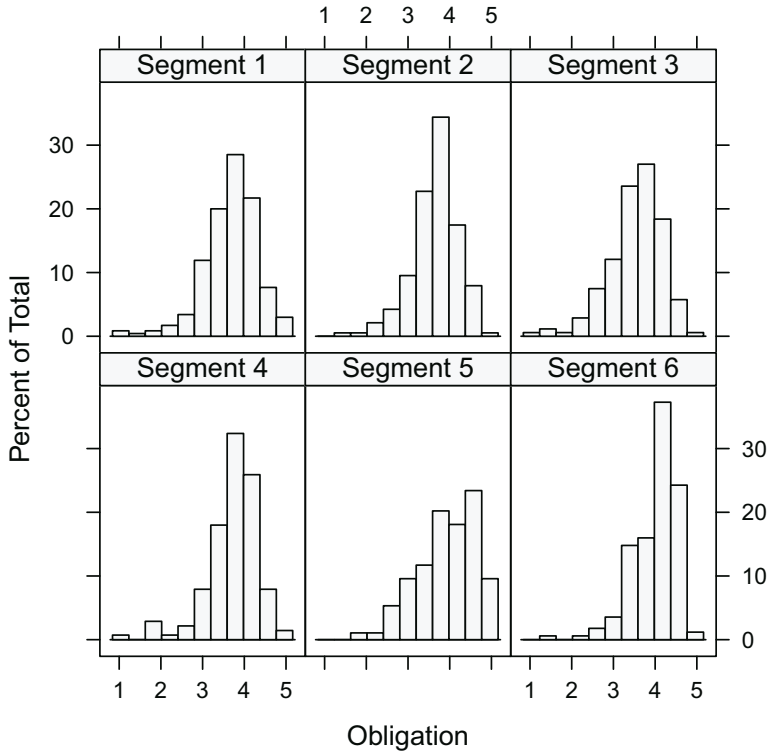
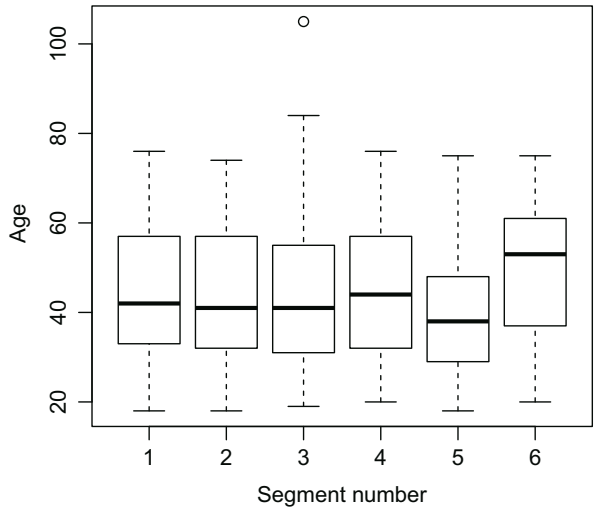


Fig. 9.6 Histograms of moral obligation to protect the environment by segment for the Australian travel motives data set

Fig. 9.7 Parallel box-and-whisker plot of age by segment for the Australian travel motives data set



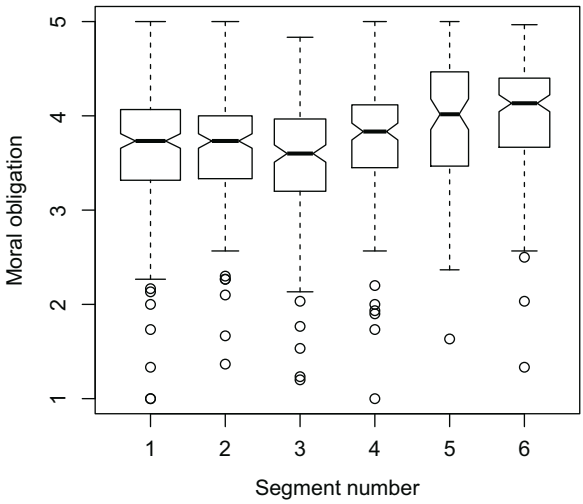


Fig. 9.8 Parallel box-and-whisker plot (with elements of statistical inference) of moral obligation to protect the environment by segment for the Australian travel motives data set

segments are far away from each other. Most of the boxes and whiskers are almost symmetric around the median, but all segments contain some outliers at the low end of moral obligation. One possible interpretation is that – while most respondents state that they feel morally obliged to protect the environment (irrespective of whether they actually do it or not) – only few openly admit to not feeling a sense of moral obligation.

We can use a modified version of the segment level stability across solutions (SLS_A) plot to trace the value of a metric descriptor variable over a series of market segmentation solutions. The modification is that additional information contained in a metric descriptor variable is plotted using different colours for the nodes:

```
R> slsaplot(vacmot.k38, nodecol = vacmotdesc$Obligation)
```

The nodes of the segment level stability across solutions (SLS_A) plot shown in Fig. 9.9 indicate each segment's mean moral obligation to protect the environment using colours. A deep red colour indicates high moral obligation. A light grey colour indicates low moral obligation.

The segment that has been repeatedly identified as a potentially attractive market segment (nature-loving tourists with an interest in the local population) appears along the bottom row. This segment consistently – across all plotted segmentation solutions – displays high moral obligation to protect the environment, followed by the segment identified as containing responses with acquiescence (yes saying) bias (segment 5 in the six-segment solution). This is not altogether surprising: if members of the acquiescence segment have an overall tendency to express agreement with survey questions (irrespective of the content), they are also likely to express agreement when asked about their moral obligation to protect the environment. Because the node colour has a different meaning in this modified segment level stability across solutions (SLS_A) plot, the shading of the edges

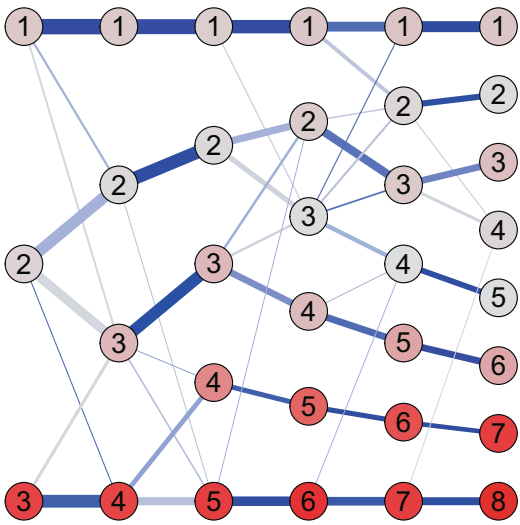


Fig. 9.9 Segment level stability across solutions (SLS_A) plot for the Australian travel motives data set for three to eight segments with nodes coloured by mean moral obligation values

represents the numeric SLS_A value. Light grey edges indicate low stability values. Dark blue edges indicate high stability values.

9.3 Testing for Segment Differences in Descriptor Variables

Simple statistical tests can be used to formally test for differences in descriptor variables across market segments. The simplest way to test for differences is to run a series of independent tests for each variable of interest. The outcome of the segment extraction step is segment membership, the assignment of each consumer to one market segment. Segment membership can be treated like any other nominal variable. It represents a nominal summary statistic of the segmentation variables. Therefore, any test for association between a nominal variable and another variable is suitable.

The association between the nominal segment membership variable and another nominal or ordinal variable (such as gender, level of education, country of origin) is visualised in Sect. 9.2.1 using the cross-tabulation of both variables as basis for the mosaic plot. The appropriate test for independence between columns and rows of a table is the χ^2-test. To formally test for significant differences in the gender distribution across the Australian travel motives segments, we use the following R command:

```
R> chisq.test(C6.Gender)

        Pearson's Chi-squared test

data:  C6.Gender
X-squared = 5.2671, df = 5, p-value = 0.3842
```

The output contains: the name of the statistical test, the data used, the value of the test statistic (in this case `X-squared`), the parameters of the distribution used to calculate the p-value (in this case the degrees of freedom (`df`) of the χ^2-distribution), and the p-value.

The p-value indicates how likely the observed frequencies occur if there is no association between the two variables (and sample size, segment sizes, and overall gender distribution are fixed). Small p-values (typically smaller than 0.05), are taken as statistical evidence of differences in the gender distribution between segments. Here, this test results in a non-significant p-value, implying that the null hypothesis is not rejected. The mosaic plot in Fig. 9.2 confirms this: no effects are visible and no cells are coloured.

The mosaic plot for segment membership and moral obligation to protect the environment shows significant association (Fig. 9.4), as does the corresponding χ^2-test:

```
R> chisq.test(with(vacmotdesc, table(C6, Obligation2)))

        Pearson's Chi-squared test

data:  with(vacmotdesc, table(C6, Obligation2))
X-squared = 96.913, df = 15, p-value = 5.004e-14
```

If the χ^2-test rejects the null hypothesis of independence because the p-value is smaller than 0.05, a mosaic plot is the easiest way of identifying the reason for rejection. The colour of the cells points to combinations occurring more or less frequently than expected under independence.

The association between segment membership and metric variables (such as age, number of nights at the tourist destinations, dollars spent on accommodation) is visualised using parallel boxplots. Any test for difference between the location (mean, median) of multiple market segments can assess if the observed differences in location are statistically significant.

The most popular method for testing for significant differences in the means of more than two groups is *Analysis of Variance* (ANOVA). To test for differences in mean moral obligation values to protect the environment (shown in Fig. 9.8) across market segments, we first inspect segment means:

```
R> C6.moblig <- with(vacmotdesc, tapply(Obligation,
+   C6, mean))
R> C6.moblig

       1        2        3        4        5        6
3.673191 3.651146 3.545977 3.724460 3.928723 4.008876
```

We can use the following analysis of variance to test for significance of differences:

```
R> aov1 <- aov(Obligation ~ C6, data = vacmotdesc)
R> summary(aov1)

             Df Sum Sq Mean Sq F value  Pr(>F)
C6            5   24.7   4.933   12.93 3.3e-12 ***
Residuals   994  379.1   0.381
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of variance performs an F-test with the corresponding test statistic given as `F value`. The `F value` compares the weighted variance between market segment means with the variance within market segments. Small values support the null hypothesis that segment means are the same. The p-value given in the output is smaller than 0.05. This means that we reject the null hypothesis that each segment has the same mean obligation. At least two market segments differ in their mean moral obligation to protect the environment.

Summarising mean values of metric descriptor variables by segment in a table provides a quick overview of segment characteristics. Adding the analysis of variance p-values indicates if differences are statistically significant. As an example, Table 9.1 presents mean values for age and moral obligation by market segment together with the analysis of variance p-values. As a robust alternative we can report median values by segment, and calculate p-values of the Kruskal-Wallis rank sum test. The Kruskal-Wallis rank sum test assumes (as null hypothesis) that all segments have the same median. This test is implemented in function `kruskal.test()` in R. `kruskal.test` is called in the same way as `aov`.

If we reject the null hypothesis of the analysis of variance, we know that segments do not have the same mean level of moral obligation. But the analysis of variance does not identify the differing segments. Pairwise comparisons between segments provide this information. The following command runs all pairwise t-tests, and reports the p-values:

```
R> with(vacmotdesc, pairwise.t.test(Obligation, C6))

    Pairwise comparisons using t tests with pooled SD

data:  Obligation and C6

  1       2       3       4       5
2 1.00000 -       -       -       -
3 0.23820 0.52688 -       -       -
```

Table 9.1 Differences in mean values for age and moral obligation between the six segments for the Australian travel motives data set together with ANOVA p-values

	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Seg. 6	Total	p-value
Age	44.61	42.66	42.31	44.42	39.37	49.62	44.17	1.699E-07
Moral obligation	3.67	3.65	3.55	3.72	3.93	4.01	3.73	3.300E-12

```
4 1.00000 1.00000 0.08980 -       -
5 0.00653 0.00387 1.8e-05 0.09398 -
6 1.2e-06 7.9e-07 1.1e-10 0.00068 1.00000

P value adjustment method: holm
```

The *p*-value of the *t*-test is the same if segment 1 is compared to segment 2, or if segment 2 is compared to segment 1. To avoid redundancy, the output only contains the *p*-values for one of these comparisons, and omits the upper half of the matrix of pairwise comparisons.

The results in the first column indicate that segment 1 does not differ significantly in mean moral obligation from segments 2, 3, and 4, but does differ significantly from segments 5 and 6. The advantage of this output is that it presents the results in very compact form. The disadvantage is that the direction of the difference cannot be seen. A parallel box-and-whisker plot reveals the direction. We see in Fig. 9.8 that segments 5 and 6 feel more morally obliged to protect the environment than segments 1, 2, 3 and 4.

The above R output for the pairwise *t*-tests shows (in the last line) that *p*-values were adjusted for *multiple testing* using the method proposed by Holm (1979). Whenever a series of tests is computed using the same data set to assess a single hypothesis, *p*-values need to be adjusted for multiple testing.

The single hypothesis in this case is that all segment means are the same. This is equivalent to the hypothesis that – for any pair of segments – the means are the same. The series of pairwise *t*-tests assesses the later hypothesis. But the *p*-value of a single *t*-test only controls for wrongly rejecting the null hypothesis that this pair has the same mean values. Adjusting the *p*-values allows to reject the null hypothesis that the means are the same for all segments if at least one of the reported *p*-values is below the significance level. After adjustment, the chance of making a wrong decision meets the expected error rate for testing this hypothesis. If the same rule is applied without adjusting the *p*-values, the error rate of wrongly rejecting the null hypothesis would be too high.

The simplest way to correct *p*-values for multiple testing is Bonferroni correction. Bonferroni correction multiplies all *p*-values by the number of tests computed and, as such, represents a very conservative approach. A less conservative and more accurate approach was proposed by Holm (1979). Several other methods are available, all less conservative than Bonferroni correction. Best known is the false discovery rate procedure proposed by Benjamini and Hochberg (1995). See `help("p.adjust")` for methods available in R.

As an alternative to calculating the series of pairwise *t*-tests, we can plot Tukey's honest significant differences (Tukey 1949; Miller 1981; Yandell 1997):

```
R> plot(TukeyHSD(aov1), las = 1)
R> mtext("Pairs of segments", side = 2, line = 3)
```

Function `mtext()` writes text into the margin of the plot. The first argument (`"Pairs of segments"`) contains the text to be included. The second argument (`"side = 2"`) specifies where the text appears. The value 2 stands for the

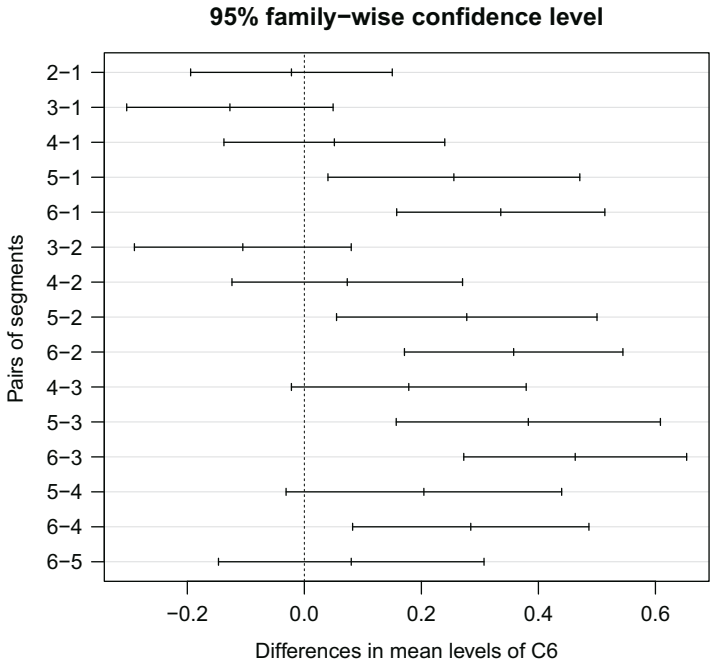


Fig. 9.10 Tukey's honest significant differences of moral obligation to behave environmentally friendly between the six segments for the Australian travel motives data set

left margin. The third argument (`"line = 3"`) specifies the distance between plot and text. The value 3 means the text is written three lines away from the box surrounding the plotting region.

Figure 9.10 shows the resulting plot. Each row represents the comparison of a pair of segments. The first row compares segments 1 and 2, the second row compares segments 1 and 3, and so on. The bottom row compares segments 5 and 6. The point estimate of the differences in mean values is located in the middle of the horizontal solid line. The length of the horizontal solid line depicts the confidence interval of the difference in mean values. The calculation of the confidence intervals is based on the analysis of variance result, and adjusted for the fact that a series of pairwise comparisons is made. If a confidence interval (horizontal solid line in the plot) crosses the vertical line at 0, the difference is not significant. All confidence intervals (horizontal solid lines in the plot) not crossing the vertical line at 0 indicate significant differences.

As can be seen from Fig. 9.10, segments 1, 2, 3 and 4 do not differ significantly from one another in moral obligation. Neither do segments 5 and 6. Segments 5 and 6 are characterised by a significantly higher moral obligation to behave environmentally friendly than the other market segments (with the only exception of segments 4 and 5 not differing significantly). As the parallel box-and-whisker

plot in Fig. 9.8 reveals, segment 4 sits between the low and high group, and does not display significant differences to segments 1–3 at the low end, and 5 at the high end of the moral obligation range.

9.4 Predicting Segments from Descriptor Variables

Another way of learning about market segments is to try to predict segment membership from descriptor variables. To achieve this, we use a *regression model* with the segment membership as categorical dependent variable, and descriptor variables as independent variables. We can use methods developed in statistics for classification, and methods developed in machine learning for supervised learning.

As opposed to the methods in Sect. 9.3, these approaches test differences in all descriptor variables simultaneously. The prediction performance indicates how well members of a market segment can be identified given the descriptor variables. We also learn which descriptor variables are critical to the identification of segment membership, especially if methods are used that simultaneously select variables.

Regression analysis is the basis of prediction models. Regression analysis assumes that a dependent variable y can be predicted using independent variables or regressors $x_1, \ldots, x_p$:

$$y \approx f(x_1, \ldots, x_p).$$

Regression models differ with respect to the function $f(\cdot)$, the distribution assumed for y, and the deviations between y and $f(x_1, \ldots, x_p)$.

The basic regression model is the linear regression model. The linear regression model assumes that function $f(\cdot)$ is linear, and that y follows a normal distribution with mean $f(x_1, \ldots, x_p)$ and variance σ^2. The relationship between the dependent variable y and the independent variables $x_1, \ldots, x_p$ is given by:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

In R, function `lm()` fits a linear regression model. We fit the model for age in dependence of segment membership using:

```
R> lm(Age ~ C6 - 1, data = vacmotdesc)

Call:
lm(formula = Age ~ C6 - 1, data = vacmotdesc)

Coefficients:
 C61   C62   C63   C64   C65   C66
44.6  42.7  42.3  44.4  39.4  49.6
```

In R, regression models are specified using a formula interface. In the formula, the dependent variable AGE is indicated on the left side of the ~. The independent variables are indicated on the right side of the ~. In this particular case, we only use segment membership C6 as independent variable. Segment membership C6 is a categorical variable with six categories, and is coded as a `factor` in the data frame `vacmotdesc`. The formula interface correctly interprets categorical variables, and fits a regression coefficient for each category. For identifiability reasons, either the intercept β_0 or one category needs to be dropped. Using `- 1` on the right hand side of ~ drops the intercept β_0. Without an intercept, each estimated coefficient is equal to the mean age in this segment. The output indicates that members of segment 5 are the youngest with a mean age of 39.4 years, and members of segment 6 are the oldest with a mean age of 49.6 years.

Including the intercept β_0 in the model formula drops the regression coefficient for segment 1. Its effect is instead captured by the intercept. The other regression coefficients indicate the mean age difference between segment 1 and each of the other segments:

```
R> lm(Age ~ C6, data = vacmotdesc)

Call:
lm(formula = Age ~ C6, data = vacmotdesc)

Coefficients:
(Intercept)          C62          C63          C64
     44.609       -1.947       -2.298       -0.191
         C65          C66
     -5.236        5.007
```

The intercept β_0 indicates that respondents in segment 1 are, on average, 44.6 years old. The regression coefficient C66 indicates that respondents in segment 6 are, on average, 5 years older than those in segment 1.

In linear regression models, regression coefficients express how much the dependent variable changes if one independent variable changes while all other independent variables remain constant. The linear regression model assumes that changes caused by changes in one independent variable are independent of the absolute level of all independent variables.

The dependent variable in the linear regression model follows a normal distribution. *Generalised linear models* (Nelder and Wedderburn 1972) can accommodate a wider range of distributions for the dependent variable. This is important if the dependent variable is categorical, and the normal distribution, therefore, is not suitable.

In the linear regression model, the mean value of y given $x_1, \ldots, x_p$ is modelled by the linear function:

$$\mathbb{E}[y|x_1, \ldots, x_p] = \mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Generalised linear models y are not limited to the normal distribution. We could, for example, use the Bernoulli distribution with y taking values 0 or 1. In this case, the mean value of y can only take values in $(0, 1)$. It is therefore not possible to describe the mean value with a linear function which can take any real value. Generalised linear models account for this by introducing a link function $g(\cdot)$. The link function transforms the mean value of y given by μ to an unlimited range indicated by η. This transformed value can then be modelled with a linear function:

$$g(\mu) = \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

η is referred to as linear predictor.

We can use the normal, Poisson, binomial, and multinomial distribution for the dependent variable in generalised linear models. The binomial or multinomial distribution are necessary for classification. A generalised linear model is characterised by the distribution of the dependent variable, and the link function. In the following sections we discuss two special cases of generalised linear models: binary and multinomial logistic regression. In these models the dependent variable follows either a binary or a multinomial distribution, and the link function is the logit function.

9.4.1 Binary Logistic Regression

We can formulate a regression model for binary data using generalised linear models by assuming that $f(y|\mu)$ is the Bernoulli distribution with success probability μ, and by choosing the logit link that maps the success probability $\mu \in (0, 1)$ onto $(-\infty, \infty)$ by

$$g(\mu) = \eta = \log\left(\frac{\mu}{1-\mu}\right).$$

Function `glm()` fits generalised linear models in R. The distribution of the dependent variable and the link function are specified by a `family`. The Bernoulli distribution with logit link is `family = binomial(link = "logit")` or `family = binomial()` because the logit link is the default. The binomial distribution is a generalisation of the Bernoulli distribution if the variable y does not only take values 0 and 1, but represents the number of successes out of a number of independent Bernoulli distributed trials with the same success probability μ.

Here, we fit the model to predict the likelihood of a consumer to belong to segment 3 given their age and moral obligation score. We specify the model using the formula interface with the dependent variable on the left of ~, and the two independent variables AGE and OBLIGATION2 on the right of ~. The dependent variable is a binary indicator of being in segment 3. This binary indicator is constructed with `I(C6 == 3)`. Function `glm()` fits the model given the formula, the data set, and the family:

```
R> f <- I(C6 == 3) ~ Age + Obligation2
R> model.C63 <- glm(f, data = vacmotdesc,
+   family = binomial())
R> model.C63

Call:  glm(formula = f, family = binomial(),
    data = vacmotdesc)

Coefficients:
  (Intercept)            Age  Obligation2Q2  Obligation2Q3
     -0.72197       -0.00842       -0.41900       -0.72285
Obligation2Q4
     -0.92526

Degrees of Freedom: 999 Total (i.e. Null);  995 Residual
Null Deviance:            924
Residual Deviance: 904         AIC: 914
```

The output contains the regression coefficients, and information on the model fit, including the degrees of freedom, the null deviance, the residual deviance, and the AIC.

The intercept in the linear regression model gives the mean value of the dependent variable if the independent variables $x_1, \ldots, x_p$ all have a value of 0. In binomial logistic regression, the intercept gives the value of the linear predictor η if the independent variables $x_1, \ldots, x_p$ all have a value of 0. The probability of being in segment 3 for a respondent with age 0 and a low moral obligation value is calculated by transforming the intercept with the inverse link function, in this case the inverse logit function:

$$g^{-1}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

Transforming the intercept value of -0.72 with the inverse logit link gives a predicted probability of 33% that a consumer of age 0 with low moral obligation is in segment 3.

The other regression coefficients in a linear regression model indicate how much the mean value of the dependent variable changes if this independent variable changes while others remain unchanged. In binary logistic regression, the regression coefficients indicate how the linear predictor changes. The changes in the linear predictor correspond to changes in the log odds of success. The odds of success are

the ratio between the probability of success μ and the probability of failure $1-\mu$. If the odds are equal to 1, success and failure are equally likely. If the odds are larger than 1, success is more likely than failure. Odds are frequently also used in betting.

The coefficient for AGE indicates that the log odds for being in segment 3 are 0.008 lower for tourists who are one year older. This means that the odds of one tourist are $e^{-0.008} = 0.992$ times the odds of another tourist if they only differ by the other tourist being one year younger. The independent variable OBLIGATION2 is a categorical variable with four different levels. The lowest category `Q1` is captured by the intercept. The regression coefficients for this variable indicate the change in log odds between the other categories and the lowest category `Q1`.

To simplify the interpretation of the coefficients and their effects, we can use package effects (Fox 2003; Fox and Hong 2009) in R. Function `allEffects` calculates the predicted values for different levels of the independent variable keeping other independent variables constant at their average value. In the case of the fitted binary logistic regression, the predicted values are the probabilities of being in segment 3. We plot the estimated probabilities to allow for easy inspection:

```
R> library("effects")
R> plot(allEffects(mod = model.C63))
```

Figure 9.11 shows how the predicted probability of being in segment 3 changes with age (on the left), and with moral obligation categories (on the right). The predicted probabilities are shown with pointwise 95% confidence bands (grey shaded areas) for metric independent variables, and with 95% confidence intervals for each category (vertical lines) for categorical independent variables. The predicted probabilities result from transforming the linear predictor with a non-linear function. The changes are not linear, and depend on the values of the other independent variables.

The plot on the left in Fig. 9.11 shows that, for a 20-year old tourist with an average moral obligation score, the predicted probability to be in segment 3 is about 20%. This probability decreases with increasing age. For 100-year old tourists the predicted probability to be in segment 3 is only slightly higher than 10%. The confidence bands indicate that these probabilities are estimated with high uncertainty. The fact that we can place into the plot a horizontal line lying completely within the grey shaded area, indicates that differences in AGE do not significantly affect the probability to be in segment 3. Dropping AGE from the regression model does not significantly decrease model fit.

The plot on the right side of Fig. 9.11 shows that the probability of being a member of segment 3 decreases with increasing moral obligation. Respondents of average age with a moral obligation value of `Q1` have a predicted probability of about 25% to be in segment 3. If these tourists of average age have the highest moral obligation value of `Q4`, they have a predicted probability of 12%. The 95% confidence intervals of the estimated effects indicate that – despite high

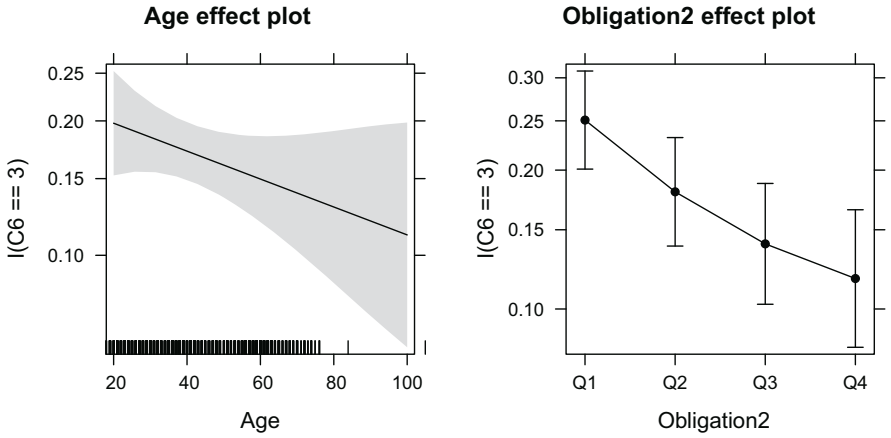


Fig. 9.11 Effect visualisation of age and moral obligation for predicting segment 3 using binary logistic regression for the Australian travel motives data set

uncertainty – probabilities do not overlap for the two most extreme values of moral obligation. This means that including moral obligation in the logistic regression model significantly improves model fit.

Summarising the fitted model provides additional insights:

```
R> summary(model.C63)

Call:
glm(formula = f, family = binomial(), data = vacmotdesc)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.835  -0.653  -0.553  -0.478   2.284

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.72197    0.28203   -2.56  0.01047 *
Age           -0.00842    0.00588   -1.43  0.15189
Obligation2Q2 -0.41900    0.21720   -1.93  0.05372 .
Obligation2Q3 -0.72285    0.23141   -3.12  0.00179 **
Obligation2Q4 -0.92526    0.25199   -3.67  0.00024 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 924.34  on 999  degrees of freedom
Residual deviance: 903.61  on 995  degrees of freedom
AIC: 913.6

Number of Fisher Scoring iterations: 4
```

The output contains the table of the estimated coefficients and their standard errors, the test statistics of a z-test, and the associated p-values. The z-test compares the fitted model to a model where this regression coefficient is set to 0. Rejecting the null hypothesis implies that the regression coefficient is not equal to 0 and this effect should be contained in the model.

This means that the null hypothesis is not rejected for AGE. We can drop AGE from the model without significantly decreasing model fit. If moral obligation is included in the model, AGE does not need to be included.

For moral obligation, three regression coefficients are fitted which capture the difference of categories `Q2`, `Q3` and `Q4` to category `Q1`. Each of the tests only compares the full model with the model with the regression coefficient of a specific category set to 0. This does not allow to decide if the model containing moral obligation performs better than the model without moral obligation. Function `Anova` from package car (Fox and Weisberg 2011) compares the model where moral obligation is dropped, and thus all regression coefficients for this variable are set to 0. We drop each of the independent variables one at a time, and compare the resulting model to the full model:

```
R> library("car")
R> Anova(model.C63)

Analysis of Deviance Table (Type II tests)

Response: I(C6 == 3)
            LR Chisq Df Pr(>Chisq)
Age             2.07  1    0.15024
Obligation2    17.26  3    0.00062 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows – for each independent variable in the model – the test statistic (`LR Chisq`), the degrees of freedom of the distribution to calculate the p-value (`Df`), and the p-value.

The test performed for the metric variable AGE is essentially the same as the z-test included in the `summary` output (use `Anova` with `test.statistic = "Wald"` for the exactly same test). The test indicates that dropping the categorical variable OBLIGATION2 would significantly reduce model fit. Moral obligation is a useful descriptor variable to predict membership in segment 3.

So far we fitted a binary logistic regression including two descriptor variables and simultaneously accounted for their association with the dependent variable. We can add additional independent variables to the binary logistic regression model. We

include all available descriptor variables in a regression model in R by specifying a dot on the right side of the ~. The variables included in the data frame in the `data` argument are then all used as independent variables (if not already used on the left of ~).

```
R> full.model.C63 <- glm(I(C6 == 3) ~ .,
+   data = na.omit(vacmotdesc), family = binomial())
```

Some descriptor variables contain missing values (`NA`). Respondents with at least one missing value are omitted from the data frame using `na.omit(vacmotdesc)`.

Including all available descriptor variables may lead to an overfitting model. An overfitting model has a misleadingly good performance, and overestimates effects of independent variables. Model selection methods exclude irrelevant independent variables. In R, function `step` performs model selection. The `step` function implements a stepwise procedure. In each step, the function evaluates if dropping an independent variable or adding an independent variable improves model fit. Model fit is assessed with the AIC. The AIC balanced goodness-of-fit with a penalty for model complexity. The function then drops or adds the variable leading to the largest improvement in AIC value. This procedure continues until no improvement in AIC is achieved by dropping or adding one independent variable.

```
R> step.model.C63 <- step(full.model.C63, trace = 0)
R> summary(step.model.C63)

Call:
glm(formula = I(C6 == 3) ~ Education + NEP +
    Vacation.Behaviour, family = binomial(),
    data = na.omit(vacmotdesc))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.051  -0.662  -0.545  -0.425   2.357

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.9359     0.6783    1.38  0.16762
Education           0.0571     0.0390    1.47  0.14258
NEP                -0.3139     0.1658   -1.89  0.05838 .
Vacation.Behaviour -0.5767     0.1504   -3.83  0.00013 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 802.23  on 867  degrees of freedom
Residual deviance: 773.19  on 864  degrees of freedom
AIC: 781.2

Number of Fisher Scoring iterations: 4
```

We suppress the printing of progress information of the iterative fitting function on screen using `trace = 0`. The selected final model is summarised. The model includes three variables: EDUCATION, NEP, and VACATION.BEHAVIOUR.

We compare the predictive performance of the model including AGE and MORAL.OBLIGATION with the model selected using `step`. A well predicting model would assign a high probability of being in segment 3 to members of segment 3 and a low probability to all other consumers. Function `predict()` returns the predicted probabilities of being in segment 3 for all consumers if the function is applied to a fitted model, and we specify `type = "response"`. Parallel boxplots visualise the distributions of predicted probabilities for consumers in segment 3, and those not in segment 3:

```
R> par(mfrow = c(1, 2))
R> prob.C63 <- predict(model.C63, type = "response")
R> boxplot(prob.C63 ~ I(C6 == 3), data = vacmotdesc,
+    ylim = 0:1, main = "", ylab = "Predicted probability")
R> prob.step.C63 <- predict(step.model.C63, type = "response")
R> boxplot(prob.step.C63 ~ I(C6 == 3),
+    data = na.omit(vacmotdesc), ylim = 0:1,
+    main = "", ylab = "Predicted probability")
```

Figure 9.12 compares the predicted probabilities of segment 3 membership for the two models. If the fitted model differentiates well between members of segment 3 and all other consumers, the boxes are located at the top of the plot (close to the value of 1) for respondents in segment 3 (TRUE), and at the bottom (close to the

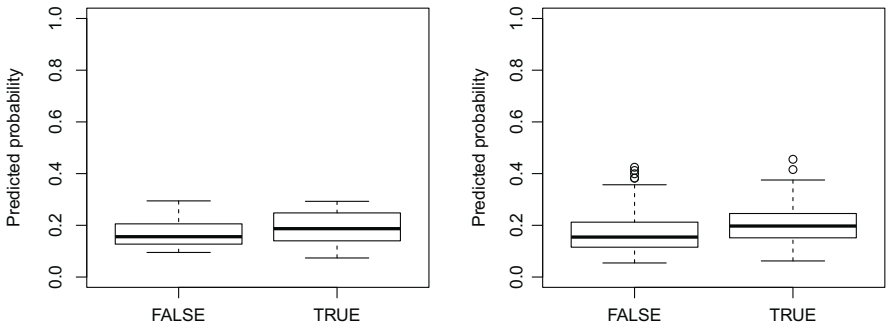


Fig. 9.12 Predicted probabilities of segment 3 membership for consumers not assigned to segment 3 (FALSE) and for consumers assigned to segment 3 (TRUE) for the Australian travel motives data set. The model containing age and moral obligation as independent variables is on the left; the model selected using stepwise variable selection on the right

value of 0) for all other consumers. We can see from Fig. 9.12 that the performance of the two fitted models is nowhere close to this optimal case. The median predicted values are only slightly higher for segment 3 in both models. The difference is larger for the model fitted using `step`, indicating that the predictive performance of this model is slightly better.

9.4.2 Multinomial Logistic Regression

Multinomial logistic regression can fit a model that predicts each segment simultaneously. Because segment extraction typically results in more than two market segments, the dependent variable y is not binary. Rather, it is categorical and assumed to follow a multinomial distribution with the logistic function as link function.

In R, function `multinom()` from package nnet (Venables and Ripley 2002) (instead of `glm`) fits a multinomial logistic regression. We specify the model in a similar way using a formula and a data frame for evaluating the formula.

```
R> library("nnet")
R> vacmotdesc$Oblig2 <- vacmotdesc$Obligation2
R> model.C6 <- multinom(C6 ~ Age + Oblig2,
+   data = vacmotdesc, trace = 0)
```

Using `trace = 0` avoids the display of progress information of the iterative fitting function.

The fitted model contains regression coefficients for each segment except for segment 1 (the baseline category). The same set of regression coefficients would result from a binary logistic regression model comparing this segment to segment 1. The coefficients indicate the change in log odds if the independent variable changes:

```
R> model.C6

Call:
multinom(formula = C6 ~ Age + Oblig2, data = vacmotdesc,
    trace = 0)

Coefficients:
  (Intercept)     Age Oblig2Q2 Oblig2Q3 Oblig2Q4
2       0.184 -0.0092    0.108   -0.026    -0.16
3       0.417 -0.0103   -0.307   -0.541    -0.34
4      -0.734 -0.0017    0.309    0.412     0.42
5      -0.043 -0.0296   -0.023   -0.039     1.33
6      -2.090  0.0212    0.269    0.790     1.65

Residual Deviance: 3384
AIC: 3434
```

The regression coefficients are arranged in matrix form. Each row contains the regression coefficients for one category of the dependent variable. Each column contains the regression coefficients for one effect of an independent variable.

The `summary()` function returns the regression coefficients and their standard errors.

```
R> summary(model.C6)

Call:
multinom(formula = C6 ~ Age + Oblig2, data = vacmotdesc,
    trace = 0)

Coefficients:
  (Intercept)     Age Oblig2Q2 Oblig2Q3 Oblig2Q4
2       0.184 -0.0092    0.108   -0.026    -0.16
3       0.417 -0.0103   -0.307   -0.541    -0.34
4      -0.734 -0.0017    0.309    0.412     0.42
5      -0.043 -0.0296   -0.023   -0.039     1.33
6      -2.090  0.0212    0.269    0.790     1.65

Std. Errors:
  (Intercept)    Age Oblig2Q2 Oblig2Q3 Oblig2Q4
2        0.34 0.0068     0.26     0.26     0.31
3        0.34 0.0070     0.26     0.27     0.31
4        0.39 0.0075     0.30     0.30     0.34
5        0.44 0.0091     0.37     0.38     0.35
6        0.42 0.0073     0.34     0.32     0.32

Residual Deviance: 3384
AIC: 3434
```

With function `Anova()` we assess if dropping a single variable significantly reduces model fit. Dropping a variable corresponds to setting all regression coefficients of this variable to 0. This means that the regression coefficients in one or several columns of the regression coefficient matrix corresponding to this variable are set to 0. Function `Anova()` tests if dropping any of the variables significantly reduces model fit. The output is essentially the same as for the binary logistic regression model:

```
R> Anova(model.C6)

Analysis of Deviance Table (Type II tests)

Response: C6
       LR Chisq Df Pr(>Chisq)
Age        35.6  5    1.1e-06 ***
Oblig2     89.0 15    1.5e-12 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output indicates that dropping any of the variables leads to a significant reduction in model fit. Applying function `step()` to a fitted model performs model selection. Starting with the full model containing all available independent variables,

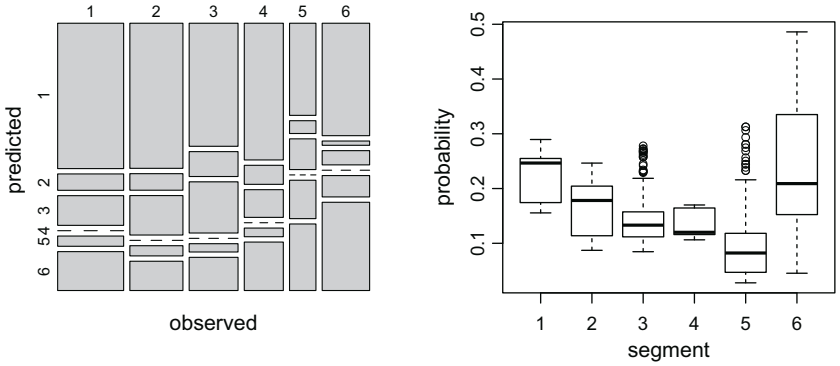


Fig. 9.13 Assessment of predictive performance of the multinomial logistic regression model including age and moral obligation as independent variables for the Australian travel motives data set. The mosaic plot of the cross-tabulation of observed and predicted segment memberships is on the left. The parallel boxplot of the predicted probabilities by segment for consumers assigned to segment 6 is on the right

the stepwise procedure returns the best-fitting model, the model which deteriorates in AIC if an independent variable is either dropped or additionally included.

We assess the predictive performance of the fitted model by comparing the predicted segment membership to the observed segment membership. Figure 9.13 shows a mosaic plot of the predicted and observed segment memberships on the left. In addition, we investigate the distribution of the predicted probabilities for each segment. Figure 9.13 shows parallel boxplots of the predicted segment probabilities for consumers assigned to segment 6 on the right:

```
R> par(mfrow = c(1, 2))
R> pred.class.C6 <- predict(model.C6)
R> plot(table(observed = vacmotdesc$C6,
+   predicted = pred.class.C6), main = "")
R> pred.prob.C6 <- predict(model.C6, type = "prob")
R> predicted <- data.frame(prob = as.vector(pred.prob.C6),
+   observed = C6,
+   predicted = rep(1:6, each = length(C6)))
R> boxplot(prob ~ predicted,
+   xlab = "segment", ylab = "probability",
+   data = subset(predicted, observed == 6))
```

By default `predict` returns the predicted classes. Adding the argument `type = "prob"` returns the predicted probabilities.

The left panel of Fig. 9.13 shows that none of the consumers are predicted to be in segment 4. Most respondents are predicted to belong to segment 1, the largest segment. The detailed results for segment 6 (right panel of Fig. 9.13) indicate that consumers from this segment have particularly low predicted probabilities to belong to segment 5.

To ease interpretation of the estimated effects, we use function `allEffects`, and plot the predicted probabilities:

```
R> plot(allEffects(mod = model.C6), layout = c(3, 2))
```

The left panel in Fig. 9.14 shows how the predicted probability to belong to any segment changes with age for a consumer with average moral obligation. The predicted probability for each segment is visualised separately. The heading indicates the segments. For example, `C6 = 1` indicates that the panel contains predicted probabilities for segment 1. Shaded grey areas indicate pointwise 95% confidence bands visualising the uncertainty of the estimated probabilities.

The predicted probability to belong to segment 6 increases with age: young respondents belong to segment 6 with a probability of less than 10%. Older respondents have a probability of about 40%. The probability of belonging to segment 5 decreases with age.

The right panel in Fig. 9.14 shows how the predicted segment membership probability changes with moral obligation values for a consumer of average age. The predicted probability to belong to segment 6 increases with increasing moral obligation value. Respondents with the lowest moral obligation value of `Q1` have a probability of about 8% to be from segment 6. This increases to 29% for respondents with a moral obligation value of `Q4`. For segment 3 the reverse is true: respondents with higher moral obligation values have lower probabilities to be from segment 3.

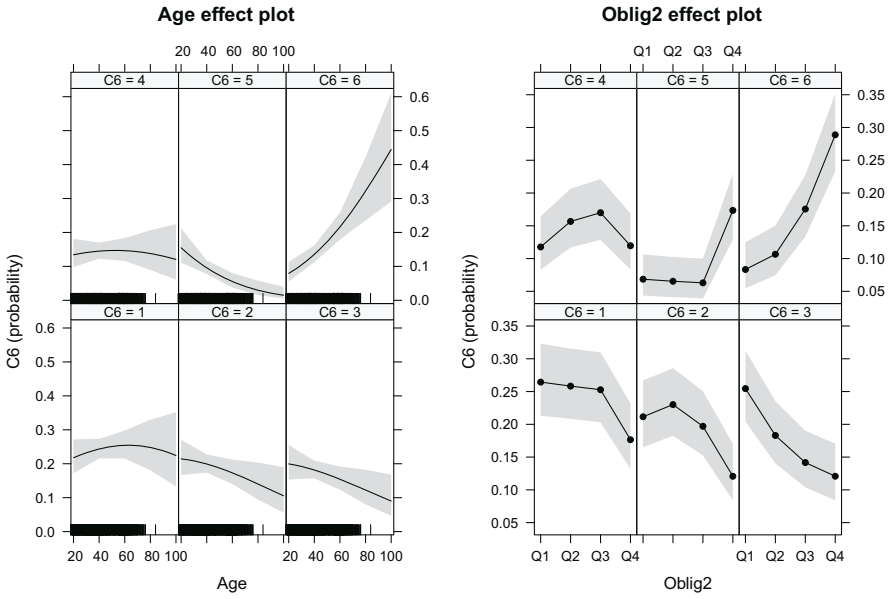


Fig. 9.14 Effect visualisation of age and moral obligation for predicting segment membership using multinomial logistic regression for the Australian travel motives data set

9.4.3 Tree-Based Methods

Classification and regression trees (CARTs; Breiman et al. 1984) are an alternative modelling approach for predicting a binary or categorical dependent variable given a set of independent variables. Classification and regression trees are a supervised learning technique from machine learning. The advantages of classification and regression trees are their ability to perform variable selection, ease of interpretation supported by visualisations, and the straight-forward incorporation of interaction effects. Classification and regression trees work well with a large number of independent variables. The disadvantage is that results are frequently unstable. Small changes in the data can lead to completely different trees.

The tree approach uses a stepwise procedure to fit the model. At each step, consumers are split into groups based on one independent variable. The aim of the split is for the resulting groups to be as pure as possible with respect to the dependent variable. This means that consumers in the resulting groups have similar values for the dependent variable. In the best case, all group members have the same value for a categorical dependent variable. Because of this stepwise splitting procedure, the classification and regression tree approach is also referred to as *recursive partitioning*.

The resulting tree (see Figs. 9.15, 9.16, and 9.17) shows the nodes that emerge from each splitting step. The node containing all consumers is the *root node*. Nodes that are not split further are *terminal nodes*. We predict segment membership by moving down the tree. At each node, we move down the branch reflecting the consumer's independent variable. When we reach the terminal node, segment membership can be predicted based on the segment memberships of consumers contained in the terminal node.

Tree constructing algorithms differ with respect to:

- Splits into two or more groups at each node (binary vs. multi-way splits)
- Selection criterion for the independent variable for the next split
- Selection criterion for the split point of the independent variable
- Stopping criterion for the stepwise procedure
- Final prediction at the terminal node

Several R packages implement tree constructing algorithms. Package rpart (Therneau et al. 2017) implements the algorithm proposed by Breiman et al. (1984). Package partykit (Hothorn and Zeileis 2015) implements an alternative tree constructing procedure that performs unbiased variable selection. This means that the procedure selects independent variables on the basis of association tests and their p-values (see Hothorn et al. 2006). Package partykit also enables visualisation of the fitted tree models.

Function `ctree()` from package partykit fits a conditional inference tree. As an example, we use the Australian travel motives data set with the six-segment solution extracted using neural gas clustering in Sect. 7.5.4. We use membership

in segment 3 as a binary dependent variable, and include all available descriptor variables as independent variables:

```
R> set.seed(1234)
R> library("partykit")
R> tree63 <- ctree(factor(C6 == 3) ~ .,
+   data = vacmotdesc)
R> tree63

Model formula:
factor(C6 == 3) ~ Gender + Age + Education +
    Income + Income2 + Occupation + State +
    Relationship.Status + Obligation + Obligation2 +
    NEP + Vacation.Behaviour + Oblig2

Fitted party:
[1] root
|   [2] Vacation.Behaviour <= 2.2: FALSE (n = 130,
     err = 32%)
|   [3] Vacation.Behaviour > 2.2
|   |   [4] Obligation <= 3.9: FALSE (n = 490, err = 19%)
|   |   [5] Obligation > 3.9: FALSE (n = 380, err = 11%)

Number of inner nodes:    2
Number of terminal nodes: 3
```

The output describes the fitted classification tree shown in Fig. 9.15. The classification tree starts with a root node containing all consumers. Next, the root note is split into two nodes (numbered 2 and 3) using the independent variable VACATION.BEHAVIOUR. The split point is 2.2. This means that consumers with a VACATION.BEHAVIOUR score of 2.2 or less are assigned to node 2. Consumers with a score higher than 2.2 are assigned to node 3. Node 2 is not split further; it becomes a terminal node. The predicted value for this particular terminal node is `FALSE`. The number of consumers in this terminal node is shown in brackets (`n = 130`), along with the proportion of wrongly classified respondents (`err = 32%`). Two thirds of consumers in this node are not in segment 3, one third is. Node 3 is split into two nodes (numbered 4 and 5) using the independent variable OBLIGATION. Consumers with an OBLIGATION score of 3.9 or less are assigned to node 4. Consumers with a higher score are assigned to node 5. The tree predicts that respondents in node 4 are not in segment 3. Node 4 contains 490 respondents; 81% of them are not in segment 3, 19% are. Most respondents in node 5 are also not in segment 3. Node 5 contains 380 respondents; 11% of them are in segment 3. The output also shows that there are 2 inner nodes (numbered 1 and 3), and 3 terminal nodes (numbered 2, 4, and 5).

Plotting the classification tree using `plot(tree63)` gives a visual representation that is easier to interpret. Figure 9.15 visualises the classification tree. The root node on the top has the number 1. The root node contains the name of the variable used for the first split (VACATION.BEHAVIOUR), as well as the *p*-value of the association test that led to the selection of this particular variable (`p <`

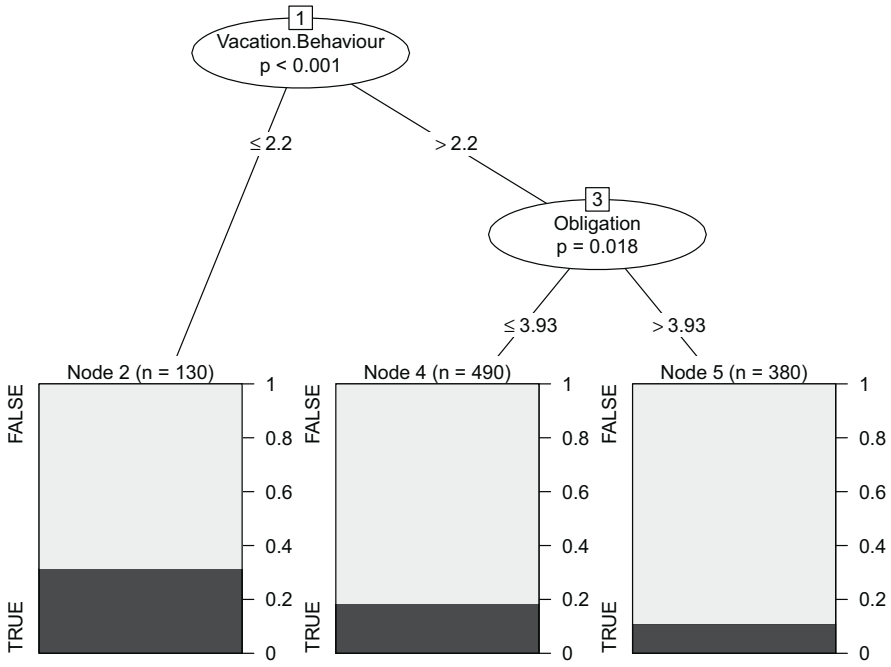


Fig. 9.15 Conditional inference tree using membership in segment 3 as dependent variable for the Australian travel motives data set

`0.001`). The lines underneath the node indicate the split or threshold value of the independent variable VACATION.BEHAVIOUR where respondents are directed to the left or right branch. Consumers with a value higher than 2.2 follow the right branch to node 3. Consumers with a value of 2.2 or less follow the left branch to node 2. These consumers are not split up further; node 2 is a terminal node. The proportion of respondents in node 2 who belong to segment 3 is shown at the bottom of the stacked bar chart for node 2. The dark grey area represents this proportion, and the label on the *y*-axis indicates that this is for the category `TRUE`. The proportion of consumers in node 2 not belonging to segment 3 is shown in light grey with label `FALSE`.

Node 3 is split further using OBLIGATION as the independent variable. The split value is 3.9. Using this split value, consumers are assigned to either node 4 or node 5. Both are terminal nodes. Stacked barplots visualise the proportion of respondents belonging to segment 3 for nodes 4 and 5.

This tree plot indicates that the group with a low mean score for environmentally friendly behaviour on vacation contains the highest proportion of segment 3 members. The group with a high score for environmental friendly behaviour and moral obligation, contains the smallest proportion of segment 3 members. The dark grey area is largest for node 1 and lowest for node 5.

Package partykit takes a number of parameters for the algorithm set by the `control` argument with function `ctree_control`. These parameters influence the tree construction by restricting nodes considered for splitting, by specifying the minimum size for terminal nodes, by selecting the test statistic for the association test, and by setting the minimum value of the criterion of the test to implement a split.

As an illustration, we fit a tree with segment 6 membership as dependent variable. We ensure that terminal nodes contain at least 100 respondents (`minbucket = 100`), and that the minimum criterion value (`mincriterion`) is 0.99 (corresponding to a p-value of smaller than 0.01). Figure 9.16 visualises this tree.

```
R> tree66 <- ctree(factor(C6 == 6) ~ .,
+   data = vacmotdesc,
+   control = ctree_control(minbucket = 100,
+     mincriterion = 0.99))
R> plot(tree66)
```

The fitted classification tree for segment 6 is more complex than that for segment 3; the number of inner and terminal nodes is larger. The stacked bar charts for the terminal nodes indicate how pure the terminal nodes are, and how the terminal nodes differ in the proportion of segment 6 members they contain. The tree algorithm tries to maximise these differences. Terminal node 11 (on the right) contains the highest proportion of consumers assigned to segment 6. Node 11 contains respondents with the highest possible value for moral obligation, and a NEP score of at least 4.

We can also fit a tree for categorical dependent variables with more than two categories with function `ctree()`. Here, the dependent variable in the formula on the left is a categorical variable. `C6` is a factor containing six levels; each level indicates the segment membership of respondents.

```
R> tree6 <- ctree(C6 ~ ., data = vacmotdesc)
R> tree6

Model formula:
C6 ~ Gender + Age + Education + Income +
   Income2 + Occupation + State + Relationship.Status +
   Obligation + Obligation2 + NEP + Vacation.Behaviour +
   Oblig2

Fitted party:
[1] root
|   [2] Oblig2 in Q1, Q2, Q3
|   |   [3] Education <= 6: 1 (n = 481, err = 73%)
|   |   [4] Education > 6: 1 (n = 286, err = 77%)
|   [5] Oblig2 in Q4
|   |   [6] Obligation <= 4.7: 6 (n = 203, err = 67%)
|   |   [7] Obligation > 4.7: 5 (n = 30, err = 57%)

Number of inner nodes:    3
Number of terminal nodes: 4
```

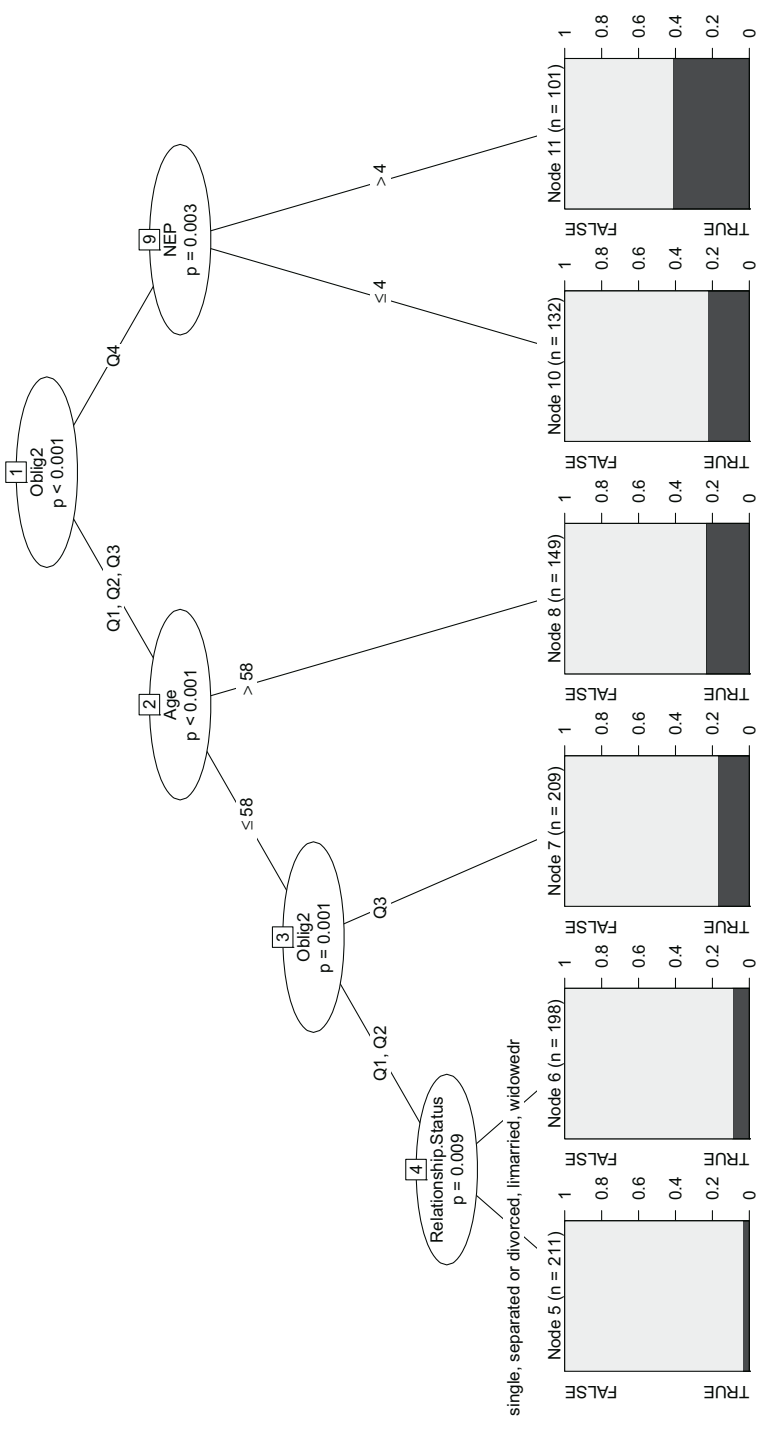


Fig. 9.16 Conditional inference tree using membership in segment 6 as dependent variable for the Australian travel motives data set

The output shows that the first splitting variable is the categorical variable indicating moral obligation (OBLIGATION2). This variable splits the root node 1 into nodes 2 and 5. Consumers with a moral obligation value of `Q1`, `Q2` and `Q3` are assigned to node 2. Consumers with a moral obligation value of `Q4` are assigned to node 7.

Node 2 is split into nodes 3 and 4 using EDUCATION as splitting variable. Consumers with an EDUCATION level of 6 or less are assigned to node 3. Node 3 is a terminal node. Most consumers in this terminal node belong to segment 1. Node 3 contains 481 respondents. Predicting segment membership as 1 for consumers in this node is wrong in 73% of cases.

Respondents with an EDUCATION level higher than 6 are assigned to node 4. Node 4 is a terminal node. The predicted segment membership for node 4 is 1. This node contains 286 respondents and 77% of them are not in segment 1.

Consumers in node 5 feel highly morally obliged to protect the environment. They are split into nodes 6 and 7 using the metric version of moral obligation as splitting variable. Node 6 contains respondents with a moral obligation value of 47 or less, and a moral obligation category value of `Q4`. Most respondents in node 6 belong to segment 6. The node contains 203 respondents; 67% are not from segment 6. Consumers with a moral obligation score higher than 4.7 are in node 7. The predicted segment membership for this node is 5. The node contains 30 consumers; 57% do not belong to segment 5.

Figure 9.17 visualises the tree. `plot(tree6)` creates this plot. Most of the plot is the same as for the classification tree with the binary dependent variable. Only the bar charts at the bottom look different. The terminal nodes show the proportion of respondents in each segment. Optimally, these bar charts for each terminal node show that nearly all consumers in that node have the same segment membership or are at least assigned to only a small number of different segments. Node 7 in Fig. 9.17 is a good example: it contains high proportions of members of segments 1 and 5, but only low proportions of members of other segments.

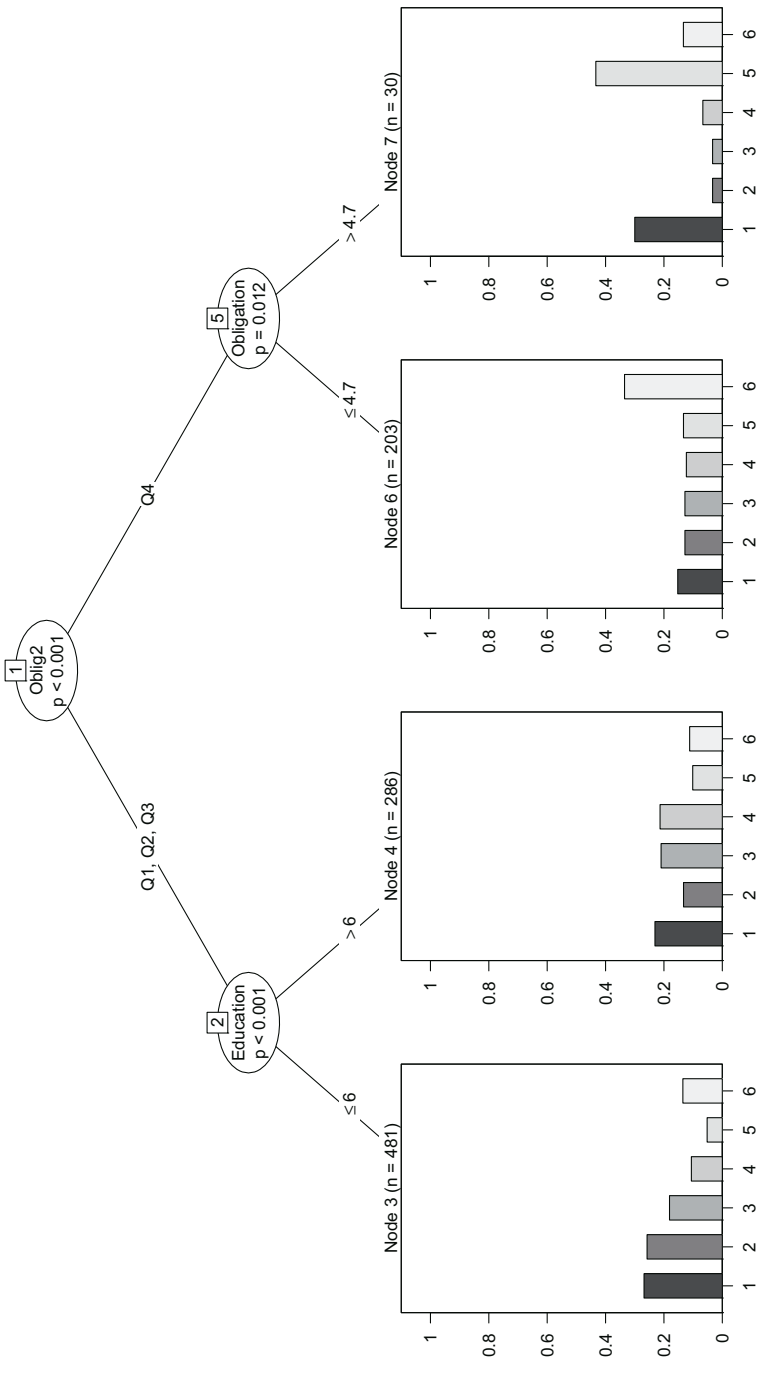


Fig. 9.17 Conditional inference tree using segment membership as dependent variable for the Australian travel motives data set

9.5 Step 7 Checklist

Task	Who is responsible?	Completed?
Bring across from Step 6 (profiling) one or a small number of market segmentation solutions selected on the basis of attractive profiles.		☐
Select descriptor variables. Descriptor variables are additional pieces of information about each consumer included in the market segmentation analysis. Descriptor variables have not been used to extract the market segments.		☐
Use visualisation techniques to gain insight into the differences between market segments with respect to descriptor variables. Make sure you use appropriate plots, for example, mosaic plots for categorical and ordinal descriptor variables, and box-and-whisker plots for metric descriptor variables.		☐
Test for statistical significance of descriptor variables.		☐
If you used separate statistical tests for each descriptor variable, correct for multiple testing to avoid overestimating significance.		☐
"Introduce" each market segment to the other team members to check how much you know about these market segments.		☐
Ask if additional insight into some segments is required to develop a full picture of them.		☐

References

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Chapman and Hall/CRC, New York

Cornelius B, Wagner U, Natter M (2010) Managerial applicability of graphical formats to support positioning decisions. Journal für Betriebswirtschaft 60(3):167–201

Dolnicar S, Leisch F (2008) An investigation of tourists' patterns of obligation to protect the environment. J Travel Res 46:381–391

Fox J (2003) Effect displays in R for generalised linear models. J Stat Softw 8(15):1–27

Fox J, Hong J (2009) Effect displays in R for multinomial and proportional-odds logit models: extensions to the effects package. J Stat Softw 32(1):1–24

Fox J, Weisberg S (2011) An R Companion to applied regression, 2nd edn. Sage, Thousand Oaks. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Friendly M (1994) Mosaic displays for multi-way contingency tables. J Amer Stat Assoc 89: 190–200

Hartigan JA, Kleiner B (1984) A mosaic of television ratings. Amer Statist 38:32–35
Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70
Hothorn T, Zeileis A (2015) partykit: a modular toolkit for recursive partytioning in R. J Mach Learn Res 16:3905–3909. http://jmlr.org/papers/v16/hothorn15a.html
Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 15(3):651–674
Miller RG (1981) Simultaneous statistical inference. Springer, Heidelberg
Nelder J, Wedderburn R (1972) Generalized linear models. J R Stat Soc A 135(3):370–384
van Raaij WF, Verhallen TMM (1994) Domain-specific market segmentation. Eur J Market 28(10):49–66
Sarkar D (2008) lattice: multivariate data visualization with R. Springer, New York
Therneau T, Atkinson B, Ripley B (2017) rpart: recursive partitioning and regression trees. https://CRAN.R-project.org/package=rpart, R package version 4.1–11
Tukey J (1949) Comparing individual means in the analysis of variance. Biometrics 5(2):99–114
Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York
Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer, New York. http://ggplot2.org
Yandell BS (1997) Practical data analysis for designed experiments. Chapman & Hall, New York

Target Market Selection and its Strategies

10.1 The Targeting Decision

Step 8 is where the rubber hits the road. Now the big decision is made: which of the many possible market segments will be selected for targeting? Market segmentation is a strategic marketing tool. The selection of one or more target segments is a long-term decision significantly affecting the future performance of an organisation. This is when the flirting and dating is over; it's time to buy a ring, pop the question, and commit.

After a *global* market segmentation solution has been chosen – typically at the end of Step 5 – a number of segments are available for detailed inspection. These segments are profiled in Step 6, and described in Step 7. In Step 8, one or more of those market segments need to be selected for targeting. The segmentation team can build on the outcome of Step 2. During Step 2, knock-out criteria for market segments have been agreed upon, and segment attractiveness criteria have been selected, and weighed to reflect the relative importance of each of the criteria to the organisation.

Optimally, the knock-out criteria have already been applied in previous steps. For example, in Step 6 market segments were profiled by inspecting their key characteristics in terms of the segmentation variables. It would have become obvious in Step 6 if a market segment is not large enough, not homogeneous or not distinct enough. It would have become obvious in Step 7 – in the process of detailed segment description using descriptor variables – if a market segment is not identifiable or reachable. And in both Steps 6 and 7, it would have become clear if a market segment has needs the organisation cannot satisfy. Imagine, for example, that the BIG SPENDING CITY TOURIST emerged as one of the very distinct and attractive segments from a market segmentation analysis, but the destination conducting the analysis is a nature based destination in outback Australia. The chances of this destination meeting the needs of the highly attractive segment of BIG SPENDING CITY TOURIST are rather slim. Optimally, therefore, all the market segments

under consideration in Step 8 should already comply with the knock-out criteria. Nevertheless, it does not hurt to double check. The first task in Step 8, therefore, is to ensure that all the market segments that are still under consideration to be selected as target markets have well and truly passed the knock-out criteria test.

Once this is done, the attractiveness of the remaining segments and the relative organisational competitiveness for these segments needs to be evaluated. In other words, the segmentation team has to ask a number of questions which fall into two broad categories:

1. Which of the market segments would the organisation most like to target? Which segment would the organisation like to commit to?
2. Which of the organisations offering the same product would each of the segments most like to buy from? How likely is it that our organisation would be chosen? How likely is it that each segment would commit to us?

Answering these two questions forms the basis of the target segment decision.

10.2 Market Segment Evaluation

Most books that discuss target market selection (e.g., McDonald and Dunbar 1995; Lilien and Rangaswamy 2003), recommend the use of a *decision matrix* to visualise relative segment attractiveness and relative organisational competitiveness for each market segment. Many versions of decision matrices have been proposed in the past, and many names are used to describe them, including: *Boston matrix* (McDonald and Dunbar 1995; Dibb and Simkin 2008) because this type of matrix was first proposed by the Boston Consulting Group; *General Electric / McKinsey matrix* (McDonald and Dunbar 1995) because this extended version of the matrix was developed jointly by General Electric and McKinsey; *directional policy matrix* (McDonald and Dunbar 1995; Dibb and Simkin 2008); *McDonald four-box directional policy matrix* (McDonald and Dunbar 1995); and *market attractiveness-business strength matrix* (Dibb and Simkin 2008). The aim of all these decision matrices along with their visualisations is to make it easier for the organisation to evaluate alternative market segments, and select one or a small number for targeting. It is up to the market segmentation team to decide which variation of the decision matrix offers the most useful framework to assist with decision making.

Whichever variation is chosen, the two criteria plotted along the axes cover two dimensions: segment attractiveness, and relative organisational competitiveness specific to each of the segments. Using the analogy of finding a partner for life: segment attractiveness is like the question Would you like to marry this person? given all the other people in the world you could marry. Relative organisational competitiveness is like the question Would this person marry you? given all the other people in the world they could marry.

In the following example, we use a generic segment evaluation plot that can easily be produced in R. To keep segment evaluation as intuitive as possible, we

label the two axes *How attractive is the segment to us?* and *How attractive are we to the segment?* We plot segment attractiveness along the x-axis, and relative organisational competitiveness along the y-axis. Segments appear as circles. The size of the circles reflects another criterion of choice that is relevant to segment selection, such as contribution to turnover or loyalty.

Of course, there is no single best measure of segment attractiveness or relative organisational competitiveness. It is therefore necessary for users to return to their specifications of what an ideal target segment looks like for them. The ideal target segment was specified in Step 2 of the market segmentation analysis. Step 2 resulted in a number of criteria of segment attractiveness, and weights quantifying how much impact each of these criteria has on the total value of segment attractiveness.

In Step 8, the target segment selection step of market segmentation analysis, this information is critical. However, the piece of information missing to be able to select a target segment, is the actual value each market segment has for each of the criteria specified to constitute segment attractiveness. These values emerge from the grouping, profiling, and description of each market segment. To determine the attractiveness value to be used in the segment evaluation plot for each segment, the segmentation team needs to assign a value for each attractiveness criterion to each segment.

The location of each market segment in the segment evaluation plot is then computed by multiplying the weight of the segment attractiveness criterion (agreed upon in Step 2) with the value of the segment attractiveness criterion for each market segment. The value of the segment attractiveness criterion for each market segment is determined by the market segmentation team based on the profiles and descriptions resulting from Steps 6 and 7. The result is a weighted value for each segment attractiveness criterion for each segment. Those values are added up, and represent a segment's overall attractiveness (plotted along the x-axis). Table 10.1 contains an example of this calculation. In this case, the organisation has chosen five segment attractiveness criteria, and has assigned importance weights to them (shown in the second column). Then, based on the profiles and descriptions of each market segment, each segment is given a rating from 1 to 10 with 1 representing the worst and 10 representing the best value. Next, for each segment, the rating is multiplied with the weight, and all weighted attractiveness values are added. Looking at segment 1, for example, determining the segment attractiveness value leads to the following calculation (where 0.25 stands for 25%): $0.25 \cdot 5 + 0.35 \cdot 2 + 0.20 \cdot 10 + 0.10 \cdot 8 + 0.10 \cdot 9 = 5.65$. The value of 5.65 is therefore the x-axis location of segment 1 in the segment evaluation plot shown in Fig. 10.1.

The exact same procedure is followed for the relative organisational competitiveness. The question asked when selecting the criteria is: *Which criteria do consumers use to select between alternative offers in the market?* Possible criteria may include attractiveness of the product to the segment in view of the benefits segment members seek; suitability of the current price to segment willingness or ability to pay; availability of distribution channels to get the product to the segment; segment awareness of the existence of the organisation or brand image of the organisation held by segment members.

Table 10.1 Data underlying the segment evaluation plot

	Weight	Seg 1	Seg 2	Seg 3	Seg 4	Seg 5	Seg 6	Seg 7	Seg 8
How attractive is the segment to us? (segment attractiveness)									
Criterion 1	25%	5	10	1	5	10	3	1	10
Criterion 2	35%	2	1	2	6	9	4	2	10
Criterion 3	20%	10	6	4	4	8	2	1	9
Criterion 4	10%	8	4	2	7	10	8	3	10
Criterion 5	10%	9	6	1	4	7	9	7	8
Total	100%	5.65	5.05	2.05	5.25	8.95	4.25	2.15	9.6
How attractive are we to the segment? (relative organisational competitiveness)									
Criterion 1	25%	2	10	10	10	1	5	2	9
Criterion 2	25%	3	10	4	6	2	4	3	8
Criterion 3	25%	4	10	8	7	3	3	1	10
Criterion 4	15%	9	8	3	9	4	5	3	9
Criterion 5	10%	1	8	6	2	1	4	4	8
Total	100%	3.7	9.5	6.55	7.3	2.2	4.15	2.35	8.9
Size		2.25	5.25	6.00	3.75	5.25	2.25	4.50	1.50

The value of each segment on the axis labelled *How attractive are we to the segment?* is calculated in the same way as the value for the attractiveness of each segment from the organisational perspective: first, criteria are agreed upon, next they are weighted, then each segment is rated, and finally the values are multiplied and summed up. The data underlying the segment evaluation plot based on the hypothetical example in Fig. 10.1 are given in Table 10.1.

The last aspect of the plot is the bubble size (contained in row "Size" in Table 10.1). Anything can be plotted onto the bubble size. Typically profit potential is plotted. Profit combines information about the size of the segment with spending and, as such, represents a critical value when target segments are selected. In other contexts, entirely different criteria may matter. For example, if a non for profit organisation uses market segmentation to recruit volunteers to help with land regeneration activities, they may choose to plot the number of hours volunteered as the bubble size.

Now the plot is complete and serves as a useful basis for discussions in the segmentation team. Using Fig. 10.1 as a basis, the segmentation team may, for example, eliminate from further consideration segments 3 and 7 because they are rather unattractive compared to the other available segments despite the fact that they have high profit potential (as indicated by the size of the bubbles). Segment 5 is obviously highly attractive and has high profit potential, but unfortunately the segment is not as fond of the organisation as the organisation is of the segment. It is unlikely, at this point in time, that the organisation will be able to cater

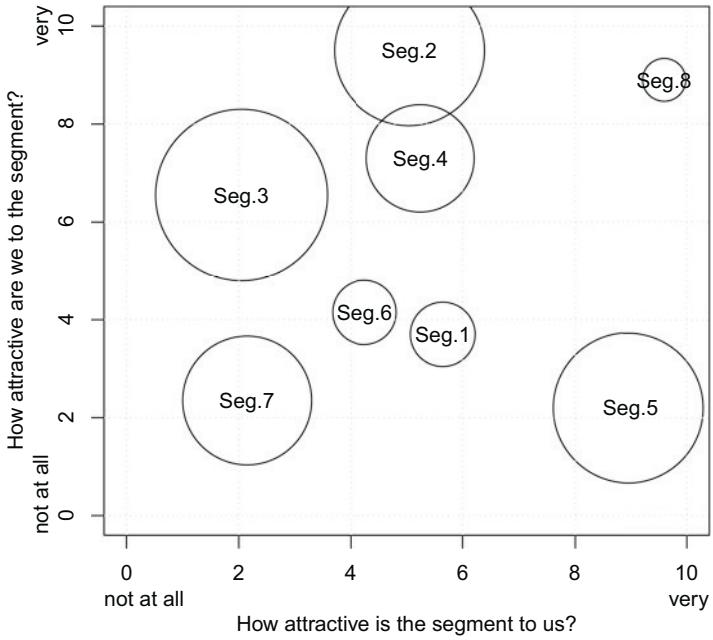


Fig. 10.1 Segment evaluation plot

successfully to segment 5. Segment 8 is excellent because it is highly attractive to the organisation, and views the organisation's offer as highly attractive. A match made in heaven, except for the fact that the profit potential is not very high. It may be necessary, therefore to consider including segment 2. Segment 2 loves the organisation, has decent profit potential, and is about equally attractive to the organisation as segments 1, 4 and 6 (all of which, unfortunately, are not very fond of the organisation's offer).

To re-create the plot in R, we store the upper half (without row "Total") of Table 10.1 in the 5×8 matrix `x`, the corresponding weights from the second column in vector `wx`, the lower half of Table 10.1 in the 5×8 matrix `y`, and weights in vector `wy`. We then create the segment evaluation plot of the decision matrix using the following commands.

```
R> library("MSA")
R> decisionMatrix(x, y, wx, wy, size = size)
```

where vector `size` controls the bubble size for each segment (e.g., profitability).

10.3 Step 8 Checklist

Task	Who is responsible?	Completed?
Convene a segmentation team meeting.		☐
Determine which of the market segments profiled in Step 6 and described in Step 7 are being considered as potential target markets.		☐
Double check that all of those remaining segments comply with the knock-out criteria of homogeneity, distinctness, size, match, identifiability and reachability. If a segment does not comply: eliminate it from further consideration.		☐
Discuss and agree on values for each market segment for each segment attractiveness criterion.		☐
Discuss and agree on values for each relative organisational competitiveness criterion for each of the market segments.		☐
Calculate each segment's overall attractiveness by multiplying the segment value with the weight for each criterion and then summing up all these values for each segment.		☐
Calculate each segment's overall relative organisational competitiveness by multiplying the segment value with the weight for each criterion and then summing up all these values for each segment.		☐
Plot the values onto a segment evaluation plot.		☐
Make a preliminary selection.		☐
If you intend to target more than one segment: make sure that the selected target segments are compatible with one another.		☐
Present the selected segments to the advisory committee for discussion and (if required) reconsideration.		☐

References

Dibb S, Simkin L (2008) Market segmentation success: making it happen! Routledge, New York

Lilien GL, Rangaswamy A (2003) Marketing engineering: computer-assisted marketing analysis and planning, 2nd edn. Prentice Hall, Upper Saddle River

McDonald M, Dunbar I (1995) Market segmentation: a step-by-step approach to creating profitable market segments. Macmillan, London

Optimizing Marketing Mix Strategy

11.1 Implications for Marketing Mix Decisions

Marketing was originally seen as a toolbox to assist in selling products, with marketers mixing the ingredients of the toolbox to achieve the best possible sales results (Dolnicar and Ring 2014). In the early days of marketing, Borden (1964) postulated that marketers have at their disposal 12 ingredients: product planning, packaging, physical handling, distribution channels, pricing, personal selling, branding, display, advertising, promotions, servicing, fact finding and analysis. Many versions of this marketing mix have since been proposed, but most commonly the marketing mix is understood as consisting of the *4Ps*: Product, Price, Promotion and Place (McCarthy 1960).

Market segmentation does not stand independently as a marketing strategy. Rather, it goes hand in hand with the other areas of strategic marketing, most importantly: positioning and competition. In fact, the segmentation process is frequently seen as part of what is referred to as the *segmentation-targeting-positioning* (STP) approach (Lilien and Rangaswamy 2003). The segmentation-targeting-positioning approach postulates a sequential process. The process starts with *market segmentation* (the extraction, profiling and description of segments), followed by *targeting* (the assessment of segments and selection of a target segment), and finally *positioning* (the measures an organisation can take to ensure that their product is perceived as distinctly different from competing products, and in line with segment needs).

Viewing market segmentation as the first step in the segmentation-targeting-positioning approach is useful because it ensures that segmentation is not seen as independent from other strategic decisions. It is important, however, not to adhere too strictly to the sequential nature of the segmentation-targeting-positioning process. It may well be necessary to move back and forward from the segmentation to the targeting step, before being in the position of making a long-term commitment to one or a small number of target segments.

Fig. 11.1 How the target segment decision affects marketing mix development

Figure 11.1 illustrates how the target segment decision – which has to be integrated with other strategic areas such as competition and positioning – affects the development of the marketing mix. For reasons of simplicity, the traditional 4Ps model of the marketing mix including Product, Price, Place and Promotion serves as the basis of this discussion. Be it twelve or four, each one of those aspects needs to be thoroughly reviewed once the target segment or the target segments have been selected.

To best ensure maximising on the benefits of a market segmentation strategy, it is important to customise the marketing mix to the target segment (see also the layers of market segmentation in Fig. 2.1 discussed on pages 11–12). The selection of one or more specific target segments may require the design of new, or the modification or re-branding of existing products (Product), changes to prices or discount structures (Price), the selection of suitable distribution channels (Place), and the development of new communication messages and promotion strategies that are attractive to the target segment (Promotion).

One option available to the organisation is to structure the entire market segmentation analysis around one of the 4Ps. This affects the choice of segmentation variables. If, for example, the segmentation analysis is undertaken to inform pricing decisions, price sensitivity, deal proneness, and price sensitivity represent suitable segmentation variables (Lilien and Rangaswamy 2003).

If the market segmentation analysis is conducted to inform advertising decisions, benefits sought, lifestyle segmentation variables, and psychographic segmentation variables are particularly useful, as is a combination of all of those (Lilien and Rangaswamy 2003).

If the market segmentation analysis is conducted for the purpose of informing distribution decisions, store loyalty, store patronage, and benefits sought when selecting a store may represent valuable segmentation variables (Lilien and Rangaswamy 2003). Typically, however, market segmentation analysis is not conducted in view of one of the 4Ps specifically. Rather, insights gained from the detailed description of the target segment resulting from Step 7 guide the organisation in how to develop or adjust the marketing mix to best cater for the target segment chosen.

11.2 Product

One of the key decisions an organisation needs to make when developing the product dimension of the marketing mix, is to specify the product in view of customer needs. Often this does not imply designing an entirely new product, but rather modifying an existing one. Other marketing mix decisions that fall under the product dimension are: naming the product, packaging it, offering or not offering warranties, and after sales support services.

The market segments obtained for the Australian vacation activities data set (see Appendix C.3) using biclustering (profiled in Fig. 7.37) present a good opportunity for illustrating how product design or modification is driven by target segment selection. Imagine, for example, being a destination with a very rich cultural heritage. And imagine having chosen to target segment 3. The key characteristics of segment 3 members in terms of vacation activities are that they engage much more than the average tourist in visiting museums, monuments and gardens (see the bicluster membership plot in Fig. 7.37). They also like to do scenic walks and visit markets. They share both of these traits with some of the other market segments. Like most other segments, they like to relax, eat out, shop and engage in sightseeing.

In terms of the product targeted at this market segment, possible product measures may include developing a new product. For example, a MUSEUMS, MONUMENTS & MUCH, MUCH MORE product (accompanied by an activities pass) that helps members of this segment to locate activities they are interested in, and points to the existence of these offers at the destination during the vacation planning process. Another opportunity for targeting this segment is that of proactively making gardens at the destination an attraction in their own right.

11.3 Price

Typical decisions an organisation needs to make when developing the price dimension of the marketing mix include setting the price for a product, and deciding on discounts to be offered.

Sticking to the example of the destination that wishes to market to segment 3 (which has emerged from a biclustering analysis of the Australian vacation activities data set), we load the bicluster solution obtained in Sect. 7.4.1:

```
R> load("ausact-bic.RData")
```

To be able to compare members of segment 3 to tourists not belonging to segment 3, we construct a binary vector containing this information from the bicluster solution. We first extract which rows (respondents) and columns (activities) are contained in a segment using:

```
R> library("biclust")
R> bcn <- biclusternumber(ausact.bic)
```

We use this information to construct a vector containing the segment membership for each consumer.

First we initialise a vector `cl12` containing only missing values (`NA`s) with the length equal to the number of consumers. Then we loop through the different clusters extracted by the biclustering algorithm, and assign the rows (respondents) contained in this cluster the corresponding cluster number in `cl12`.

```
R> data("ausActiv", package = "MSA")
R> cl12 <- rep(NA, nrow(ausActiv))
R> for (k in seq_along(bcn)) {
+   cl12[bcn[[k]]$Rows] <- k
+  }
```

The resulting segment membership vector contains numbers 1 to 12 because biclustering extracted 12 clusters. It also contains missing values because biclustering does not assign all consumers to a cluster. We obtain the number of consumers assigned to each segment, and the number of consumers not assigned by tabulating the vector:

```
R> table(cl12, exclude = NULL)

cl12
   1    2    3    4    5    6    7    8    9   10   11   12
  50   57   67   73   61   83   52   65   51   53   80   60
<NA>
 251
```

The argument `exclude = NULL` ensures that `NA` values are included in the frequency table.

Based on the segment membership vector, we create a binary variable indicating if a consumer is assigned to segment 3 or not. We do this by selecting those as being in segment 3 who are not `NA` (`!is.na(cl12)`), and where the segment membership value is equal to 3.

```
R> cl12.3 <- factor(!is.na(cl12) & cl12 == 3,
+   levels = c(FALSE, TRUE),
+   labels = c("Not Segment 3", "Segment 3"))
```

The categories are specified in the second argument `levels`. Their names are specified in the third argument `labels`.

Additional information on consumers is available in the data frame `ausActivDesc` in package MSA. We use the following command to load the data, and create a parallel boxplot of the variable SPEND PER PERSON PER DAY split by membership in segment 3:

```
R> data("ausActivDesc", package = "MSA")
R> boxplot(spendpppd ~ cl12.3, data = ausActivDesc,
+   notch = TRUE, varwidth = TRUE, log = "y",
+   ylab = "AUD per person per day")
```

The additional arguments specify that confidence intervals for the median estimates should be included (`notch = TRUE`), box widths should reflect group sizes (`varwidth = TRUE`), that the *y*-axis should be on the log scale because of the right-skewness of the distribution (`log = "y"`), and that a specific label should be included for the *y*-axis (`ylab`).

Figure 11.2 shows the expenditures of segment 3 members on the right, and those of all other consumers on the left. Ideally, we would have information about actual expenditures across a wide range of expenditure categories, or information about price elasticity, or reliable information about the segment's willingness to pay for a range of products. But the information contained in Fig. 11.2 is still valuable. It illustrates how the price dimension can be used to best possibly harvest the targeted marketing approach.

As can be seen in Fig. 11.2, members of segment 3 have higher vacation expenditures per person per day than other tourists. This is excellent news for the tourist destination; it does not need to worry about having to offer the MUSEUMS, MONUMENTS & MUCH, MUCH MORE product at a discounted price. If anything, the insights gained from Fig. 11.2 suggest that there is potential to attach a premium price to this product.

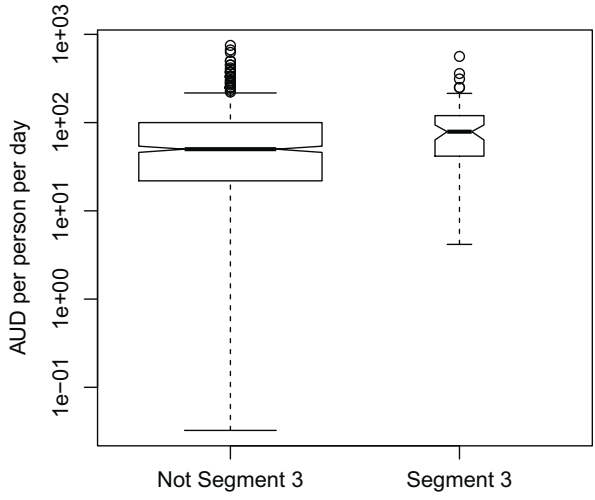


Fig. 11.2 Total expenditures in Australian dollars (AUD) for the last domestic holiday for tourists in segment 3 and all other tourists

11.4 Place

The key decision relating to the place dimension of the marketing mix is how to distribute the product to the customers. This includes answering questions such as: should the product be made available for purchase online or offline only or both; should the manufacturer sell directly to customers; or should a wholesaler or a retailer or both be used.

Returning to the example of members of segment 3 and the destination with a rich cultural heritage: the survey upon which the market segmentation analysis was based also asked survey respondents to indicate how they booked their accommodation during their last domestic holiday. Respondents could choose multiple options. This information is place valuable; knowing the booking preferences of members of segment 3 enables the destination to ensure that the MUSEUMS, MONUMENTS & MUCH, MUCH MORE product is bookable through these very distribution channels.

We can use `propBarchart` from package flexclust to visualise stated booking behaviour. First we load the package. Then we call function `propBarchart()` with the following arguments: `ausActivDesc` contains the data, `g = cl12.3` specifies segment membership, and `which` indicates the columns of the data to be used. We select all columns with column names starting with `"book"`. Function `grep` based on *regular expressions* extracts those columns. For more details see the help page of `grep`. Alternatively, we can use `which = startsWith(names(ausActivDesc), "book")` instead of `which = grep("^book", names(ausActivDesc))`.

```
R> library("flexclust")
R> propBarchart(ausActivDesc, g = cl12.3,
+   which = grep("^book", names(ausActivDesc)),
+   layout = c(1, 1), xlab = "percent", xlim = c(-2, 102))
```

The additional arguments specify: that only one panel should be included in each plot (`layout = c(1, 1)`), the label for the x-axis (`xlab`), and the limits for the x-axis (`xlim`). Figure 11.3 shows the resulting plot for members in segment 3.

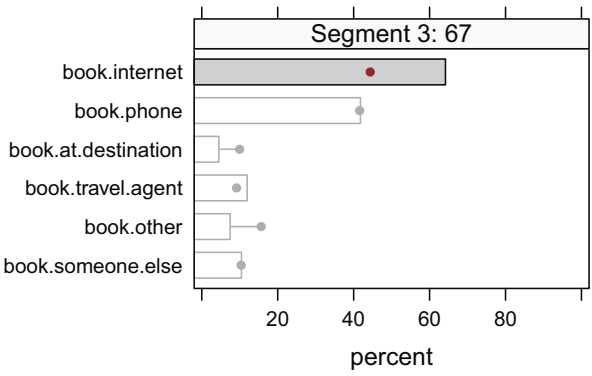


Fig. 11.3 Hotel booking avenues used for the last domestic holiday by segment 3 and by the average tourist

Figure 11.3 indicates that members of segment 3 differ from other tourists in terms of how they booked their hotel on their last domestic vacation: they book their hotel online much more frequently than the average tourist. This information has clear implications for the place dimension of the marketing mix. There must be an online booking option available for the hotel. It would be of great value to also collect information about the booking of other products, services and activities by members of segment 3 to see if most of their booking activity occurs online, or if their online booking behaviour is limited to the accommodation.

11.5 Promotion

Typical promotion decisions that need to be made when designing a marketing mix include: developing an advertising message that will resonate with the target market, and identifying the most effective way of communicating this message. Other tools in the promotion category of the marketing mix include public relations, personal selling, and sponsorship.

Looking at segment 3 again: we need to determine the best information sources for reaching members of segment 3 so we can inform them about the MUSEUMS, MONUMENTS & MUCH, MUCH MORE product. We answer this question by comparing the information sources they used for the last domestic holiday, and by investigating their preferred TV stations.

We obtain a plot comparing the use of the different information sources to choose a destination for their last domestic holiday with the same command as used for Fig. 11.3, except that we use the variables starting with `"info"`:

```
R> propBarchart(ausActivDesc, g = cl12.3,
+   which = grep("^info", names(ausActivDesc)),
+   layout = c(1, 1), xlab = "percent",
+   xlim = c(-2, 102))
```

As Fig. 11.4 indicates, members of segment 3 rely – more frequently than other tourists – on information provided by tourist centres when deciding where to spend their vacation. This is a very distinct preference in terms of information sources. One way to use this insight to design the promotion component of the marketing mix is to have specific information packs on the MUSEUMS, MONUMENTS & MUCH, MUCH MORE product available both in hard copy in the local tourist information centre at the destination as well as making it available online on the tourist information centre's web page.

The mosaic plot in Fig. 11.5 shows TV channel preference. We generate Fig. 11.5 with the command:

```
R> par(las = 2)
R> mosaicplot(table(cl12.3, ausActivDesc$TV.channel),
+   shade = TRUE, xlab = "", main = "")
```

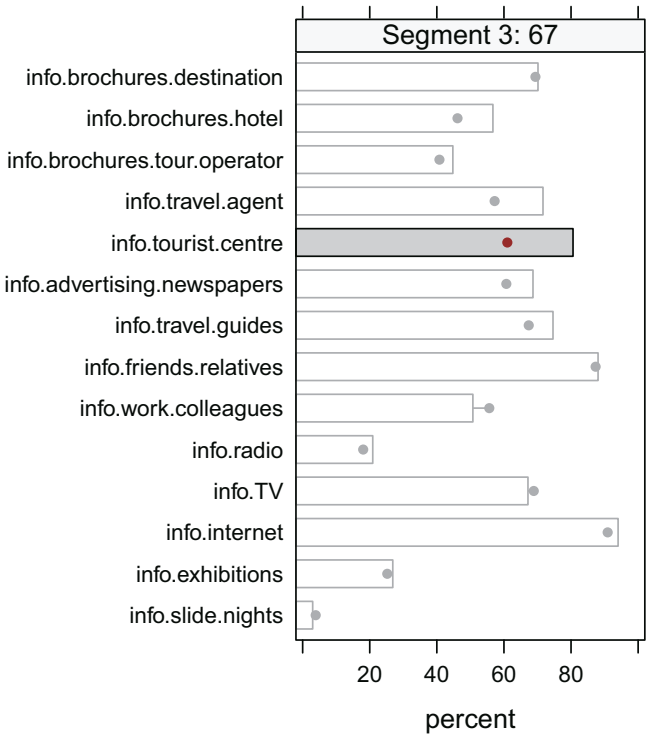


Fig. 11.4 Information sources used by segment 3 and by the average tourist.

We use `par(las = 2)` to ensure that axis labels are vertically aligned for the x-axis, and horizontally aligned for the y-axis. This makes it easier to fit the channel names onto the plot.

Figure 11.5 points to another interesting piece of information about segment 3. Its members have a TV channel preference for Channel 7, differentiating them from other tourists. Again, it is this kind of information that enables the destination to develop a media plan ensuring maximum exposure of members of segment 3 to the targeted communication of, for example, a MUSEUMS, MONUMENTS & MUCH, MUCH MORE product.

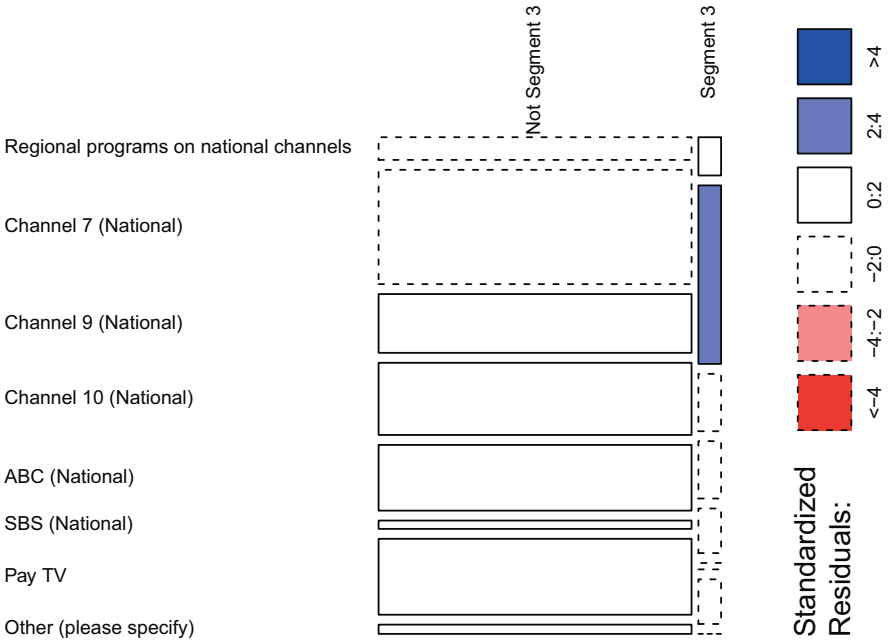


Fig. 11.5 TV station most frequently watched by segment 3 and all other tourists

11.6 Step 9 Checklist

Task	Who is responsible?	Completed?
Convene a segmentation team meeting.		☐
Study the profile and the detailed description of the target segment again carefully.		☐
Determine how the product-related aspects need to be designed or modified to best cater for this target segment.		☐
Determine how the price-related aspects need to be designed or modified to best cater for this target segment.		☐
Determine how the place-related aspects need to be designed or modified to best cater for this target segment.		☐
Determine how the promotion-related aspects need to be designed or modified to best cater for this target segment.		☐
Review the marketing mix in its entirety.		☐
If you intend to target more than one segment: repeat the above steps for each of the target segments. Ensure that segments are compatible with one another.		☐
Present an outline of the proposed marketing mix to the advisory committee for discussion and (if required) modification.		☐

References

Borden NH (1964) The concept of the marketing mix. J Advert Res 4:2–7

Dolnicar S, Ring A (2014) Tourism marketing research – past, present and future. Ann Tour Res 47:31–47

Lilien GL, Rangaswamy A (2003) Marketing engineering: computer-assisted marketing analysis and planning, 2nd edn. Prentice Hall, Upper Saddle River

McCarthy JE (1960) Basic marketing: a managerial approach. Richard D. Irwin, Homewood

Segment Evolution and Monitoring

12.1 Ongoing Tasks in Market Segmentation

Market segmentation analysis does not end with the selection of the target segment, and the development of a customised marketing mix. As Lilien and Rangaswamy (2003, p. 103) state segmentation must be viewed as an ongoing strategic decision process. Haley (1985, p. 261) elaborates as follows: The world changes . . . virtually the only practical option for an intelligent marketer is to monitor his or her market continuously. After the segmentation strategy is implemented, two additional tasks need to be performed on an ongoing basis:

1. The effectiveness of the segmentation strategy needs to be evaluated. Much effort goes into conducting the market segmentation analysis, and customising the marketing mix to best satisfy the target segment's needs. These efforts should result in an increase in profit, or an increase in achievement of the organisational mission. If they did not, the market segmentation strategy failed.
2. The market is not static. Consumers change, the environment, and actions of competitors change. As a consequence, a process of ongoing monitoring of the market segmentation strategy must be devised. This monitoring process can range from a regular review by the segmentation team, to a highly automatised data mining system alerting the organisation to any relevant changes to the size or nature of the target segment.

12.2 Evaluating the Success of the Segmentation Strategy

The aim of evaluating the effectiveness of the market segmentation strategy is to determine whether developing a customised marketing mix for one or more segments did achieve the expected benefits for the organisation. In the short term,

the primary desired outcome for most organisations will be increased profit. For non for profit organisations it may be some other performance criterion, such as the amount of donations raised or number of volunteers recruited. These measures can be monitored continuously to allow ongoing assessment of the segmentation strategy. In addition, taking a longer term perspective, the effectiveness of targeted positioning could be measured. For example, a tracking study would provide insight about how the organisation is perceived in the market place. If the segmentation strategy is successful, the organisation should increasingly be perceived as being particularly good at satisfying certain needs. If this is the case, the organisation should derive a competitive advantage from this specialised positioning because the target segment will perceive it as one of their preferred suppliers.

12.3 Stability of Segment Membership and Segment Hopping

A number of studies have investigated change of market segment membership of respondents over time (Boztug et al. 2015). In the context of banking, Calantone and Sawyer (1978) find that – over a two-year period of time – fewer than one third of bank customers remained in the same benefit segment. Similarly, Yuspeh and Fein (1982) conclude that only 40% of the respondents in their study fell into the same market segment two years later. Farley et al. (1987) estimate that half of all households change in a two-year period when segmented on the basis of their consumption patterns. Müller and Hamm (2014) confirm the low stability of segment membership over time in a three-year study. Paas et al. (2015) analyse the long-term developments of financial product portfolio segments in several European countries over more than three decades. They use only cross-sectional data sets for the different time points, but are able to identify changes in segment structure at country level over time, implying instability of segment membership.

Changes in segment membership are problematic if (1) segment sizes change (especially if the target segment shrinks), and if (2) the nature of segments changes in terms of either segmentation or descriptor variables. Changes in segment size may require a fundamental rethinking of the segmentation strategy. Changes in segment characteristics could be addressed through a modification of the marketing mix.

The changes discussed so far represent a relative slow evolution of the segment landscape. In some product categories, segment members change segments regularly, they *segment hop*. Segment hopping does not occur spuriously. It can be caused by a number of factors. For example, the same product may be used in different situations, and different product features may matter in those different situations; consumers may seek variety; or they may react to different promotional offers. Haley (1985) already discussed the interaction of consumption occasions and benefits sought, recommending to use both aspects to ensure maximum insight.

For example, the following scenario is perfectly plausible: a family spends their vacation camping. Their key travel motives are to experience nature, to get away from the hustle and bustle of city living, and to engage in outdoor activities. The

family stays for two weeks, but their expenditures per person per day are well below those of an average tourist. Imagine that one of the parents, say the mother, is asked – after the family camping trip – to complete a survey about their last vacation. Data from this survey is used in a market segmentation analysis and the mother is assigned to the segment of NATURE LOVING FAMILIES ON A TIGHT BUDGET. A month later, the mother and the father celebrate their anniversary. They check into a luxury hotel in a big city for one night only, indulge in a massage and spa treatment, and enjoy a very fancy and very expensive dinner. Now the mother is again asked to complete the same survey. Suddenly, she is classified as a BIG SPENDING, SHORT STAY CITY TOURIST.

These tourists segment hop. This phenomenon has previously been observed and segment hopping consumers have been referred to as *centaurs* (Wind et al. 2002) or *hybrid consumers* (Wind et al. 2002; Ehrnrooth and Grönroos 2013).

Consumer hybridity of this kind – or segment hopping – has been discussed in Bieger and Laesser (2002), and empirically demonstrated in the tourism context by Boztug et al. (2015). The latter study estimates that 57% of the Swiss population display a high level of segment hopping in terms of travel motives, and that 39% segment hop across vacation expenditure segments.

Ha et al. (2002) model segment hopping using Markov chains. They use self-organizing maps (SOMs) to extract segments from a customer relationship management database; and Markov chains to model changes in segment membership over time. Lingras et al. (2005) investigate segment hopping using a modified self-organizing maps (SOMs) algorithm. They study segment hopping among supermarket customers over a period of 24 weeks; consumers are assigned to segments for every four week period and their switching behaviour is modelled.

Another possible interpretation of the empirical observation of segment hopping is that there may be a distinct market segment of *segment hoppers*. This notion has first been investigated by Hu and Rau (1995) who find segment hoppers to share a number of socio-economic and demographic characteristics. Boztug et al. (2015) also ask if segment hoppers are a segment in their own right, concluding that segment hoppers (in their tourism-related data set from Swiss residents) are older, describe themselves more frequently as calm, modest, organised and colourless, and more frequently obtain travel-related information from advertisements.

Accepting that segment hopping occurs has implications for market segmentation analysis, and the translation of findings from market segmentation analysis into marketing action. Most critically, we cannot assume that consumers are well behaved and stay in the segments. Optimally, we could estimate how many segment members are hoppers. Those may need to be excluded or targeted in a very specific way. Returning to our example: once the annual vacation pattern of the camping family is understood, we may be able to target information about luxury hotels at this family as they return from the camping trip.

12.4 Segment Evolution

Segments evolve. Like any characteristic of markets, market segments change over time. The environments in which the organisation operates, and actions taken by competitors change. Haley (1985), the father of benefit segmentation, says that not following-up a segmentation study means sacrificing a substantial part of the value it is able to generate. Haley (1985) proceeds to recommend a tracking system to ensure that any changes are identified as early as possible and acted upon. Haley refers to the tracking system as an *early warning system* activating action only if an irregularity is detected. Or, as Cahill (2006) puts it (p. 38): Keep testing, keep researching, keep measuring. People change, trends change, values change, everything changes.

A number of reasons drive genuine change of market segments, including: evolution of consumers in terms of their product savviness or their family life cycle; the availability of new products in the category; and the emergence of disruptive innovations changing a market in its entirety.

To be able to assess potential segment evolution correctly, we need to know the baseline stability of market segments. The discussions in Sects. 2.3, 7.5.3, and 7.5.4 demonstrate that – due to the general lack of natural segments in empirical consumer data – most segmentation solutions and segments are unstable, even if segment extraction is repeated a few seconds later with data from the same population and the same extraction algorithm. It is critical, therefore, to conduct stability analysis at both the global level and the segment level to determine the baseline stability. Only if this information is available, can instability over time be correctly interpreted.

Assuming that genuine segment evolution is taking place, a number of approaches can simultaneously extract segments, *and* model segment evolution over time. The MONIC framework developed by Spiliopoulou et al. (2006) allows the following segment evolution over time: segments can remain unchanged, segments can be merged, existing segments can be split up, segments can disappear, and completely new segments can emerge. This method uses a series of segmentation solutions over time, and compares those next to each other in time. For the procedure to work automatically, repeated measurements for at least a subset of the segment members have to be available for neighbouring points in time; the data needs to be truly longitudinal.

A similar approach is used by Oliveira and Gama (2010). In their framework, the following taxonomy is used for changes in segments over time:

- Birth: a new segment emerges.
- Death: an existing segment disappears.
- Split: one segment is split up.
- Merge: segments are merged.
- Survival: a segment remains almost unchanged.

The procedure can only be automated if the same consumers are repeatedly segmented over time; data must be truly longitudinal. The application by Oliveira

and Gama (2010) uses three successive years, and, in their study, the clustered objects are not consumers, but economic activity sectors. If different objects are available in different years (as is the case in typical repeat cross-sectional survey studies), the framework can still be used, but careful matching of segments based on their profiles is required.

To sum up: ignoring dynamics in market segments is very risky. It can lead to customising product, price, promotion and place to a segment that existed a few years ago, but has since changed its expectations or behaviours. It is critical, therefore, to determine stability benchmarks initially, and then set up a process to continuously monitor relevant market dynamics.

Being the first organisation to adapt to change is a source of competitive advantage. And, in times of big data where fresh information about consumers becomes available by the second, the source of competitive advantage will increasingly shift from the ability to adapt to the capability to identify relevant changes quickly. Relevant changes include changes in segment needs, changes in segment size, changes in segment composition, changes in the alternatives available to the segment to satisfy their needs as well as general market changes, like recessions.

McDonald and Dunbar (1995, p. 10) put it very nicely in their definition of market segmentation: Segmentation is a creative and iterative process, the purpose of which is to satisfy consumer needs more closely and, in so doing, create competitive advantage for the company. It is defined by the customers' needs, not the company's, and should be re-visited periodically.

Example: Winter Vacation Activities

To illustrate monitoring of market segments over time, we use the data set on winter activities of tourists to Austria in 1997/98 (see Appendix C.2). We used this data set in Sect. 7.2.4.2 to illustrate bagged clustering. Here, we use a reduced set of 11 activities as segmentation variables. These 11 activities include all the key winter sports (such as alpine skiing), and a few additional activities which do not reflect the main purpose of people's vacation. Importantly, we have the same information about winter activities available for the 1991/92 winter season. These two data sets are repeat cross-sectional – rather than truly longitudinal – because different tourists participated in the two survey waves.

Package MSA contains both data sets (`wi91act`, `wi97act`). We can load the data, and calculate the overall means for all activities for 1991/92 and 1997/98 using the following R commands:

```
R> data("winterActiv2", package = "MSA")
R> p91 <- colMeans(wi91act)
R> round(100 * p91)

       alpine skiing cross-country skiing          ski touring
                  71                    18                    9
```

```
          ice-skating        sleigh riding                hiking
                    6                   16                    30
             relaxing              shopping         sight-seeing
                   51                   25                    11
              museums            pool/sauna
                    6                   30
```

```
R> p97 <- colMeans(wi97act)
R> round(100 * p97)
```

```
       alpine skiing cross-country skiing          ski touring
                  68                    9                    3
          ice-skating        sleigh riding                hiking
                   5                   14                   29
             relaxing              shopping         sight-seeing
                  74                   55                   30
              museums            pool/sauna
                  14                   47
```

The resulting output lists the winter activities, along with the percentage of tourists in the entire sample who engage in those activities. We visualise differences in these percentages across the two survey waves using a dot chart (Fig. 12.1). The vertical grid line crosses the x-axis at zero; dots along the vertical line indicate that there is no difference in the percentage of tourists engaging in that particular winter activity between survey waves 1991/92 and 1997/98. The following R code generates the dot chart of sorted differences, and adds a vertical dashed line at zero (`abline()` with line type `lty = 2`):

```
R> dotchart(100 * sort(p97 - p91),
+   xlab = paste("difference",
+   "in percentages undertaking activity in '91 and '97"))
R> abline(v = 0, lty = 2)
```

Figure 12.1 indicates that the aggregate increase in pursuing a specific activity is largest for shopping (shown at the top of the plot): the percentage of tourists going shopping during their winter vacation increased by 30% points from 1991/92 to 1997/98. The largest decrease in aggregate activity level occurs for cross-country skiing. For a number of other activities – ice-skating, hiking, sleigh riding, and alpine skiing – the percentages are almost identical in both waves.

So far we explored the data at aggregate level. To account for heterogeneity, we extract market segments using the data from the 1991/92 winter season. In a first step we conduct stability analysis across a range of segmentation solutions. Stability analysis indicates that natural market segments do not exist; the stability results do not offer a firm recommendation about the best number of segments to extract. Based on the manual inspection of a number of alternative segmentation solutions with different numbers of market segments, we select the six-segment solution for further inspection.

We extract the six-segment solution for the 1991/92 winter season data using the standard k-means partitioning clustering algorithm:

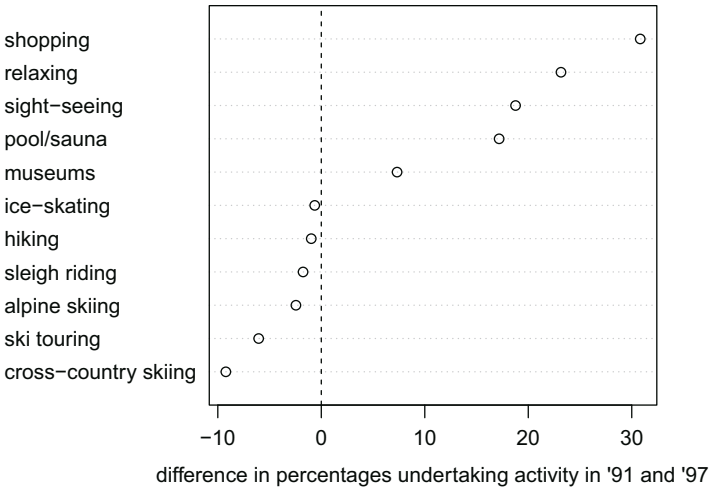


Fig. 12.1 Difference in the percentage of tourists engaging in 11 winter vacation activities during their vacation in Austria in 1991/92 and 1997/98

```
R> library("flexclust")
R> set.seed(1234)
R> wi91act.k6 <- stepcclust(wi91act, k = 6, nrep = 20)
```

where `k` specifies the number of segments to extract, and `nrep` specifies the number of random restarts.

We then use the following R code to generate a segment profile plot for the 1991/92 data. We highlight marker variables (`shade = TRUE`), and specify for each panel label to start with `"Segment "`:

```
R> barchart(wi91act.k6, shade = TRUE,
+   strip.prefix = "Segment ")
```

Figure 12.2 contains the resulting segment profile plot. We see that market segment 1 is distinctly different from the other segments because members of this segment like to go hiking, sight-seeing, and visiting museums during their winter vacation in Austria. Members of market segment 2 engage in alpine skiing (although not much more frequently than the average tourist in the sample), and go to the pool/sauna. Members of market segment 3 like skiing and relaxing; members of segment 4 are all about alpine skiing; members of segment 5 engage in a wide variety of vacation activities, as do members of segment 6.

To monitor whether – six years later – this same market segmentation solution is still a good basis for target marketing by the Austrian National Tourism Organisation, we explore changes in the segmentation solution in the 1997/98 data set. We first use the segmentation solution for 1991/92 to predict segment memberships in 1997/98. Then we assess differences in segment sizes by determining the percentages of tourists assigned to each of the segments for the two waves:

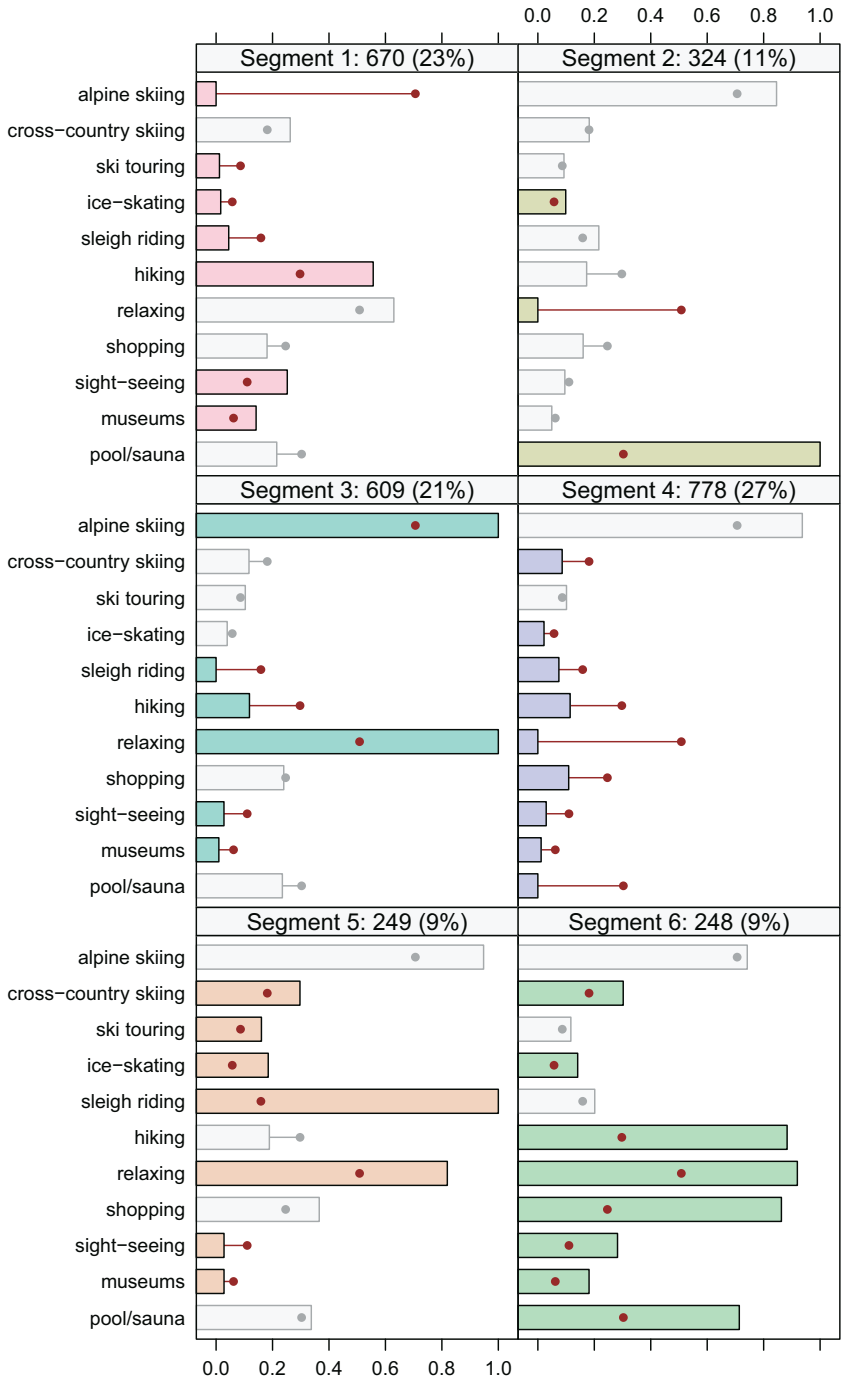


Fig. 12.2 Segment profile plot for the six-segment solution of winter vacation activities in 1991/92

```
R> size91 <- table(clusters(wi91act.k6))
R> size97 <- table(clusters(wi91act.k6,
+   newdata = wi97act))
R> round(prop.table(rbind(size91, size97), 1) * 100)

        1  2  3  4 5  6
size91 23 11 21 27 9  9
size97 22  7 29 12 9 21
```

The comparison of segment sizes indicates that segments 1 and 5 are relatively stable in size, whereas segments 4 and 6 change substantially. We use a χ^2-test to test if these differences could have occurred by chance:

```
R> chisq.test(rbind(size91, size97))

        Pearson's Chi-squared test

data:  rbind(size91, size97)
X-squared = 375.35, df = 5, p-value < 2.2e-16
```

The χ^2-test indicates that segment sizes did indeed change significantly. We can visualise the comparison in a mosaic plot (Fig. 12.3):

```
R> mosaicplot(rbind("1991" = size91, "1997" = size97),
+   ylab = "Segment", shade = TRUE, main = "")
```

The mosaic plot indicates that some segments (1 and 5) did not change in size, that segment 4 shrunk, and that segment 6 nearly doubled. Depending on the target segment chosen initially, these results can be good or bad news for the Austrian National Tourism Organisation. If we also had descriptor variables available for both periods of time, we could also study differences in those characteristics.

In a second step we assess the evolution of market segments. We extract segments from the 1997/98 data. Optimally, we would use truly longitudinal data (containing

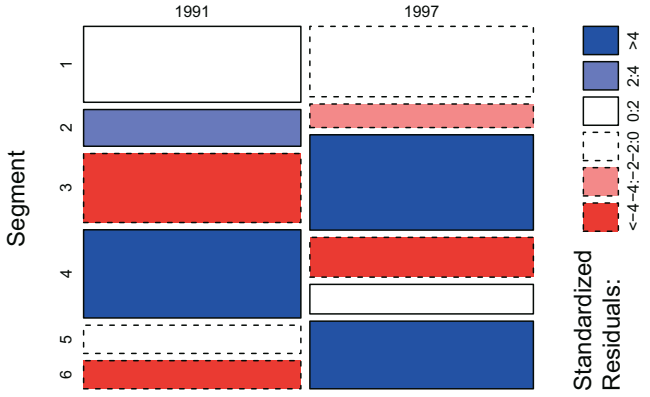


Fig. 12.3 Mosaic plot comparing segment sizes in 1991/92 and 1997/98 based on the segmentation solution for winter activities in 1991/92

responses from the same tourists at both points in time). Longitudinal data would allow keeping the segment assignment of tourists fixed, and assessing whether segment profiles changed over time. Given that only repeat cross-section data are available, we extract new segments using centroids (cluster centres, segment representatives) from the 1991/92 segmentation to start off the segment extraction for the 1997/98 data. We obtain the new segmentation solution using the previous centroids as initial values (argument `k`) for *k*-means clustering of the 1997/98 data using:

```
R> wi97act.k6 <- cclust(wi97act,
+   k = parameters(wi91act.k6))
```

The following R command generates the segment profile plot for the market segmentation solution of the 1997/98 data:

```
R> barchart(wi97act.k6, shade = TRUE,
+   strip.prefix = "Segment ")
```

We see in Fig. 12.4 that the resulting segmentation solution is very similar to that based on the 1991/92 data. We can conclude that the nature of tourist segments has not changed; the same types of tourist segments still come to Austria six years later.

Segment evolution is visible in the variable shopping, pursued to a large extent by tourists in segment 6 and nearly half of all tourists. The aggregate analysis already pointed to this increase in shopping activity: a quarter of winter tourists to Austria went shopping in 1991/92; more than half did so in 1997/98. This change might be explained by the liberalisation of opening hours for shops in Austria in 1992.

Another obvious difference is the change in segment sizes. Segment 4 (interested primarily in alpine skiing) contained 27% of tourists in 1991, but only 13% in 1997. Segments 3 and 6 increased substantially in size, suggesting that more people combine alpine skiing with relaxation, and more people engage in a broader portfolio of winter activities.

These changes in segment sizes have implications for the Austrian National Tourism Organisation. While in 1991/92 a third of winter tourists to Austria would have been quite satisfied to ski, eat and sleep, the Austrian National Tourism Organisation would be well advised six years later to offer tourists a wider range of activities.

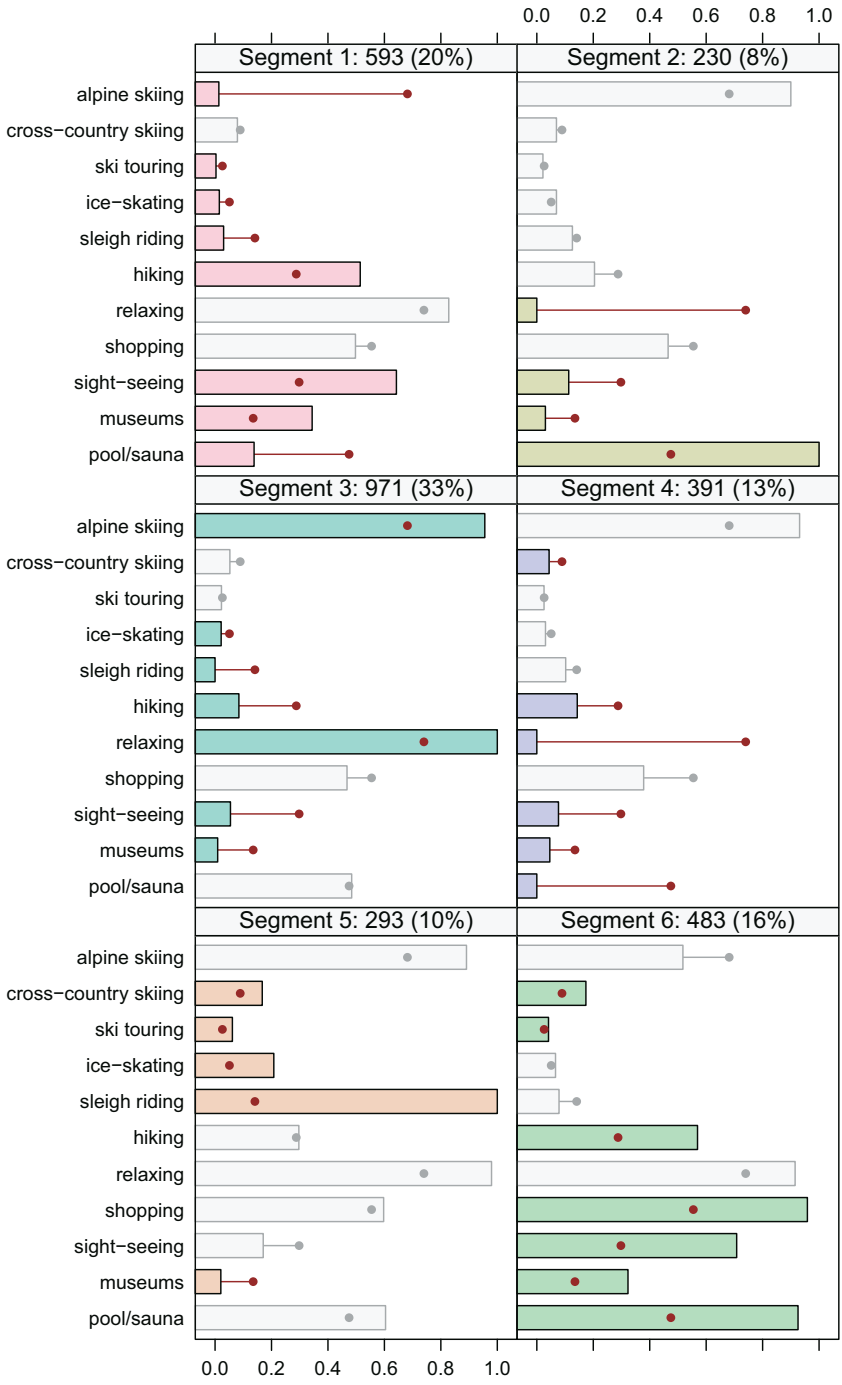


Fig. 12.4 Segment profile plot for the six-segment solution of winter vacation activities in 1997/98

12.5 Step 10 Checklist

Task	Who is responsible?	Completed?
Convene a segmentation team meeting.		☐
Determine which indicators of short-term and long-term success will be used to evaluate the market segmentation strategy.		☐
Operationalise how segmentation success indicators will be measured and how frequently.		☐
Determine who will be responsible for collecting data on these indicators.		☐
Determine how often the segmentation team will re-convene to review the indicators.		☐
Determine which indicators will be used to capture market dynamics.		☐
Remind yourself of the baseline *global stability* to ensure that the source of instability is attributed to the correct cause.		☐
Remind yourself of the baseline *segment level stability* to ensure that the source of instability is attributed to the correct cause.		☐
Operationalise how market dynamics indicators will be measured and how frequently.		☐
Determine who will be responsible for collecting data on market dynamics.		☐
Determine how often the segmentation team will re-convene to review the market dynamics indicators or whether the collecting unit will pro-actively alert the segmentation team if a meeting is required.		☐
Develop an adaptation checklist specifically for your organisation of things that need to happen quickly across the affected organisational units if a critical change is detected.		☐
Run the indicators, measures of indicators, reviewing intervals and the draft adaptation checklist past the advisory committee for approval or (if necessary) modification.		☐

References

Bieger T, Laesser C (2002) Market segmentation by motivation: the case of Switzerland. J Travel Res 41(1):68–76

Boztug Y, Babakhani N, Laesser C, Dolnicar S (2015) The hybrid tourist. Ann Tour Res 54:190–203

Cahill DJ (2006) Lifestyle market segmentation. Haworth Press, New York

Calantone RJ, Sawyer AG (1978) The stability of benefit segments. J Mark Res 15(3):395–404

Ehrnrooth H, Grönroos C (2013) The hybrid consumer: exploring hybrid consumption behaviour. Manag Decis 51(9):1793–1820

Farley JU, Winer RS, Lehmann DR (1987) Stability of membership in market segments identified with a disaggregate consumption model. J Bus Res 15(4):313–328

Ha SH, Bae SM, Park SC (2002) Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case. Comput Ind Eng 43:801–820

Haley RI (1985) Developing effective communications strategy – a benefit segmentation approach. Wiley, New York

Hu MY, Rau PA (1995) Stability of usage segments, membership shifts across segments and implications for marketing strategy: an empirical examination. Mid-Atl J Bus 31(2):161–177

Lilien GL, Rangaswamy A (2003) Marketing engineering: computer-assisted marketing analysis and planning, 2nd edn. Prentice Hall, Upper Saddle River

Lingras P, Hogo M, Snorek M, West C (2005) Temporal analysis of clusters of supermarket customers: conventional versus interval set approach. Inf Sci 172:215–240

McDonald M, Dunbar I (1995) Market segmentation: a step-by-step approach to creating profitable market segments. Macmillan, London

Müller H, Hamm U (2014) Stability of market segmentation with cluster analysis – a methodological approach. Food Qual Prefer 34:70–78

Oliveira M, Gama J (2010) MEC – monitoring clusters' transitions. In: Proceedings of the fifth starting AI researchers' symposium (STAIRS 2010), vol 222. IOS Press, pp 212–225

Paas LJ, Bijmolt THA, Vermunt JK (2015) Long-term developments of respondent financial product portfolios in the EU: a multilevel latent class analysis. Metron 73:249–262

Spiliopoulou M, Ntoutsi I, Theodoridis Y, Schult R (2006) MONIC: modeling and monitoring cluster transitions. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data Mining. ACM, New York, pp 706–711

Wind Y, Mahajan V, Gunther RE (2002) Convergence marketing – strategies for reaching the new hybrid consumer. Prentice Hall, Upper Saddle River

Yuspeh S, Fein G (1982) Can segments be born again? J Advert Res 22(3):13–22

Appendix A
Case Study: Fast Food

The purpose of this case study is to offer another illustration of market segmentation analysis using a different empirical data set.

This data set was collected originally for the purpose of comparing the validity of a range of different answer formats in survey research investigating brand image. Descriptions of the data are available in Dolnicar and Leisch (2012), Dolnicar and Grün (2014), and Grün and Dolnicar (2016). Package MSA contains the sections of the data used in this case study.

For this case study, imagine that you are McDonald's, and you would want to know if consumer segments exist that have a distinctly different image of McDonald's. Understanding such systematic differences of brand perceptions by market segments informs which market segments to focus on, and what messages to communicate to them. We can choose to focus on market segments with a positive perception, and strengthen the positive perception. Or we can choose to focus on a market segment that currently perceives McDonald's in a negative way. In this case, we want to understand the key drivers of the negative perception, and modify them.

A.1 Step 1: Deciding (not) to Segment

McDonald's can take the position that it caters to the entire market and that there is no need to understand systematic differences across market segments. Alternatively, McDonald's can take the position that, despite their market power, there is value in investigating systematic heterogeneity among consumers and harvest these differences using a differentiated marketing strategy.

A.2 Step 2: Specifying the Ideal Target Segment

McDonald's management needs to decide which key features make a market segment attractive to them. In terms of knock-out criteria, the target segment or target segments must be homogeneous (meaning that segment members are similar to one another in a key characteristic), distinct (meaning that members of the segments differ substantially from members of other segments in a key characteristic), large enough to justify the development and implementation of a customised marketing mix, matching the strengths of McDonald's (meaning, for example, that they must be open to eating at fast food restaurants rather than rejecting them outright), identifiable (meaning that there must be some way of spotting them among other consumers) and, finally, reachable (meaning that channels of communication and distribution need to exist which make it possible to aim at members of the target segment specifically).

In terms of segment attractiveness criteria, the obvious choice would be a segment that has a positive perception of McDonald's, frequently eats out and likes fast food. But McDonald's management could also decide that they not only wish to solidify their position in market segments in which they already hold high market shares, but rather wish to learn more about market segments which are currently not fond of McDonald's; try to understand which perceptions are responsible for this; and attempt to modify those very perceptions.

Given that the fast food data set in this case study contains very little information beyond people's brand image of McDonald's, the following attractiveness criteria will be used: liking McDonald's and frequently eating at McDonald's. These segment attractiveness criteria represent key information in Step 8 where they inform target segment selection.

A.3 Step 3: Collecting Data

The data set contains responses from 1453 adult Australian consumers relating to their perceptions of McDonald's with respect to the following attributes: YUMMY, CONVENIENT, SPICY, FATTENING, GREASY, FAST, CHEAP, TASTY, EXPENSIVE, HEALTHY, and DISGUSTING. These attributes emerged from a qualitative study conducted in preparation of the survey study. For each of those attributes, respondents provided either a YES response (indicating that they feel McDonald's possesses this attribute), or a NO response (indicating that McDonald's does not possess this attribute).

In addition, respondents indicated their AGE and GENDER. Had this data been collected for a real market segmentation study, additional information – such as details about their dining out behaviour, and their use of information channels – would have been collected to enable the development of a richer and more detailed description of each market segment.

A.4 Step 4: Exploring Data

First we explore the key characteristics of the data set by loading the data set and inspecting basic features such as the variable names, the sample size, and the first three rows of the data:

```
R> library("MSA")
R> data("mcdonalds", package = "MSA")
R> names(mcdonalds)

 [1] "yummy"           "convenient"      "spicy"
 [4] "fattening"       "greasy"          "fast"
 [7] "cheap"           "tasty"           "expensive"
[10] "healthy"         "disgusting"      "Like"
[13] "Age"             "VisitFrequency" "Gender"

R> dim(mcdonalds)

[1] 1453   15

R> head(mcdonalds, 3)

  yummy convenient spicy fattening greasy fast cheap tasty
1    No        Yes    No       Yes     No  Yes   Yes    No
2   Yes        Yes    No       Yes    Yes  Yes   Yes   Yes
3    No        Yes   Yes       Yes    Yes  Yes    No   Yes
  expensive healthy disgusting Like Age     VisitFrequency
1       Yes      No          No   -3  61 Every three months
2       Yes      No          No   +2  51 Every three months
3       Yes     Yes          No   +1  62 Every three months
  Gender
1 Female
2 Female
3 Female
```

As we can see from the output, the first respondent believes that McDonald's is not yummy, convenient, not spicy, fattening, not greasy, fast, cheap, not tasty, expensive, not healthy and not disgusting. This same respondent does not like McDonald's (rating of −3), is 61 years old, eats at McDonald's every three months and is female.

This quick glance at the data shows that the segmentation variables (perception of McDonald's) are verbal, not numeric. This means that they are coded using the words YES and NO. This is not a suitable format for segment extraction. We need numbers, not words. To get numbers, we store the segmentation variables in a separate matrix, and convert them from verbal YES/NO to numeric binary.

First we extract the first eleven columns from the data set because these columns contain the segmentation variables, and convert the data to a matrix. Then we identify all YES entries in the matrix. This results in a logical matrix with entries TRUE and FALSE. Adding 0 to the logical matrix converts TRUE to 1, and FALSE to 0. We check that we transformed the data correctly by inspecting the average value of each transformed segementation variable.

```
R> MD.x <- as.matrix(mcdonalds[, 1:11])
R> MD.x <- (MD.x == "Yes") + 0
R> round(colMeans(MD.x), 2)
```

```
     yummy convenient      spicy  fattening     greasy
      0.55       0.91       0.09       0.87       0.53
      fast      cheap      tasty  expensive    healthy
      0.90       0.60       0.64       0.36       0.20
disgusting
      0.24
```

The average values of the transformed binary numeric segmentation variables indicate that about half of the respondents (55%) perceive McDonald's as YUMMY, 91% believe that eating at McDonald's is CONVENIENT, but only 9% think that McDonald's food is SPICY.

Another way of exploring data initially is to compute a principal components analysis, and create a perceptual map. A perceptual map offers initial insights into how attributes are rated by respondents and, importantly, which attributes tend to be rated in the same way. Principal components analysis is not computed to reduce the number of variables. This approach – also referred to as factor-cluster analysis – is inferior to clustering raw data in most instances (Dolnicar and Grün 2008). Here, we calculate principal components because we use the resulting components to rotate and project the data for the perceptual map. We use unstandardised data because our segmentation variables are all binary.

```
R> MD.pca <- prcomp(MD.x)
R> summary(MD.pca)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     0.7570 0.6075 0.5046 0.3988 0.33741
Proportion of Variance 0.2994 0.1928 0.1331 0.0831 0.05948
Cumulative Proportion  0.2994 0.4922 0.6253 0.7084 0.76787
                          PC6     PC7     PC8     PC9
Standard deviation     0.3103 0.28970 0.27512 0.26525
Proportion of Variance 0.0503 0.04385 0.03955 0.03676
Cumulative Proportion  0.8182 0.86201 0.90156 0.93832
                         PC10    PC11
Standard deviation     0.24884 0.23690
Proportion of Variance 0.03235 0.02932
Cumulative Proportion  0.97068 1.00000
```

Results from principal components analysis indicate that the first two components capture about 50% of the information contained in the segmentation variables. The following command returns the factor loadings:

```
R> print(MD.pca, digits = 1)
```

```
Standard deviations (1, .., p=11):
 [1] 0.8 0.6 0.5 0.4 0.3 0.3 0.3 0.3 0.3 0.2 0.2
```

```
Rotation (n x k) = (11 x 11):
               PC1   PC2   PC3    PC4    PC5   PC6   PC7
yummy        0.477 -0.36  0.30 -0.055 -0.308  0.17 -0.28
convenient   0.155 -0.02  0.06  0.142  0.278 -0.35 -0.06
spicy        0.006 -0.02  0.04 -0.198  0.071 -0.36  0.71
fattening   -0.116  0.03  0.32  0.354 -0.073 -0.41 -0.39
greasy      -0.304  0.06  0.80 -0.254  0.361  0.21  0.04
fast         0.108  0.09  0.06  0.097  0.108 -0.59 -0.09
cheap        0.337  0.61  0.15 -0.119 -0.129 -0.10 -0.04
tasty        0.472 -0.31  0.29  0.003 -0.211 -0.08  0.36
expensive   -0.329 -0.60 -0.02 -0.068 -0.003 -0.26 -0.07
healthy      0.214 -0.08 -0.19 -0.763  0.288 -0.18 -0.35
disgusting -0.375  0.14  0.09 -0.370 -0.729 -0.21 -0.03
               PC8    PC9   PC10   PC11
yummy        0.01 -0.572  0.110  0.045
convenient  -0.11  0.018  0.666 -0.542
spicy        0.38 -0.400  0.076  0.142
fattening    0.59  0.161  0.005  0.251
greasy      -0.14  0.003 -0.009  0.002
fast        -0.63 -0.166 -0.240  0.339
cheap        0.14 -0.076 -0.428 -0.489
tasty       -0.07  0.639 -0.079  0.020
expensive    0.03 -0.067 -0.454 -0.490
healthy      0.18  0.186  0.038  0.158
disgusting -0.17  0.072  0.290 -0.041
```

The loadings indicate how the original variables are combined to form principal components. Loadings guide the interpretation of principal components. In our example, the two segmentation variables with the highest loadings (in absolute terms) for principal component 2 are CHEAP and EXPENSIVE, indicating that this principal component captures the price dimension. We project the data into the principal component space with `predict`. The following commands rotate and project consumers (in grey) into the first two principal components, plot them and add the rotated and projected original segmentation variables as arrows:

```
R> library("flexclust")
R> plot(predict(MD.pca), col = "grey")
R> projAxes(MD.pca)
```

Figure A.1 shows the resulting perceptual map. The attributes CHEAP and EXPENSIVE play a key role in the evaluation of McDonald's, and these two attributes are assessed quite independently of the others. The remaining attributes align with what can be interpreted as positive versus negative perceptions: FATTENING, DISGUSTING and GREASY point in the same direction in the perceptual chart, indicating that respondents who view McDonald's as FATTENING, DISGUSTING are also likely to view it as GREASY. In the opposite direction are the positive attributes FAST, CONVENIENT, HEALTHY, as well as TASTY and YUMMY. The observations along the EXPENSIVE versus CHEAP axis cluster around three values: a group of consumers at the top around the arrow pointing to CHEAP, a group of respondents

Fig. A.1 Principal components analysis of the fast food data set

at the bottom around the arrow pointing to EXPENSIVE, and a group of respondents in the middle.

These initial exploratory insights represent valuable information for segment extraction. Results indicate that some attributes are strongly related to one another, and that the price dimension may be critical in differentiating between groups of consumers.

A.5 Step 5: Extracting Segments

Step 5 is where we extract segments. To illustrate a range of extraction techniques, we subdivide this step into three sections. In the first section, we will use standard *k*-means analysis. In the second section, we will use finite mixtures of binary distributions. In the third section, we will use finite mixtures of regressions.

A.5.1 Using k-Means

We calculate solutions for two to eight market segments using standard *k*-means analysis with ten random restarts (argument `nrep`). We then relabel segment numbers such that they are consistent across segmentations.

```
R> set.seed(1234)
R> MD.km28 <- stepFlexclust(MD.x, 2:8, nrep = 10,
+    verbose = FALSE)
R> MD.km28 <- relabel(MD.km28)
```

We extract between two and eight segments because we do not know in advance what the best number of market segments is. If we calculate a range of solutions, we can compare them and choose the one which extracts segments containing similar consumers which are distinctly different from members of other segments.

We compare different solutions using a scree plot:

```
R> plot(MD.km28, xlab = "number of segments")
```

where `xlab` specifies the label of the x-axis.

The scree plot in Fig. A.2 has no distinct elbow: the sum of distances within market segments drops slowly as the number of market segments increases. We expect the values to decrease because more market segments automatically mean that the segments are smaller and, as a consequence, that segment members are more similar to one another. But the much anticipated point where the sum of distances drops dramatically is not visible. This scree plot does not provide useful guidance on the number of market segments to extract.

A second approach to determining a good number of segments is to use stability-based data structure analysis. Stability-based data structure analysis also indicates whether market segments occur naturally in the data, or if they have to be artificially constructed. Stability-based data structure analysis uses stability across replications as criterion to offer this guidance. Imagine using a market segmentation solution which cannot be reproduced. Such a solution would give McDonald's management little confidence in terms of investing substantial resources into a market segmentation strategy. Assessing the stability of segmentation solutions across repeated calculations (Dolnicar and Leisch 2010) ensures that unstable, random solutions are not used.

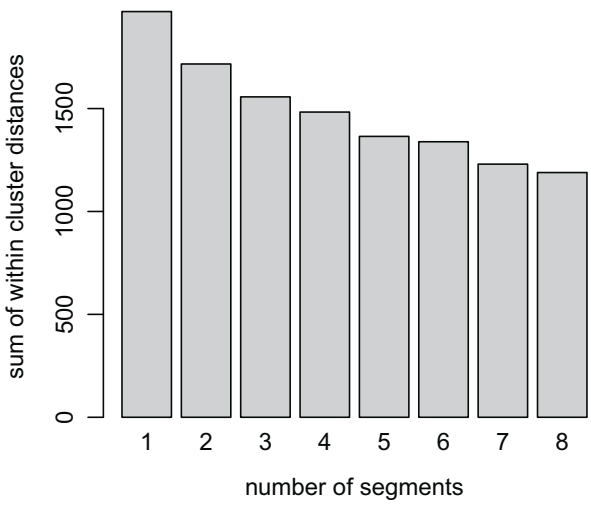


Fig. A.2 Scree plot for the fast food data set

Global stability is the extent to which the same segmentation solution emerges if the analysis is repeated many times using bootstrap samples (randomly drawn subsets) of the data. Global stability is calculated using the following R code, which conducts the analysis for each number of segments (between two and eight) using 2 × 100 bootstrap samples (argument `nboot`) and ten random initialisations (argument `nrep`) of *k*-means for each sample and number of segments:

```
R> set.seed(1234)
R> MD.b28 <- bootFlexclust(MD.x, 2:8, nrep = 10,
+   nboot = 100)
```

We obtain the global stability boxplot shown in Fig. A.3 using:

```
R> plot(MD.b28, xlab = "number of segments",
+   ylab = "adjusted Rand index")
```

The vertical boxplots show the distribution of stability for each number of segments. The median is indicated by the fat black horizontal line in the middle of the box. Higher stability is better.

Inspecting Fig. A.3 points to the two-, three- and four-segment solutions as being quite stable. However, the two- and three-segment solutions do not offer a very differentiated view of the market. Solutions containing a small number of segments typically lack the market insights managers are interested in. Once we increase the number of segments to five, average stability drops quite dramatically. The four-segment solution thus emerges as the solution containing the most market segments which can still be reasonably well replicated if the calculation is repeated multiple times.

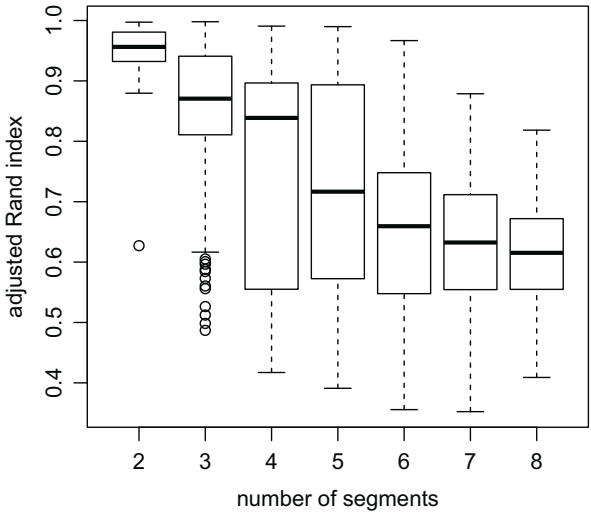


Fig. A.3 Global stability of *k*-means segmentation solutions for the fast food data set

We gain further insights into the structure of the four-segment solution with a gorge plot:

```
R> histogram(MD.km28[["4"]], data = MD.x, xlim = 0:1)
```

None of the segments shown in Fig. A.4 is well separated from the other segments, and proximity to at least one other segment is present as indicated by the similarity values all being between 0.3 and 0.7.

The analysis of global stability is based on a comparison of segmentation solutions with the same number of segments. Another way of exploring the data before committing to the final market segmentation solution is to inspect how segment memberships change each time an additional market segment is added, and to assess segment level stability across solutions. This information is contained in the segment level stability across solutions (SLS_A) plot created by `slsaplot(MD.km28)` and shown in Fig. A.5.

Thick green lines indicate that many members of the segment to the left of the line move across to the segment on the right side of the line. Segment 2 in the two-segment solution (in the far left column of the plot) remains almost unchanged until the four-segment solution, then it starts losing members. Looking at the segment level stability across solutions (SLS_A) plot in Fig. A.5 in view of the earlier determination that the four-segment solution looks good, it can be concluded that segments 2, 3 and 4 are nearly identical to the corresponding segments in

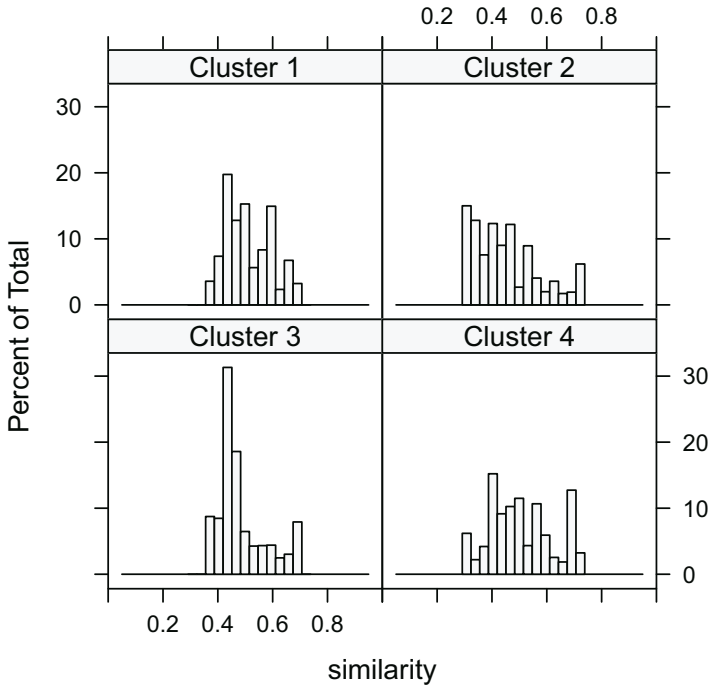


Fig. A.4 Gorge plot of the four-segment *k*-means solution for the fast food data set

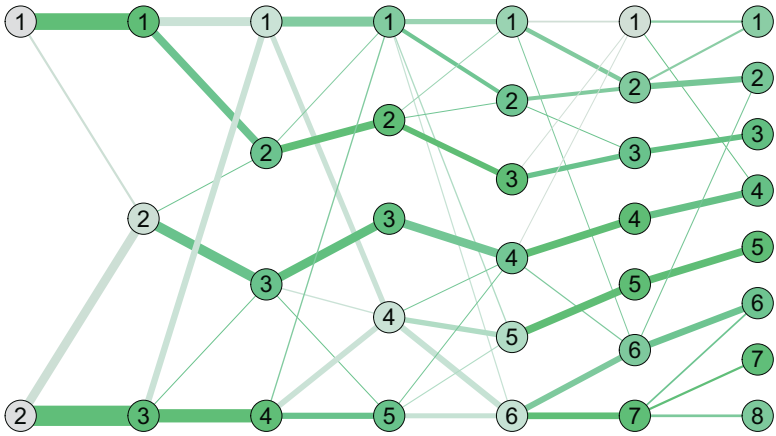


Fig. A.5 Segment level stability across solutions (SLS_A) plot from two to eight segments for the fast food data set

the three- and five-segment solution. They display high stability across solutions with different numbers of segments. Segment 1 in the four-segment solution is very different from both the solutions with one fewer and one more segments. Segment 1 draws members from two segments in the three-segment solution, and splits up again into two segments contained in the five-segment solution. This highlights that – while the four-segment solution might be a good overall segmentation solution – segment 1 might not be a good target segment because of this lack of stability.

After this exploration, we select the four-segment solution and save it in an object of its own:

```
R> MD.k4 <- MD.km28[["4"]]
```

By definition, global stability assesses the stability of a segmentation solution in its entirety. It does not investigate the stability of each market segment. We obtain the stability of each segment by calculating segment level stability within solutions (SLS_W):

```
R> MD.r4 <- slswFlexclust(MD.x, MD.k4)
```

We plot the result with limits 0 and 1 for the y-axis (`ylim`) and customised labels for both axes (`xlab`, `ylab`) using:

```
R> plot(MD.r4, ylim = 0:1, xlab = "segment number",
+   ylab = "segment stability")
```

Figure A.6 shows the segment level stability within solutions for the four-segment solution. Segment 1 is the least stable across replications, followed by segments 4 and 2. Segment 3 is the most stable. The low stability levels for segment 1 are not unexpected given the low stability this segment has when comparing segment level stability across solutions (see Fig. A.5).

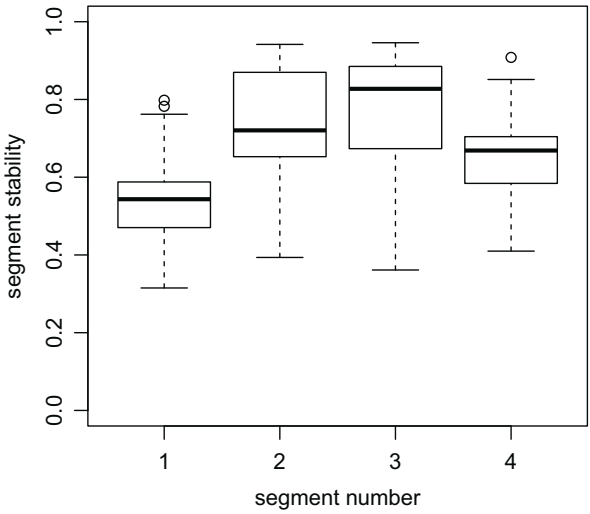


Fig. A.6 Segment level stability within solutions (SLS_W) plot for the fast food data set

A.5.2 Using Mixtures of Distributions

We calculate latent class analysis using a finite mixture of binary distributions. The mixture model maximises the likelihood to extract segments (as opposed to minimising squared Euclidean distance, as is the case for *k*-means). The call to `stepFlexmix()` extracts two to eight segments (`k = 2:8`) using ten random restarts of the EM algorithm (`nrep`), `model = FLXMCmvbinary()` for a segment-specific model consisting of independent binary distributions and no intermediate output about progress (`verbose = FALSE`).

```
R> library("flexmix")
R> set.seed(1234)
R> MD.m28 <- stepFlexmix(MD.x ~ 1, k = 2:8, nrep = 10,
+    model = FLXMCmvbinary(), verbose = FALSE)
R> MD.m28

Call:
stepFlexmix(MD.x ~ 1, model = FLXMCmvbinary(),
    k = 2:8, nrep = 10, verbose = FALSE)

  iter converged k k0    logLik      AIC      BIC      ICL
2   32      TRUE 2  2 -7610.848 15267.70 15389.17 15522.10
3   43      TRUE 3  3 -7311.534 14693.07 14877.92 15077.96
4   33      TRUE 4  4 -7111.146 14316.29 14564.52 14835.95
5   61      TRUE 5  5 -7011.204 14140.41 14452.01 14806.54
6   49      TRUE 6  6 -6956.110 14054.22 14429.20 14810.65
7   97      TRUE 7  7 -6900.188 13966.38 14404.73 14800.16
8  156      TRUE 8  8 -6872.641 13935.28 14437.01 14908.52
```

We plot the information criteria with a customised label for the y-axis to choose a suitable number of segments:

```
R> plot(MD.m28,
+    ylab = "value of information criteria (AIC, BIC, ICL)")
```

Figure A.7 plots the information criteria values AIC, BIC and ICL on the y-axis for the different number of components (segments) on the x-axis. As can be seen, the values of all information criteria decrease quite dramatically until four components (market segments) are reached. If the information criteria are strictly applied based on statistical inference theory, the ICL recommends – by a small margin – the extraction of seven market segments. The BIC also points to seven market segments. The AIC values continue to decrease beyond seven market segments, indicating that at least eight components are required to suitably fit the data.

The visual inspection of Fig. A.7 suggests that four market segments might be a good solution if a more pragmatic point of view is taken; this is the point at which the decrease in the information criteria flattens visibly. We retain the four-component solution and compare it to the four-cluster k-means solution presented in Sect. A.5.1 using a cross-tabulation:

```
R> MD.m4 <- getModel(MD.m28, which = "4")
R> table(kmeans = clusters(MD.k4),
+    mixture = clusters(MD.m4))

       mixture
kmeans   1   2   3   4
     1   1 191 254  24
```

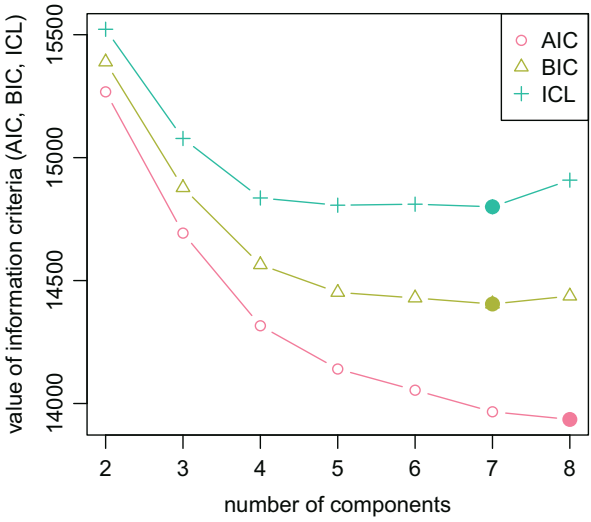


Fig. A.7 Information criteria for the mixture models of binary distributions with 2 to 8 components (segments) for the fast food data set

```
2 200   0  25  32
3   0  17   0 307
4   0 384   2  16
```

Component (segment) members derived from the mixture model are shown in columns, cluster (segment) members derived from *k*-means are shown in rows. Component 2 of the mixture model draws two thirds all of its members (384) from segment 4 of the *k*-means solution. In addition, 191 members are recruited from segment 1. This comparison shows that the stable segments in the *k*-means solution (numbers 2 and 3) are almost identical to segments (components) 1 and 4 of the mixture model. This means that the two segmentation solutions derived using very different extraction methods are actually quite similar.

The result becomes even more similar if the mixture model is initialised using the segment memberships of the *k*-means solution `MD.k`:

```
R> MD.m4a <- flexmix(MD.x ~1, cluster = clusters(MD.k4),
+   model = FLXMCmvbinary())
R> table(kmeans = clusters(MD.k4),
+   mixture = clusters(MD.m4a))

      mixture
kmeans   1   2   3   4
     1 278   1  24 167
     2  26 200  31   0
     3   0   0 307  17
     4   2   0  16 384
```

This is interesting because all algorithms used to extract market segments are exploratory in nature. Typically, therefore, they find a local optimum or global optimum of their respective target function. The EM algorithm maximises the log-likelihood. The log-likelihood values for the two fitted mixture models obtained using the two different ways of initialisation are:

```
R> logLik(MD.m4a)

'log Lik.' -7111.152 (df=47)

R> logLik(MD.m4)

'log Lik.' -7111.146 (df=47)
```

indicating that the values are very close, with random initialisations leading to a slightly better result.

If two completely different ways of initialising the mixture model, namely (1) ten random restarts and keeping the best, and (2) initialising the mixture model using the *k*-means solution, yield almost the same result, this gives more confidence that the result is a global optimum or a reasonably close approximation to the global optimum. It also is a re-assurance for the *k*-means solution, because the extracted segments are essentially the same. The fact that the two solutions are not identical is not of concern. Neither of the solutions is correct or incorrect. Rather, both of them need to be inspected and may be useful to managers.

A.5.3 Using Mixtures of Regression Models

Instead of finding market segments of consumers with similar perceptions of McDonald's, it may be interesting to find market segments containing members whose love or hate for McDonald's is driven by similar perceptions. This segmentation approach would enable McDonald's to modify critical perceptions selectively for certain target segments in view of improving love and reducing hate.

We extract such market segments using finite mixtures of linear regression models, also called latent class regressions. Here, the variables are not all treated in the same way. Rather, one dependent variable needs to be specified which captures the information predicted using the independent variables. We choose as dependent variable y the degree to which consumers love or hate McDonald's. The dependent variable contains responses to the statement I LIKE MCDONALDS. It is measured on an 11-point scale with endpoints labelled I LOVE IT! and I HATE IT!. The independent variables x are the perceptions of McDonald's. In this approach the segmentation variables can be regarded as unobserved, and consisting of the regression coefficients. This means market segments consist of consumers for whom changes in perceptions have similar effects on their liking of McDonald's.

First we create a numerical dependent variable by converting the ordinal variable LIKE to a numeric one. We need a numeric variable to fit mixtures of linear regression models. The categorical variable has 11 levels, from I LOVE IT!(+5) with numeric code 1 to I HATE IT!(-5) with numeric code 11. Computing 6 minus the numeric code will result in $6 - 11 = -5$ for I HATE IT!-5, $6 - 10 = -4$ for `"-4"`, etc.:

```
R> rev(table(mcdonalds$Like))
```

```
I hate it!-5           -4           -3           -2
         152           71           73           59
          -1            0           +1           +2
          58          169          152          187
          +3           +4 I love it!+5
         229          160          143
```

```
R> mcdonalds$Like.n <- 6 - as.numeric(mcdonalds$Like)
R> table(mcdonalds$Like.n)
```

```
 -5  -4  -3  -2  -1   0   1   2   3   4   5
152  71  73  59  58 169 152 187 229 160 143
```

Then we can either create a model formula for the regression model manually by typing the eleven variable names, and separating them by plus signs. Or we can automate this process in R by first collapsing the eleven independent variables into a single string separated by plus signs, and then pasting the dependent variable `Like.n` to it. Finally, we convert the resulting string to a formula.

```
R> f <- paste(names(mcdonalds)[1:11], collapse = "+")
R> f <- paste("Like.n ~ ", f, collapse = "")
R> f <- as.formula(f)
R> f
```

```
Like.n ~ yummy + convenient + spicy + fattening + greasy +
    fast + cheap + tasty + expensive + healthy + disgusting
```

We fit a finite mixture of linear regression models with the EM algorithm using `nrep = 10` random starts and `k = 2` components. We ask for the progress of the EM algorithm not to be visible on screen during estimation (`verbose = FALSE`):

```
R> set.seed(1234)
R> MD.reg2 <- stepFlexmix(f, data = mcdonalds, k = 2,
+   nrep = 10, verbose = FALSE)
R> MD.reg2

Call:
stepFlexmix(f, data = mcdonalds, k = 2, nrep = 10,
    verbose = FALSE)

Cluster sizes:
  1   2
630 823

convergence after 68 iterations
```

Mixtures of regression models can only be estimated if certain conditions on the x and y variables are met (Hennig 2000; Grün and Leisch 2008b). Even if these conditions are met, estimation problems can occur. In this section we restrict the fitted mixture model to two components. Fitting a mixture model with more components to the data would lead to problems during segment extraction.

Using the degree of loving or hating McDonald's as dependent variable will cause problems if we want to extract many market segments because the dependent variable is not metric. It is ordinal where we use the assigned scores with values -5 to $+5$. Having an ordinal variable implies that groups of respondents exist in the data who all have the exactly same value for the dependent variable. This means that we can extract, for example, a group consisting only of respondents who gave a score of $+5$. The regression model for this group perfectly predicts the value of the dependent variable if the intercept equals $+5$ and the other regression coefficients are set to zero. A mixture of regression models containing this component would have an infinite log-likelihood value and represent a degenerate solution. Depending on the starting values, the EM algorithm might converge to a segmentation solution containing such a component. The more market segments are extracted, the more likely is the EM algorithm to converge against such a degenerate solution.

The fitted mixture model contains two linear regression models, one for each component. We assess the significance of the parameters of each regression model with:

```
R> MD.ref2 <- refit(MD.reg2)
R> summary(MD.ref2)

$Comp.1
               Estimate Std. Error  z value  Pr(>|z|)
(Intercept)   -4.347851   0.252058 -17.2494 < 2.2e-16 ***
```

```
yummyYes        2.399472   0.203921  11.7667 < 2.2e-16 ***
convenientYes   0.072974   0.148060   0.4929  0.622109
spicyYes       -0.070388   0.175200  -0.4018  0.687864
fatteningYes   -0.544184   0.183931  -2.9586  0.003090 **
greasyYes       0.079760   0.115052   0.6933  0.488152
fastYes         0.361220   0.170346   2.1205  0.033964 *
cheapYes        0.437888   0.157721   2.7763  0.005498 **
tastyYes        5.511496   0.216265  25.4850 < 2.2e-16 ***
expensiveYes    0.225642   0.150979   1.4945  0.135037
healthyYes      0.208154   0.149607   1.3913  0.164121
disgustingYes -0.562942   0.140337  -4.0114 6.037e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Comp.2
              Estimate Std. Error z value  Pr(>|z|)
(Intercept)   -0.90694    0.41921 -2.1635  0.030505 *
yummyYes       2.10884    0.18731 11.2586 < 2.2e-16 ***
convenientYes  1.43443    0.29576  4.8499 1.235e-06 ***
spicyYes      -0.35793    0.23745 -1.5074  0.131715
fatteningYes  -0.34899    0.21932 -1.5912  0.111556
greasyYes     -0.47748    0.15015 -3.1800  0.001473 **
fastYes        0.42103    0.23223  1.8130  0.069837 .
cheapYes      -0.15675    0.20698 -0.7573  0.448853
tastyYes      -0.24508    0.23428 -1.0461  0.295509
expensiveYes  -0.11460    0.21312 -0.5378  0.590745
healthyYes     0.52806    0.18761  2.8146  0.004883 **
disgustingYes -2.07187    0.21011 -9.8611 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the stars in the far right column, we see that members of segment 1 (component 1) like McDonald's if they perceive it as YUMMY, NOT FATTENING, FAST, CHEAP, TASTY. and NOT DISGUSTING. Members of segment 2 (component 2) like McDonald's if they perceive it as YUMMY, CONVENIENT, NOT GREASY, HEALTHY, and NOT DISGUSTING.

Comparing the regression coefficients of the two components (segments) is easier using a plot. Argument `significance` controls the shading of bars to reflect the significance of parameters:

```
R> plot(MD.ref2, significance = TRUE)
```

Figure A.8 shows regression coefficients in dark grey if the corresponding estimate is significant. The default significance level is $\alpha = 0.05$, and multiple testing is not accounted for. Insignificant coefficients are light grey. The horizontal lines at the end of the bars give a 95% confidence interval for each regression coefficient of each segment.

We interpret Fig. A.8 as follows: members of segment 1 (component 1) like McDonald's if they perceive it as yummy, fast, cheap and tasty, but not fattening

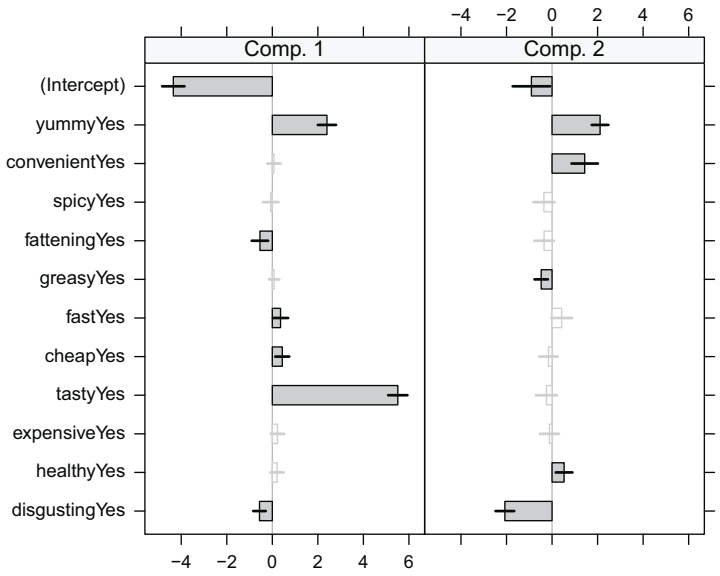


Fig. A.8 Regression coefficients of the two-segment mixture of linear regression models for the fast food data set

and disgusting. For members of segment 1, liking McDonald's is not associated with their perception of whether eating at McDonald's is convenient, and whether food served at McDonald's is healthy. In contrast, perceiving McDonald's as convenient and healthy is important to segment 2 (component 2). Using the perception of healthy as an example: if segment 2 is targeted, it is important for McDonald's to convince segment members that McDonald's serves (at least some) healthy food items. The health argument is unnecessary for members of segment 1. Instead, this segment wants to hear about how good the food tastes, and how fast and cheap it is.

A.6 Step 6: Profiling Segments

The core of the segmentation analysis is complete: market segments have been extracted. Now we need to understand what the four-segment k-means solution means. The first step in this direction is to create a segment profile plot. The segment profile plot makes it easy to see key characteristics of each market segment. It also highlights differences between segments. To ensure the plot is easy to interpret, similar attributes should be positioned close to one another. We achieve this by calculating a hierarchical cluster analysis. Hierarchical cluster analysis used on attributes (rather than consumers) identifies – attribute by attribute – the most similar ones.

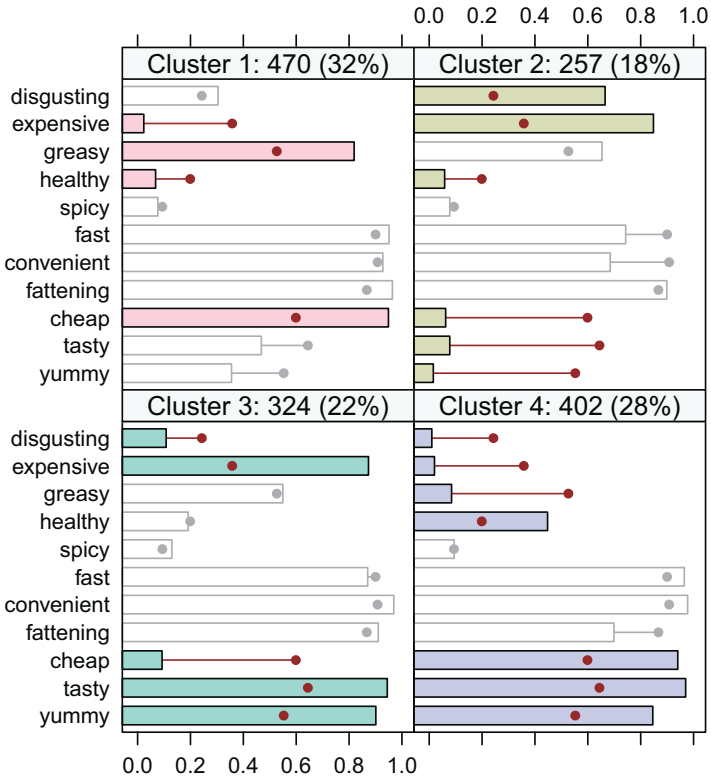


Fig. A.9 Segment profile plot for the four-segment solution for the fast food data set

```
R> MD.vclust <- hclust(dist(t(MD.x)))
```

The ordering of the segmentation variables identified by hierarchical clustering is then used (argument `which`) to create the segment profile plot. Marker variables are highlighted (`shade = TRUE`):

```
R> barchart(MD.k4, shade = TRUE,
+    which = rev(MD.vclust$order))
```

Figure A.9 is easy for McDonald's managers to interpret. They can see that there are four market segments. They can also see the size of each market segment. The smallest segment (segment 2) contains 18% of consumers, the largest (segment 1) 32%. The names of the segmentation variables (attributes) are written on the left side of the plot. The horizontal lines with the dot at the end indicate the percentage of respondents in the entire sample who associate each perception with McDonald's. The bars plot the percentage of respondents *within each segment* who associate each perception with McDonald's. Marker variables are coloured differently for each segment. All other variables are greyed out. Marker variables differ from the overall

sample percentage either by more than 25% points in absolute terms, or by more than 50% in relative terms.

To understand the market segments, McDonald's managers need to do two things: (1) compare the bars for each segment with the horizontal lines to see what makes each segment distinct from all consumers in the market; and (2) compare bars across segments to identify differences between segments.

Looking at Fig. A.9, we see that segment 1 thinks McDonald's is cheap and greasy. This is a very distinct perception. Segment 2 views McDonald's as disgusting and expensive. This is also a very distinct perception, setting apart members of this segment from all other consumers. Members of segment 3 share the view that McDonald's is expensive, but also think that the food served at McDonald's is tasty and yummy. Finally, segment 4 is all praise: members of this market segment believe that McDonald's food is tasty, yummy and cheap and at least to some extent healthy.

Another visualisation that can help managers grasp the essence of market segments is the segment separation plot shown in Fig. A.10. The segment separation plot can be customised with additional arguments. We choose not to plot the hulls around the segments (`hull = FALSE`), to omit the neighbourhood graph (`simlines = FALSE`), and to label both axes (`xlab, ylab`):

```
R> plot(MD.k4, project = MD.pca, data = MD.x,
+   hull = FALSE, simlines = FALSE,
+   xlab = "principal component 1",
+   ylab = "principal component 2")
R> projAxes(MD.pca)
```

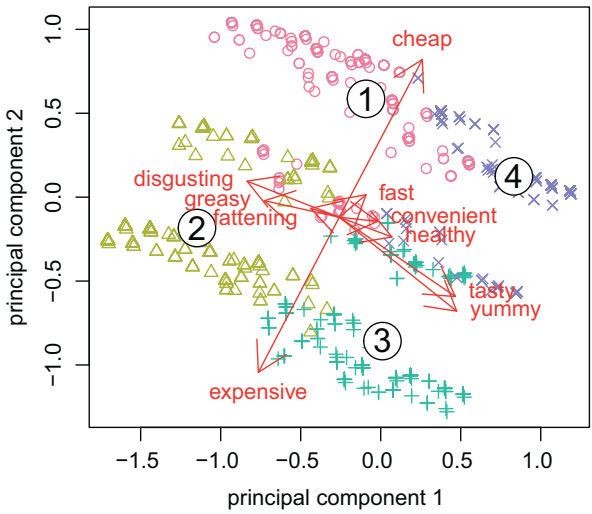


Fig. A.10 Segment separation plot using principal components 1 and 2 for the fast food data set

Figure A.10 looks familiar because we have already used principal components analysis to explore data in Step 4 (Fig. A.1). Here, the centres of each market segment are added using black circles containing the segment number. In addition, observations are coloured to reflect segment membership.

As can be seen, segments 1 and 4 both view McDonald's as cheap, with members of segment 4 holding – in addition – some positive beliefs and members of segment 1 associating McDonald's primarily with negative attributes. At the other end of the price spectrum, segments 2 and 3 agree that McDonald's is not cheap, but disagree on other features with segment 2 holding a less flattering view than members of segment 3.

At the end of Step 6 McDonald's managers have a good understanding of the nature of the four market segments in view of the information that was used to create these segments. Apart from that, they know little about the segments. Learning more about them is the key aim of Step 7.

A.7 Step 7: Describing Segments

The fast food data set is not typical for data collected for market segmentation analysis because it contains very few descriptor variables. Descriptor variables – additional pieces of information about consumers – are critically important to gaining a good understanding of market segments. One descriptor variable available in the fast food data set is the extent to which consumers love or hate McDonald's. Using a simple mosaic plot, we can visualise the association between segment membership and loving or hating McDonald's.

To do this, we first extract the segment membership for each consumer for the four-segment solution. Next we cross-tabulate segment membership and the love-hate variable. Finally, we generate the mosaic plot with cells colours indicating the deviation of the observed frequencies in each cell from the expected frequency if variables are not associated (`shade = TRUE`). We do not require a title for our mosaic plot (`main = ""`), but we would like the x-axis to be labelled (`xlab`):

```
R> k4 <- clusters(MD.k4)
R> mosaicplot(table(k4, mcdonalds$Like), shade = TRUE,
+   main = "", xlab = "segment number")
```

The mosaic plot in Fig. A.11 plots segment number along the x-axis, and loving or hating McDonald's along the y-axis. The mosaic plot reveals a strong and significant association between those two variables. Members of segment 1 (depicted in the first column) rarely express love for McDonald's, as indicated by the top left boxes being coloured in red. In stark contrast, members of segment 4 are significantly more likely to love McDonald's (as indicated by the dark blue boxes in the top right of the mosaic plot). At the same time, these consumers are less likely to hate McDonald's (as indicated by the very small red boxes at the bottom right of the plot). Members of segment 2 appear to have the strongest negative feelings towards McDonald's; their likelihood of hating McDonald's is extremely high (dark

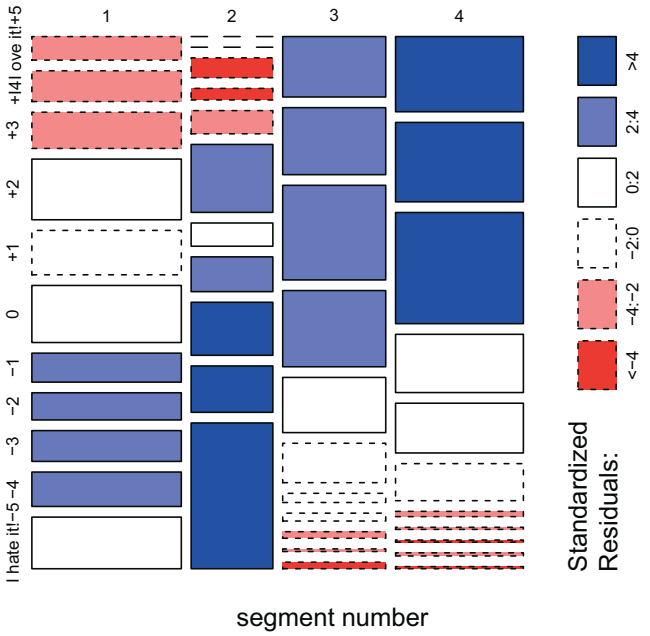


Fig. A.11 Shaded mosaic plot for cross-tabulation of segment membership and I LIKE IT for the fast food data set

blue boxes at the bottom of the second column), and nearly none of the consumers in this segment love McDonald's (tiny first and second box at the top of column two, then dark red third and fourth box).

The fast food data contains a few other basic descriptor variables, such as gender and age. Figure A.12 shows gender distribution across segments. We generate this figure using the command:

```
R> mosaicplot(table(k4, mcdonalds$Gender), shade = TRUE)
```

Market segments are plotted along the x-axis. The descriptor variable (gender) is plotted along the y-axis. The mosaic plot offers the following additional insights about our market segments: segment 1 and segment 3 have a similar gender distribution as the overall sample. Segment 2 contains significantly more men (as depicted by the larger blue box for the category male, and the smaller red box for the category female in the second column of the plot). Members of segment 4 are significantly less likely to be men (smaller red box at the top of the fourth column).

Because age is metric – rather than categorical – we use a parallel box-and-whisker plot to assess the association of age with segment membership. We generate Fig. A.13 using the R command `boxplot(mcdonalds$Age ~ k4, varwidth = TRUE, notch = TRUE)`.

Figure A.13 plots segments along the x-axis, and age along the y-axis. We see immediately that the notches do not overlap, suggesting significant differences in average age across segments. A more detailed inspection reveals that members of

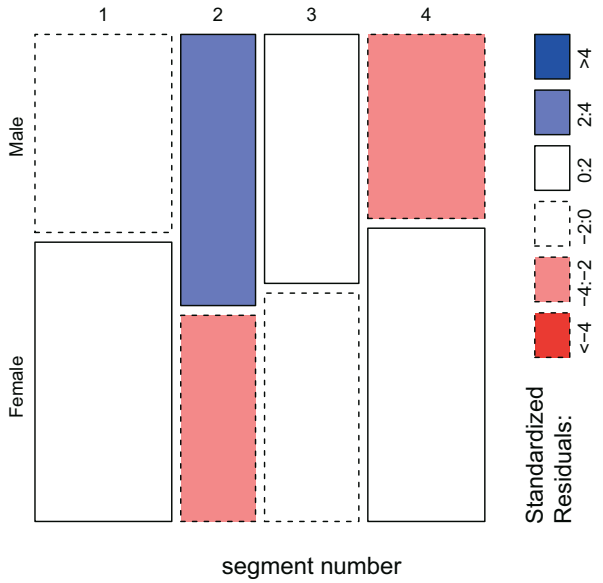


Fig. A.12 Shaded mosaic plot for cross-tabulation of segment membership and gender for the fast food data set

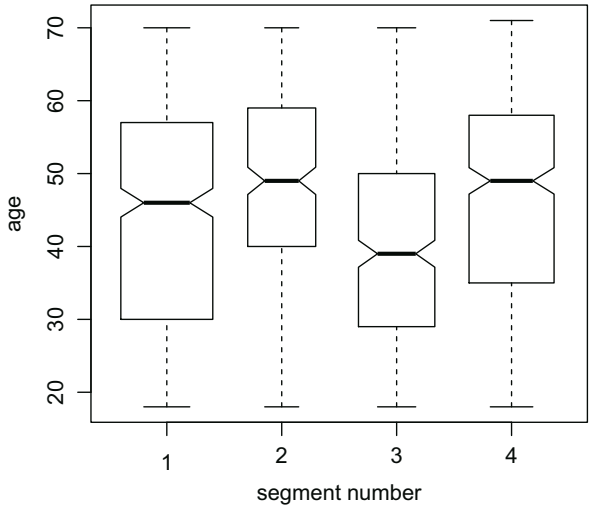


Fig. A.13 Parallel box-and-whisker plot of age by segment for the fast food data set

segment 3 – consumers who think McDonald's is yummy and tasty, but expensive – are younger than the members of all other segments. The parallel box-and-whisker plot shows this by (1) the box being in lower position; and (2) the notch

in the middle of the box being lower and not overlapping with the notches of the other boxes.

To further characterise market segments with respect to the descriptor variables, we try to predict segment membership using descriptor variables. We do this by fitting a conditional inference tree with segment 3 membership as dependent variable, and all available descriptor variables as independent variables:

```
R> library("partykit")
R> tree <- ctree(
+   factor(k4 == 3) ~ Like.n + Age +
+   VisitFrequency + Gender,
+   data = mcdonalds)
R> plot(tree)
```

Figure A.14 shows the resulting classification tree. The independent variables used in the tree are LIKE.N, AGE and VISITFREQUENCY. GENDER is not used to split the respondents into groups. The tree indicates that respondents who like McDonald's, and are young (node 10), or do not like McDonald's, but visit it more often than once a month (node 8), have the highest probability to belong to segment 3. In contrast, respondents who give a score of -4 or worse for liking McDonald's, and visit McDonald's once a month at most (node 5), are almost certainly not members of segment 3.

Optimally, additional descriptor variables would be available. Of particular interest would be information about product preferences, frequency of eating at a fast food restaurant, frequency of dining out in general, hobbies and frequently used information sources (such as TV, radio, newspapers, social media). The availability of such information allows the data analyst to develop a detailed description of each market segment. A detailed description, in turn, serves as the basis for tasks conducted in Step 9 where the perfect marketing mix for the selected target segment is designed.

A.8 Step 8: Selecting (the) Target Segment(s)

Using the knock-out criteria and segment attractiveness criteria specified in Step 2, users of the market segmentation (McDonald's managers) can now proceed to develop a segment evaluation plot.

The segment evaluation plot in Fig. A.15 is extremely simplified because only a small number of descriptor variables are available for the fast food data set. In Fig. A.15 the frequency of visiting McDonald's is plotted along the x-axis. The extent of liking or hating McDonald's is plotted along the y-axis. The bubble size represents the percentage of female consumers.

We can obtain the values required to construct the segment evaluation plot using the following commands. First, we compute the mean value of the visiting frequency of McDonald's for each segment.

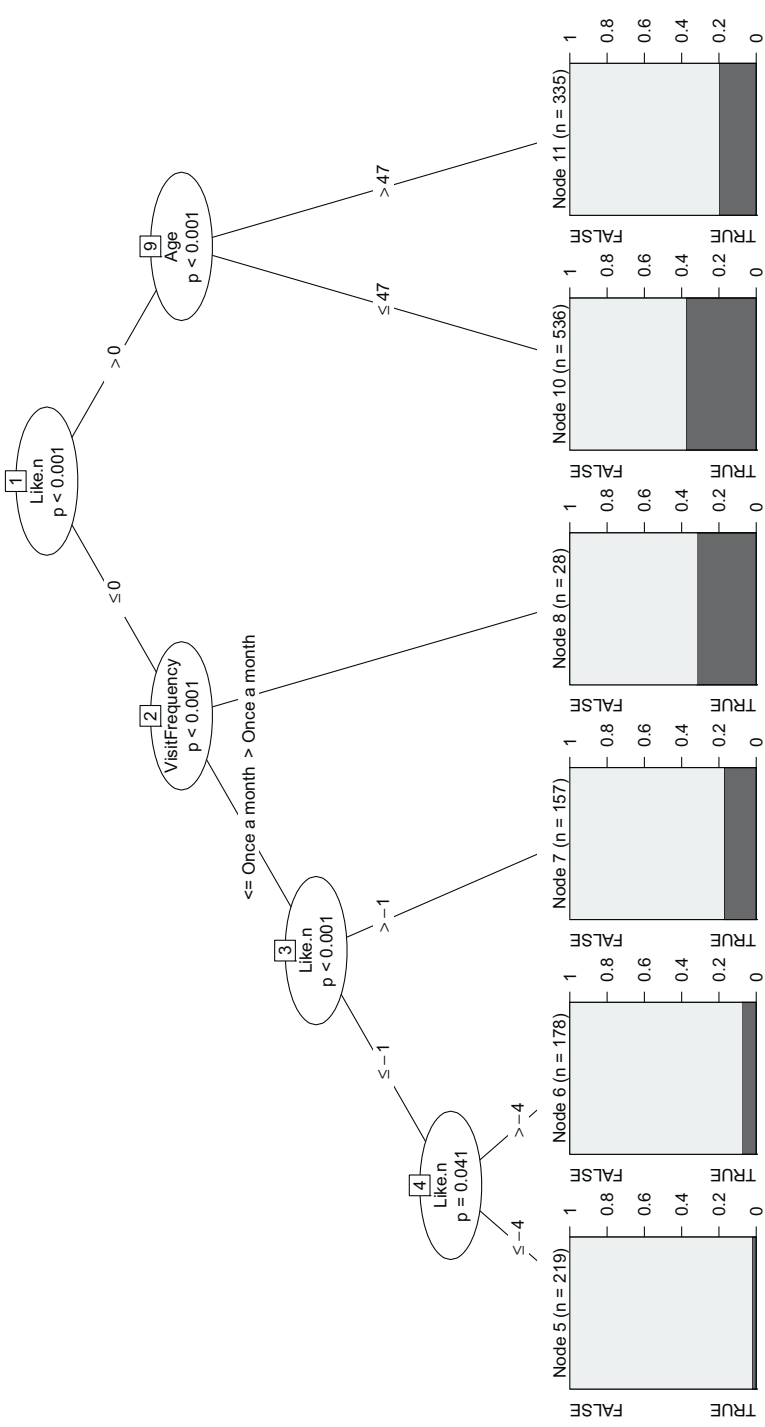


Fig. A.14 Conditional inference tree using segment 3 membership as dependent variable for the fast food data set

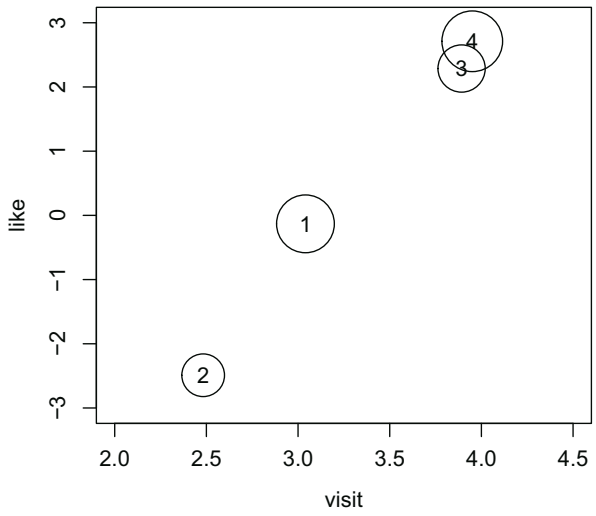


Fig. A.15 Example of a simple segment evaluation plot for the fast food data set

```
R> visit <- tapply(as.numeric(mcdonalds$VisitFrequency),
+   k4, mean)
R> visit
```

```
       1        2        3        4
3.040426 2.482490 3.891975 3.950249
```

Function `tapply()` takes as arguments a variable (here VISITFREQUENCY converted to numeric), a grouping variable (here segment membership `k4`), and a function to be used as a summary statistic for each group (here `mean`). A numeric version of liking McDonald's is already stored in LIKE.N. We can use this variable to compute mean segment values:

```
R> like <- tapply(mcdonalds$Like.n, k4, mean)
R> like
```

```
         1          2          3          4
-0.1319149 -2.4902724  2.2870370  2.7114428
```

We need to convert the variable GENDER to numeric before computing mean segment values:

```
R> female <- tapply((mcdonalds$Gender == "Female") + 0,
+   k4, mean)
R> female
```

```
        1         2         3         4
0.5851064 0.4319066 0.4783951 0.6144279
```

Now we can create the segment evaluation plot using the following commands:

```
R> plot(visit, like, cex = 10 * female,
+   xlim = c(2, 4.5), ylim = c(-3, 3))
R> text(visit, like, 1:4)
```

Argument `cex` controls the size of the bubbles. The scaling factor of 10 is a result of manual experimentation. Arguments `xlim` and `ylim` specify the ranges for the axes.

Figure A.15 represents a simplified example of a segment evaluation plot. Market segments 3 and 4 are located in the attractive quadrant of the segment evaluation plot. Members of these two segments like McDonald's and visit it frequently. These segments need to be retained, and their needs must be satisfied in the future. Market segment 2 is located in the least attractive position. Members of this segment hate McDonald's, and rarely eat there, making them unattractive as a potential market segment. Market segment 1 does not currently perceive McDonald's in a positive way, and feels that it is expensive. But in terms of loving McDonald's and visitation frequency, members of market segment 1 present as a viable target segment. Marketing action could attempt to address the negative perceptions of this segment, and re-inforce positive perceptions. As a result, McDonald's may be able to broaden its customer base.

The segment evaluation plot serves as a useful decision support tool for McDonald's management to discuss which of the four market segments should be targeted and, as such, become the focus of attention in Step 9.

A.9 Step 9: Customising the Marketing Mix

In Step 9 the marketing mix is designed. If, for example, McDonald's managers decide to focus on segment 3 (young customers who like McDonald's, think the food is yummy and tasty, but perceive it as pretty expensive), they could choose to offer a MCSUPERBUDGET line to cater specifically to the price expectations of this segment (4Ps: Price). The advantage of such an approach might be that members of segment 3 develop to become loyal customers who, as they start earning more money, will not care about the price any more and move to the regular McDonald's range of products. To not cannibalise the main range, the product features of the MCSUPERBUDGET range would have to be distinctly different (4Ps: Product). Next, communication channels would have to be identified which are heavily used by members of segment 3 to communicate the availability of the MCSUPERBUDGET line (4Ps: Promotion). Distribution channels (4Ps: Place) would have to be the same given that all McDonald's food is sold in McDonald's outlets. But McDonald's management could consider having a MCSUPERBUDGET lane where the wait in the queue might be slightly longer in an attempt not to cannibalise the main product line.

A.10 Step 10: Evaluation and Monitoring

After the market segmentation analysis is completed, and all strategic and tactical marketing activities have been undertaken, the success of the market segmentation strategy has to be evaluated, and the market must be carefully monitored on a continuous basis. It is possible, for example, that members of segment 3 start earning more money and the MCSUPERBUDGET line is no longer suitable for them. Changes can occur within existing market segments. But changes can also occur in the larger marketplace, for example, if new competitors enter the market. All potential sources of change have to be monitored in order to detect changes which require McDonald's management to adjust their strategic or tactical marketing in view of new market circumstances.

Appendix B
R and R Packages

B.1 What Is R?

B.1.1 A Short History of R

R started in 1992 as a small software project initiated by Ross Ihaka and Robert Gentleman. A first open source version was made available in 1995. In 1997 the R Core Development Team was formed. The R Core Development Team consists of about 20 members, including the two inventors of R, who maintain the base distribution of R. R implements a variation of a programming language called S (as in *Statistics*) which was developed by John Chambers and colleagues in the 1970s and 1980s. Chambers was awarded the Association for Computing Machinery (ACM) Software Systems Award in 1998 for S, which was predicted will forever alter the way people analyse, visualise, and manipulate data (ACM 1999). Chambers also serves as member of the R Core Development Team.

R is open source software; anyone can download the source code for R from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org at no cost. More importantly, CRAN makes available executables for Linux, Apple MacOS and Microsoft Windows. CRAN is a network of dozens of servers distributed across many countries across all continents to minimise download time.

Over the last two decades, R has become what some call the "lingua franca of computational statistics" (de Leeuw and Mair 2007, p. 2). Initially only known to specialists, R is now used for teaching and research in universities all over the world. R is particularly attractive to educational institutions because it reduces software licence fees and trains students in a language they can use after their studies independently of the software their employer uses. R has also been adopted enthusiastically by businesses and organisations across a wide range of industries.

Entering the single letter R in a web search engine, returns as top hits the R homepage (https://www.R-project.org), and the Wikipedia entry for R, highlighting the substantial global interest in R.

B.1.2 R Packages

R organises its functionality in so-called *packages*. The most fundamental package is called base, without which R cannot work. The base package has no statistical functionality itself, its only purpose is to handle data, interact with the operating system, and load other packages. The first thing a new R user needs to install, therefore, is the *base system of R* which contains the interpreter for the R programming language, and a selection of numeric and graphic statistical methods for a wide range of data analysis applications.

Each R package can be thought of as a book. A collection of R packages is a library. Packages come in three priority categories:

Base packages: Base packages are fundamental packages providing computational infrastructure. The base packages datasets, graphics and stats provide data sets used in examples, a comprehensive set of data visualisation functions (scatter plots, bar plots, histograms, ...), and a comprehensive set of statistical methods (descriptive statistics, classical tests, linear and generalized linear models, clustering, distribution functions, random number generators, ...). All base packages are maintained by the R Core Development Team, and are contained in the base system of R.

Recommended packages: To provide even more statistical methods in every R installation, installers of the software for most operating systems also include a set of so-called recommended packages with more specialised functionality. Examples include lattice for conditioning plots (Sarkar 2008), mgcv for generalised additive models (Wood 2006), and nlme for mixed effects models (Pinheiro et al. 2017).

Contributed packages: The vast majority of R packages is contributed by the R user community. *Contributed* packages are not necessarily of lower quality, but – as opposed to recommended packages – they are not automatically distributed with every R installation. The wide array of contributed packages, and the continuing increase in the number of those packages, make R particularly attractive, as they represent an endless resource of code.

Not surprisingly, therefore, the backbone of R's success is that everybody can contribute to the project by developing their own packages. In December 2017 some 12,000 extension packages were available on CRAN. Many more R packages are available on private web pages or other repositories. These offer a wide variety of data analytic methodology, several of which can be used for market segmentation and are introduced in this book. R packages can be automatically installed and updated from CRAN using commands like `install.packages()` or `update.packages()`, respectively. Packages can be loaded into an R session using the command `library("pkgname")`.

A typical R package is a collection of R code and data sets together with help pages for both the R code and the data sets. Not all packages have both components; some contain only code, others only data sets. In addition, packages can contain manuals, vignettes or test code for quality assurance.

B.1.3 Quality Control

The fact that R is available for free could be misinterpreted as an indicator of low quality or lack of quality control. Very popular and competitive software projects like the Firefox browser or the Android smartphone operating system are also open source. Successful large open source projects usually have rigid measures for quality control, and R is no exception.

Every change to the R base code is only accepted if a long list of tests is passed successfully. These tests compare calculations pre-stored with earlier versions of R with results from the current version, making sure that $2 + 2$ is still 4 and not all of a sudden 3 or 5. All examples in all help pages are executed to see if the code runs without errors. A battery of tests is also run on every R package on CRAN on a daily basis for the current release and development versions of R for various versions of four different operating systems (Windows, MacOS, Linux, Solaris). The results of all these checks and the R bug repository can be browsed by the interested public online.

B.1.4 User Interfaces for R

Most R users do not interact with R using the interface provided by the base installation. Rather, they choose one of several alternatives, depending on operating system and level of sophistication. The basic installation for Windows has menus for opening R script files (text files with R commands), installing packages from CRAN or opening help pages and manuals shipped with R.

If new users want to start learning R without typing commands, several graphical user interfaces offer direct access to statistical methods using point and click. The most comprehensive and popular graphical user interface (GUI) for R is the R Commander (Fox 2017); it has a menu structure similar to that of IBM SPSS (IBM Corporation 2016). The R Commander can be installed using the command `install.packages("Rcmdr")`, and started from within R using `library("Rcmdr")`. The R Commander has been translated to almost 20 languages. The R Commander can also be extended; other R packages can add new menus and sub-menus to the interface.

Once a user progresses to interacting with R using commands, it becomes helpful to use a text editor with syntax support. The Windows version of R has a small script editor, but more powerful editors exist. Note that Microsoft Word and similar

programs are not text editors and not suitable for the task. R does not care if a command is bold, italic, small or large. All that matters is that commands are syntactically valid (for example: all parentheses that are opened, must be closed). Text editors for programming languages assist data analysts in ensuring syntactic validity by, for example, highlighting the opening parenthesis when one is closed. Numerous text editors now support the R language, and several can connect to a running R process. In such cases, R code is entered into a window of the text editor and can be sent by keyboard shortcuts or pressing buttons to R for evaluation.

If a new user does not have a preferred editor, a good recommendation is to use RStudio which is freely available for all major operating systems from https://www.RStudio.com. A popular choice for Linux users is to run R inside the Emacs editor using the Emacs extension package ESS (Emacs Speaks Statistics) available at https://ess.R-project.org/.

B.2 R Packages Used in the Book

B.2.1 MSA

Package MSA is the companion package to this book. It contains most of the data sets used in the book and all R code as demos:

`step-4`:	Exploring data.
`step-5-2`:	Extracting segments: distance-based clustering (hierarchical, partitioning, hybrid approaches).
`step-5-3`:	Extracting segments: model-based clustering (finite mixture models).
`step-5-4`:	Extracting segments: algorithms with variable selection (biclustering, VSBD).
`step-5-5`:	Extracting segments: data structure analysis (cluster indices, gorge plots, global and segment-level stability analysis).
`step-6`:	Profiling segments (segment profile plot, segment separation plot).
`step-7`:	Describing segments (graphics and inference).
`step-8`:	Selecting (the) target segment(s) (segment evaluation plot).
`step-9`:	Customising the marketing mix.
`step-10`:	Evaluation and monitoring.
`case-study`:	Case study: fast food (all 10 steps).

For example, to run all code from Step 4, use the command
`demo("step-4", package = "MSA")` in R. For a detailed summary of all data sets see Appendix C. In addition the package also contains functions written as part of the book:

`clusterhulls()`:	Plot data with cluster hulls.
`decisionMatrix()`:	Segment evaluation plot.
`twoStep()`:	Infer segment membership for two-step clustering.
`vsbd()`:	Select variables for binary clustering using the algorithm proposed by Brusco (2004).

B.2.2 flexclust

flexclust is the R package for partitioning cluster analysis, stability-based data structure analysis and segment visualisation (Leisch 2006, 2010; Dolnicar and Leisch 2010, 2014, 2017). The most important functions and methods for the book are:

`barchart()`:	Segment profile plot.
`bclust()`:	Bagged clustering.
`bootFlexclust()`:	Global stability analysis.
`cclust()`:	*k*-means, hard competitive learning, neural gas.
`plot()`:	Segment separation plot.
`priceFeature()`:	Artificial mobile phone data.
`slsaplot()`:	Segment level stability across solutions.
`slswFlexclust()`:	Segment level stability within solutions.
`stepcclust()`:	Repeated random initialisations of `cclust()`.
`stepFlexclust()`:	Repeated random initialisations of a given clustering algorithm.

B.2.3 flexmix

flexmix is the R package for flexible finite mixture modelling (Leisch 2004; Grün and Leisch 2007; Grün and Leisch 2008b). The most important functions for the book are:

`flexmix()`:	Finite mixtures of distributions and regression models.
`stepFlexmix()`:	Repeated random initialisations of `flexmix()`.

B.2.4 Other Packages

The following R packages were also used for computations and visualisations in the book. Base packages are not listed because they are part of every R installation

and do not need to be downloaded from CRAN individually. Packages are listed in alphabetical order:

biclust:	A collection of several bi-clustering procedures.
car:	A collection of tools for applied regression.
cluster:	A collection of methods for cluster analysis including the calculation of dissimilarity matrices.
deldir:	Compute and plot a Voronoi partition corresponding to a clustering (Turner 2017).
effects:	Visualise effects for regression models.
kohonen:	Self-organising maps (SOMs).
lattice:	Trellis graphics.
mclust:	Model-based clustering with multivariate normal distributions.
mlbench:	A collection of benchmark data sets from the UCI Machine Learning Repository (Leisch and Dimitriadou 2012).
nnet:	Software for feed-forward neural networks with a single hidden layer, and for multinomial log-linear models.
partykit:	A toolkit for recursive partitioning.
xtable:	Convert R tables and model summaries to HTML or LaTeX (Dahl 2016).

Appendix C
Data Sets Used in the Book

C.1 Tourist Risk Taking

Year of data collection: 2015.

Location: Australia.

Sample size: 563.

Sample: Adult Australian residents.

Screening: Respondents must have undertaken at least one holiday in the last year which involved staying away from home for at least four nights.

Segmentation variables used in the book:
Six variables on frequency of risk taking. Respondents were asked: *Which risks have you taken in the past?*

- Recreational risks (e.g., rock-climbing, scuba diving)
- Health risks (e.g., smoking, poor diet, high alcohol consumption)
- Career risks (e.g., quitting a job without another to go to)
- Financial risks (e.g., gambling, risky investments)
- Safety risks (e.g., speeding)
- Social risks (e.g., standing for election, publicly challenging a rule or decision)

Response options provided to respondents (integer code in parentheses):

- NEVER (1)
- RARELY (2)
- QUITE OFTEN (3)
- OFTEN (4)
- VERY OFTEN (5)

Descriptor variables used in this book: None.

Purpose of data collection: Academic research into improving market segmentation methodology as well as the potential usefulness of peer-to-peer accommodation networks for providing emergency accommodation in case of a disaster hitting a tourism destination.

Data collected by: Academic researchers using a permission based online panel.

Ethics approval: #2015001433 (The University of Queensland, Australia).

Funding source: Australian Research Council (DP110101347).

Prior publications using this data: Hajibaba and Dolnicar (2017); Hajibaba et al. (2017).

Availability: Data set `risk` in R package MSA and online at http://www.MarketSegmentationAnalysis.org.

C.2 Winter Vacation Activities

Year of data collection: Winter tourist seasons 1991/92 and 1997/98.

Location: Austria.

Sample size: 2878 (1991/92), 2961 (1997/98).

Sample: Adult tourists spending their holiday in Austria.

Sampling: Quota sampling by state and accommodation used.

Screening: Tourists to capital cities are excluded.

Segmentation variables used in the book:
Twenty seven binarised travel activities for season 1997/98, a subset of eleven binarised travel activities is also available for season 1991/92 and marked by asterisks (*). Numeric codes are 1 for DONE and 0 for NOT DONE.

- Alpine skiing *
- Cross-country skiing *
- Snowboarding
- Carving
- Ski touring *
- Ice-skating *
- Sleigh riding *
- Tennis
- Horseback riding
- Going to a spa
- Using health facilities

- Hiking *
- Going for walks
- Organized excursions
- Excursions
- Relaxing *
- Going out in the evening
- Going to discos/bars
- Shopping *
- Sight-seeing *
- Museums *
- Theater/opera
- Visiting a "Heurigen"
- Visiting concerts
- Visiting "Tyrolean evenings"
- Visiting local events
- Going to the pool/sauna *

Descriptor variables used in this book: None.

Purpose of data collection: These data sets are from two waves of the Austrian National Guest Survey conducted in three-yearly intervals by the Austrian National Tourism Organisation to gain market insight for the purpose of strategy development. The format of data collection has since changed.

Data collected by: Austrian Society for Applied Research in Tourism (ASART) for the Austrian National Tourism Organisation (Österreich Werbung).

Funding source: Austrian National Tourism Organisation (Österreich Werbung).

Prior publications using this data: Dolnicar and Leisch (2003).

Availability: Data sets `winterActiv` and `winterActiv2` (containing the two objects `wi91act` and `wi97act`) in R package MSA and online at http://www.MarketSegmentationAnalysis.org.

C.3 Australian Vacation Activities

Year of data collection: 2007.

Location: Australia.

Sample size: 1003.

Sample: Adult Australian residents.

Segmentation variables used in the book:
Forty five binarised vacation activities, integer codes are 1 for DONE and 0 for NOT DONE.

- Bush or rainforest walks (BUSHWALK)
- Visit the beach (including swimming and sunbathing) (BEACH)
- Visit farms/tour countryside (FARM)
- Whale/dolphin watching (in the ocean) (WHALE)
- Visit botanical or other public gardens (GARDENS)
- Going camping (CAMPING)
- Swimming (beach, pool or river) (SWIMMING)
- Snow activities (e.g. snowboarding/skiing) (SKIING)
- Tennis (TENNIS)
- Horse riding (RIDING)
- Cycling (CYCLING)
- Hiking/Climbing (HIKING)
- Exercise/gym /swimming (at a local pool or river) (EXERCISING)
- Play golf (GOLF)
- Go fishing (FISHING)
- Scuba diving/Snorkelling (SCUBADIVING)
- Surfing (SURFING)
- Four wheel driving (FOURWHEEL)
- Adventure activities (e.g. bungee jumping, hang gliding, white water rafting, etc.) (ADVENTURE)
- Other water sports (e.g. sailing, windsurfing, kayaking, waterskiing/wake boarding, etc.) (WATERSPORT)
- Attend theatre, concerts or other performing arts (THEATRE)
- Visit history/heritage buildings, sites or monuments (MONUMENTS)
- Experience aboriginal art/craft and cultural displays (CULTURAL)
- Attend festivals/fairs or cultural events (FESTIVALS)
- Visit museums or art galleries (MUSEUM)
- Visit amusements/theme parks (THEMEPARK)
- Charter boat/cruise/ferry ride (CHARTERBOAT)
- Visit a health or beauty spa/get a massage (SPA)
- Going for scenic walks or drives/general sightseeing (SCENICWALKS)
- Going to markets (street/weekend/art/craft markets) (MARKETS)
- Go on guided tour or excursion (GUIDEDTOURS)
- Visit industrial tourism attractions (e.g. breweries, mines, wineries) (INDUSTRIAL)
- Visit wildlife parks/zoos/aquariums (WILDLIFE)
- Visit attractions for the children (CHILDRENATT)
- General sightseeing (SIGHTSEEING)
- Visit friends & relatives (FRIENDS)
- Pubs, clubs, discos, etc. (PUBS)
- Picnics/BBQ's (BBQ)
- Go shopping (pleasure) (SHOPPING)
- Eating out in reasonably priced places (EATING)
- Eating out in upmarket restaurants (EATINGHIGH)
- Watch movies (MOVIES)
- Visit casinos (CASINO)

- Relaxing/doing nothing (RELAXING)
- Attend an organised sporting event (SPORTEVENT)

Descriptor variables used in the book:

- Fourteen information sources for vacation planning, integer codes are 1 (indicating use) and 0 (indicating no use).
  - Destination information brochures (INFO.BROCHURES.DESTINATION)
  - Brochures from hotels (INFO.BROCHURES.HOTEL)
  - Brochures from tour operator (INFO.BROCHURES.TOUR.OPERATOR)
  - Information from travel agent (INFO.TRAVEL.AGENT)
  - Information from tourist info centre (INFO.TOURIST.CENTRE)
  - Advertisements in newspapers/journals (INFO.ADVERTISING.NEWSPAPERS)
  - Travel guides/books/journals (INFO.TRAVEL.GUIDES)
  - Information given by friends and relatives (INFO.FRIENDS.RELATIVES)
  - Information given by work colleagues (INFO.WORK.COLLEAGUES)
  - Radio programs (INFO.RADIO)
  - TV programs (INFO.TV)
  - Internet (INFO.INTERNET)
  - Exhibitions/fairs (INFO.EXHIBITIONS)
  - Slide nights (INFO.SLIDE.NIGHTS)
- Six sources to book accommodation, integer codes are 1 (used during last Australian holiday) and 0 (not used).
  - Internet (BOOK.INTERNET)
  - Phone (BOOK.PHONE)
  - Booked on arrival at destination (BOOK.AT.DESTINATION)
  - Travel agent (BOOK.TRAVEL.AGENT)
  - Other (BOOK.OTHER)
  - Someone else in my travel party booked it (BOOK.SOMEONE.ELSE)
- Spend per person per day during the last Australian holiday (numeric in AUD) (SPENDPPPD).
- TV channel watched most often (TV.CHANNEL).

Purpose of data collection: PhD thesis.

Data was collected by: Katie Cliff (née Lazarevski).

Funding source: Australian Research Council (DP0557769).

Ethics approval: HE07/068 (University of Wollongong, Australia).

Prior publications using this data: Cliff (2009), Dolnicar et al. (2012).

Availability: Data sets `ausActiv` and `ausActivDesc` in R package MSA and online at http://www.MarketSegmentationAnalysis.org.

C.4 Australian Travel Motives

Year of data collection: 2006.

Location: Australia.

Sample size: 1000.

Sample: Adult Australian residents.

Segmentation variables used in the book:
Twenty travel motives, integer codes are 1 (for applies) and 0 (for does not apply).

- I want to rest and relax (REST AND RELAX)
- I am looking for luxury and want to be spoilt (LUXURY / BE SPOILT)
- I want to do sports (DO SPORTS)
- This holiday means excitement, a challenge and special experience to me (EXCITEMENT, A CHALLENGE)
- I try not to exceed my planned budget for this holiday (NOT EXCEED PLANNED BUDGET)
- I want to realise my creativity (REALISE CREATIVITY)
- I am looking for a variety of fun and entertainment (FUN AND ENTERTAINMENT)
- Good company and getting to know people is important to me (GOOD COMPANY)
- I use my holiday for the health and beauty of my body (HEALTH AND BEAUTY)
- I put much emphasis on free-and-easy-going (FREE-AND-EASY-GOING)
- I spend my holiday at a destination, because there are many entertainment facilities (ENTERTAINMENT FACILITIES)
- Being on holiday I do not pay attention to prices and money (NOT CARE ABOUT PRICES)
- I am interested in the life style of the local people (LIFE STYLE OF THE LOCAL PEOPLE)
- The special thing about my holiday is an intense experience of nature (INTENSE EXPERIENCE OF NATURE)
- I am looking for cosiness and a familiar atmosphere (COSINESS/FAMILIAR ATMOSPHERE)
- On holiday the efforts to maintain unspoilt surroundings play a major role for me (MAINTAIN UNSPOILT SURROUNDINGS)
- It is important to me that everything is organised and I do not have to care about anything (EVERYTHING ORGANISED)
- When I choose a holiday-resort, an unspoilt nature and a natural landscape plays a major role for me (UNSPOILT NATURE/NATURAL LANDSCAPE)
- Cultural offers and sights are a crucial factor (CULTURAL OFFERS)
- I go on holiday for a change to my usual surroundings (CHANGE OF SURROUNDINGS)

The three numeric descriptor variables OBLIGATION, NEP, VACATION.BEHAVIOUR (see below) are also used as segmentation variables to illustrate the use of model-based methods.

Descriptor variables used in the book:

- Gender (FEMALE, MALE)
- Age (numeric)
- Education (numeric, minimum 1, maximum 8)
- Income (LESS THAN $30,000, $30,001 TO $60,000, $60,001 TO $90,000, $90,001 TO $120,000, $120,001 TO $150,000, $150,001 TO $180,000, $180,001 TO $210,000, $210,001 TO $240,000, MORE THAN $240,001)
- Re-coded income (<30K, 30–60 K, 60–90 K, 90–120 K, >120K)
- Occupation (CLERICAL OR SERVICE WORKER, PROFESSIONAL, UNEMPLOYED, RETIRED, MANAGER OR ADMINISTRATOR, SALES, TRADESPERSON, SMALL BUSINESS OWNER, HOME-DUTIES, TRANSPORT WORKER, LABOURER)
- State (NSW, VIC, QLD, SA, WA, TAS, NT, ACT)
- Relationship status (SINGLE, MARRIED, SEPARATED OR DIVORCED, LIVING WITH A PARTNER, WIDOWED)
- Stated moral obligation to protect the environment (OBLIGATION: numeric, minimum 1, maximum 5).
- Re-coded stated moral obligation to protect the environment (OBLIGATION2: re-coded ordered factor by quartiles: Q1, Q2, Q3, Q4).
- Mean New Ecological Paradigm (NEP) scale value (NEP: numeric, minimum 1, maximum 5).
- Mean environmental friendly behaviour score when on vacation (VACATION.BEHAVIOUR: numeric, minimum 1, maximum 5).

Purpose of data collection: Academic research into public acceptance of water from alternative sources.

Data was collected by: Academic researchers using a permission based online panel.

Funding source: Australian Research Council (DP0557769).

Ethics approval: HE08/328 (University of Wollongong, Australia).

Prior publications using this data: Dolnicar and Leisch (2008a,b).

Availability: Data set `vacmot` (containing the three objects `vacmot`, `vacmot6` and `vacmotdesc`) in R package flexclust and online at http://www.MarketSegmentationAnalysis.org.

C.5 Fast Food

Year of data collection: 2009.

Location: Australia.

Sample size: 1453.

Sample: Adult Australian residents.

Segmentation variables used in the book:
Eleven attributes on the perception of McDonald's measured on a binary scale, all categorical with levels YES and NO.

- yummy
- convenient
- spicy
- fattening
- greasy
- fast
- cheap
- tasty
- expensive
- healthy
- disgusting

The descriptor variable LIKE (see below) is also used as dependent variable when fitting a mixture of linear regression models.

Descriptor variables used in the book:

- Age (numeric)
- Gender (FEMALE, MALE)
- Love or hate McDonald's restaurants (LIKE: measured using a bipolar 11-point scale with levels I LOVE IT!+5, +4,. . . , −4, I HATE IT!−5)
- Visiting frequency of McDonald's restaurants (VISITFREQUENCY: measured on a 6-point scale with levels NEVER, ONCE A YEAR, EVERY THREE MONTHS, ONCE A MONTH, ONCE A WEEK, MORE THAN ONCE A WEEK)

Purpose of data collection: Comparative study of the stability of survey responses in dependence of answer formats offered to respondents.

Data was collected by: Sara Dolnicar, John Rossiter.

Funding source: Australian Research Council (DP0878423).

Ethics approval: HE08/331 (University of Wollongong, Australia).

Prior publications using this data:
Dolnicar and Leisch (2012), Dolnicar and Grün (2014), Grün and Dolnicar (2016).

Availability: Data set `mcdonalds` in R package MSA, and online at http://www.MarketSegmentationAnalysis.org.

Glossary

Adjusted Rand index: The adjusted Rand index measures how similar two market segmentation solutions are while correcting for agreement by chance. The adjusted Rand index is 1 if two market segmentation solutions are identical and 0 if the agreement between the two market segmentation solutions is the same as expected by chance.

A priori market segmentation: Also referred to as commonsense segmentation or convenience group segmentation, this segmentation approach uses only one (or a very small number) of segmentation variables to group consumers into segments. The segmentation variables are known in advance, and determine the nature of market segments. For example, if age is used, age segments are the result. The success of a priori market segmentation depends on the relevance of the chosen segmentation variable, and on the detailed description of resulting market segments. A priori market segmentation is methodologically simpler than a posteriori or post hoc or data-driven market segmentation, but is not necessarily inferior. If the segmentation variable is highly relevant, it may well represent the optimal approach to market segmentation for an organisation.

A posteriori market segmentation: Also referred to as data-driven market segmentation or post hoc segmentation, a posteriori market segmentation uses a set of segmentation variables to extract market segments. Segmentation variables used are typically similar in nature, for example, a set of vacation activities. The nature of the resulting segmentation solution is known in advance (for example: vacation activity segmentation). But, in contrast to commonsense segmentation, the characteristics of the emerging segments with respect to the segmentation variables are not known in advance. Resulting segments need to be both profiled and described in detail before one or a small number of target segments are selected.

Artificial data: Artificial data is data created by a data analyst. The properties of artificial data – such as the number and shape of market segments contained – are known. Artificial data is critical to the development and comparative assessment

of methods in market segmentation analysis because alternative methods can be evaluated in terms of their ability to reveal the true structure of the data. The true structure of empirical consumer data is never known.

Attractiveness criteria: See segment attractiveness criteria.

Behavioural segmentation: Behavioural segmentation is the result of using information about human behaviour as segmentation variable(s). Examples include scanner data from supermarkets, or credit card expenditure data.

Bootstrapping: Bootstrapping is a statistical term for random sampling with replacement. Bootstrapping is useful in market segmentation to explore randomness when only a single data sample is available. Bootstrapping plays a key role in stability-based data structure analysis, which helps to prevent the selection of an inferior, not replicable segmentation solution.

Box-and-whisker plot: The box-and-whisker plot (or boxplot) visualises the distribution of a unimodal metric variable. Parallel boxplots allow to compare the distribution of metric variables across market segments. It is a useful tool for the description of market segments using metric descriptor variables, such as age, or dollars spent.

Centroid: The mathematical centre of a cluster (market segment) used in distance-based partitioning clustering or segment extraction methods such as *k*-means. The centroid can be imagined as the prototypical segment member; the best representative of all members of the segment.

Classification: Classification is the statistical problem of learning a prediction algorithm where the predicted variable is a nominal variable. Classification is also referred to as *supervised learning* in machine learning. Logistic regression or recursive partitioning algorithms are examples for classification algorithms. Classification algorithms can be used to describe market segments.

Commonsense segmentation: See a priori market segmentation.

Constructive segmentation: The concept of constructive segmentation has to be used when the segmentation variables are found (in stability-based data structure analysis) to contain no structure. As a consequence of the lack of data structure, repeated segment extractions lead to different market segmentation solutions. This is not optimal, but from a managerial point of view it still often makes sense to treat groups of consumers differently. Therefore, in constructive market segmentation, segments are artificially constructed. The process of constructive market segmentation requires collaboration of the data analyst and the user of the market segmentation solution. The data analyst's role is to offer alternative segmentation solutions. The user's role is to assess which of the many possible groupings of consumers is most suitable for the segmentation strategy of the organisation.

Convenience group market segmentation: See a priori market segmentation.